Multimodal and Multitask Representation Learning For Perceiving Embodied Interactions

by

Md Mofijul Islam

A dissertation document submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
in
Systems Engineering

Department of Systems and Information Engineering
University of Virginia
June, 2023

Doctoral Committee:

Tariq Iqbal (Advisor), Assistant Professor, Systems and Information Engineering
Laura Barnes (Chair), Professor, Systems and Information Engineering
Aidong Zhang, Professor, Computer Science
Sara Riggs, Associate Professor, Systems and Information Engineering
Dan Bohus, Senior Principal Researcher, Microsoft Research

Multimodal and Multitask Representation Learning For Perceiving Embodied Interactions

---

A

Dissertation

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

---

in partial fulfillment

of the requirements for the degree

Doctor of Philosophy

by

Md Mofijul Islam

August 2023

APPROVAL SHEET

This

Dissertation

is submitted in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

Author: Md Mofijul Islam

This Dissertation has been read and approved by the examing committee:

Advisor: Tariq Iqbal

Advisor:

Committee Member: Laura Barnes

Committee Member: Aidong Zhang

Committee Member: Sara Riggs

Committee Member: Dan Bohus

Committee Member:

Committee Member:

Accepted for the School of Engineering and Applied Science:

Jennifer L. West, School of Engineering and Applied Science

August 2023

ABSTRACT

Humans inherently use multimodal data, such as verbal utterances and nonverbal gestures, to interact with others in shared physical environments. To develop AI systems to seamlessly interact with humans, it is also essential to understand how people behave and interact in these environments. Understanding human behavior and interactions (verbal and nonverbal) is paramount to ensuring seamless interactions between human and AI systems. Most recent approaches to understanding human behavior and interactions use unimodal data, such as using only visual data or only verbal utterances. For example, the majority of existing works solely use verbal utterance to comprehend embodied interactions and visual data to perceive human behavior. However, relying on unimodal data leads to single-point failure and can not ensure a robust perception of human action and comprehension of human interactions. For example, visual occlusion or visual data from low light conditions can limit the model to accurately recognize activities. Similarly, a verbal utterance is insufficient to determine an object if a visual scene contains two identical objects. Therefore, a robust model of understanding human interactions needs to incorporate multimodal information.

Developing these multimodal models for multiple tasks of perceiving human-embodied interactions requires addressing several fundamental challenges. For example, extracting salient representations from missing and noisy data modalities. Fusing, aligning, and extracting complementary representations from multiple heterogeneous modalities is challenging due to the disparate feature distributions and feedforward learning architecture. Similarly, learning salient representations from multiple verbal and visual perspectives needs to be addressed to effectively comprehend multimodal embodied interactions with verbal and nonverbal gestures. In my Ph.D. research, I developed robust models to perceive human behavior and embodied interaction using multimodal data to address these challenges.

First, we have developed multimodal learning models to robustly perceive human actions using multimodal sensor data, such as visual, depth, skeleton, and physical sensor data. These models can extract salient and complementary representations from heterogeneous modalities. Moreover, our proposed models can prioritize the modalities and extract salient representations from missing and noisy sensor modalities, whereas prior models could not effectively extract salient representations from heterogeneous and noisy sensor data. Additionally, we have developed a novel cooperative multitask learning model that can help to extract complementary multimodal representations using auxiliary information. Our extensive experimental results suggest that our proposed multimodal learning models outperform state-of-the-art models in recognizing human actions.

Second, comprehending embodied interaction can be studied by designing several fundamental tasks, such as understanding referring expression, comprehending embodied question answering, and determining perspective in an interaction. As collecting real-world data is costly and the existing simulator could not generate human multimodal interactions (verbal and nonverbal gestures), we have developed an embodied simulator, which we can use to generate synthetic multimodal human interactions and datasets. We can use these generated datasets to train and diagnose models for comprehending interactions with verbal utterances and nonverbal gestures.

The existing models of comprehending human interactions are designed to understand only verbal interaction from a single perspective and use the visual scene as context. However, people use multiple verbal and visual perspectives in real-world interactions (speaker and observer perspectives). Moreover, our experimental analysis suggests that perspective awareness in the learning models is crucial to comprehend embodied interactions. We have developed a perspective-aware learning model to understand human instructions with verbal utterances and non-verbal gestures. Our experimental analysis suggests that our proposed model can effectively extract salient multimodal representations to comprehend embodied interactions.

Additionally, the existing models and datasets of visual question answering use the visual scene as a context to answer verbal questions. However, humans use multimodal expressions (verbal utterances and nonverbal gestures) to ask questions in real-world settings. We have developed an embodied question answering (EQA) dataset and designed new tasks to develop models for comprehending question-answering interactions in embodied settings. As the existing models are designed to answer verbal questions, these models are less suitable for comprehending EQA tasks. We have developed learning models to extract aligned representations from multiple verbal and visual perspectives to answer questions with multimodal expressions.

While we can generate diverse synthetic interactions using our simulators, these interactions may differ from real-world human interactions, such as variations in pointing gesture and eye gaze patterns, different camera angles, object arrangements, and diverse environments. Thus, we have curated a large-scale embodied interaction dataset with multimodal data (verbal utterances and nonverbal gestures) in real-world settings. We have evaluated baseline multimodal learning models on this real-world dataset. The existing multimodal model aligns multiple representations and thus loses information across modalities. To address this challenge, we have proposed a reinforced residual representation-based multimodal learning model for extracting robust multimodal representations to comprehend human interactions in real-world settings. Our experimental results suggest that our proposed model with guided attention-based reinforced residual representation outperforms the baseline visual-language models in various challenging evaluation settings.

Our multimodal learning models and datasets can help to develop and evaluate models for various tasks, such as embodied question answering, visual-language-based navigation, and human-robot interactions. These models can be extended to improve human interactions in both virtual and real-world settings, including providing improved user experiences for people with disabilities. Furthermore, these models can enhance the user experience of AI assistants, like Amazon Alexa and Microsoft Cortana, strengthening their applications in online shopping, video gaming, and personalized online learning. Lastly, the findings from our works, proposed models, benchmarks, datasets, and embodied simulators can serve as a valuable tool for the research community, fostering the development and evaluation of learning multimodal models for Human-AI Interaction systems.

ACKNOWLEDGEMENTS

My utmost appreciation and respect are extended to my advisor, Professor Tariq Iqbal. Working under his guidance has been a transformative experience that I will cherish. Professor Iqbal did not just supervise my work; he mentored me, fostering an environment of freedom where I could navigate the complexities of my chosen research areas. He allowed me to explore my intellectual curiosities, challenging me to engage with the work that I found the most intriguing. His mentorship went beyond the confines of academics, always inspiring me to strive for excellence in all aspects of life. His firm belief in my capabilities and the gentle nudges towards exploring uncharted territories have immensely enriched my academic journey. His depth of knowledge, keen insights, and intellectual rigor have set a high benchmark for me. I am eternally grateful to him for this invaluable experience and his unwavering support throughout my research endeavor. Professor Iqbal's mentorship has been a cornerstone of my doctoral journey, shaping me into the researcher I am today.

The contributions of my committee members, Laura Barnes (Chair, Professor, Engineering Systems and Environment), Aidong Zhang (Professor, Computer Science), Sara Riggs (Associate Professor, Engineering Systems and Environment), and Dan Bohus (Senior Principal Researcher, Microsoft Research), were pivotal to the success of my thesis. Their expert advice, constructive criticisms, and insightful feedback have significantly shaped the work presented in this thesis.

I owe an immense debt of gratitude to my family. My parents, along with my sister, have made countless sacrifices to ensure that I could focus on my academic journey. My mother, in particular, has devoted her life to fostering my growth and success. The unwavering support of my family has been instrumental in my journey. Their sacrifices, love, and encouragement have been the bedrock upon which my academic pursuits have been built. My mother, with her exceptional strength and selflessness, sacrificed her own ambitions so that I could focus entirely on my studies. Her faith in my potential has been a beacon of hope, guiding me through the most challenging times. My uncles, Masud and Mamun, aunts, and grandparents have shown me enormous support and have played an integral part in my academic growth. Their wisdom, guidance, and steadfast encouragement have been invaluable in navigating the complex and often daunting world of academia. Their collective sacrifices and unconditional love have fuelled my motivation and determination, giving me the strength to persevere through the demanding process of completing my Ph.D. journey. I am forever indebted to them for their contribution to my achievements.

My heart brims with gratitude for my wife, whose sacrifice and support are beyond measure. She gave up her own flourishing career in Bangladesh to accompany me to the United States, a testament to her profound love and commitment. The sacrifices she made, and the tireless work she put into maintaining our home all aimed to create an environment where I could concentrate solely on my research. She is, without a doubt, the bedrock upon which much of my success stands. Not only has she been my life partner, but she also unwittingly took on the role of a co-advisor. Her perspective, despite her lack of domain expertise, brought a unique freshness to my research. Our countless brainstorming sessions, where she listened to my ideas and contributed her

own, often raised intriguing questions. Tackling these queries not only enriched my understanding but also solidified my projects. Her faith in my capabilities, her constant push for me to pursue more rigorous and qualitative work, and her unwavering encouragement during my darkest hours kept me motivated and focused. Her role in my academic journey, though invisible to many, has been monumental, and I am profoundly grateful for her immeasurable support.

My journey would have been far less rewarding without the continuous support of my friends. Their constant encouragement and faith in my abilities have been my source of strength and resilience. Among them, Anik, Swakkhar, Rana, Sanjay, Tanveer, Naim, Safat, Swapnil, Sabbir, Tonni, Zakaria, Sirat, Masum, Ridwan, Faysal, Arup, Novia, Kamrul, Tasnim, and Jahidul deserve special mention. Their kindness, understanding, and constant support helped me endure the most challenging phases of my doctoral journey. While they each brought unique perspectives and enriching companionship, collectively, they created an environment of solidarity and motivation that was instrumental to my success. I also extend my gratitude to the many other friends not mentioned here, each of whom has contributed to my journey in their own significant ways.

My lab mates— Samin, Haley, Alexi, Sujan, Reza, Shahid, Ganesh, Shahira, Srikar, and Keyan— have been integral to my research pursuits. The collaborative spirit, mutual support, and intellectual exchange within our lab significantly enhanced my academic journey. The countless discussions, brainstorming sessions, and collaborative problem-solving episodes we shared provided invaluable learning opportunities and a solid foundation for my research projects. Their constructive criticism, fresh perspectives, and relentless motivation have immensely contributed to my growth as a researcher. The journey would have been much harder and certainly less enjoyable without their camaraderie and support. I am profoundly thankful for their contributions to my academic achievements.

In the deepest chambers of my heart, a special place is reserved for my beloved Uncle, Tanvir Ahamed Mamu (I called him Boro Mama). His recent departure from this world on May 13, 2023, left a void that no accomplishment or accolade can fill. The thought of his absence during my Ph.D. defense is a pang that never ceases to stir a profound sense of loss within me. I had envisioned him by my side, sharing the joy of this milestone, echoing my triumph with his warm and proud smile. His departure, however, has left an emptiness that my success, ironically, has amplified. He was the first person who inspired me to delve into the intriguing world of computation. His enthusiasm for my aspirations, his undying belief in my abilities, and his unwavering support have been my guiding lights throughout this journey. In his memory, I carry forth the passion he instilled in me and the resilience he exemplified. While the grief of his absence weighs heavy, I am consoled by the knowledge that my accomplishments would have filled him with pride. Boro Mama, your influence and love remain indelible in my heart.

From the bottom of my heart, I extend my gratitude to everyone who has supported and contributed to my academic journey. This thesis is not just a reflection of my work but the combined efforts and support of you all.

DEDICATION

This thesis is wholeheartedly dedicated to my beloved parents, whose love, guidance, and sacrifices have lit the path of my academic journey. Their unwavering support and faith in me have been my greatest source of strength and inspiration.

And to my late Uncle, Tanvir Ahamed Mamun (Boro Mama), whose memory remains an indomitable beacon of motivation and determination. Even in his absence, his belief in my potential continues to push me toward greater heights. This work stands as a tribute to his enduring influence on my life and career.

To them, I dedicate each discovery and accomplishment encapsulated in this thesis. They are the bedrock upon which my academic achievements rest.

Table of Contents

List of Figures

List of Tables

Chapter 1
INTRODUCTION

Humans and most animals utilize a multisensory system (visual, non-visual, and somatosensory) to discern events, perceive actions, and comprehend interactions [1]–[4]. Multisensory systems provide complementary stimuli, which allows for a more holistic perception [5], [6]. Similar intuition guides the multimodal machine learning community to extract comprehensive feature representations for improving various applications, such as activity recognition [7]–[11], gesture recognition [12]–[14], affective-states recognition [15], video classification [16], [17], image captioning [18], [19], referring expression comprehension [20]–[25], and visual question answering [26], [27].

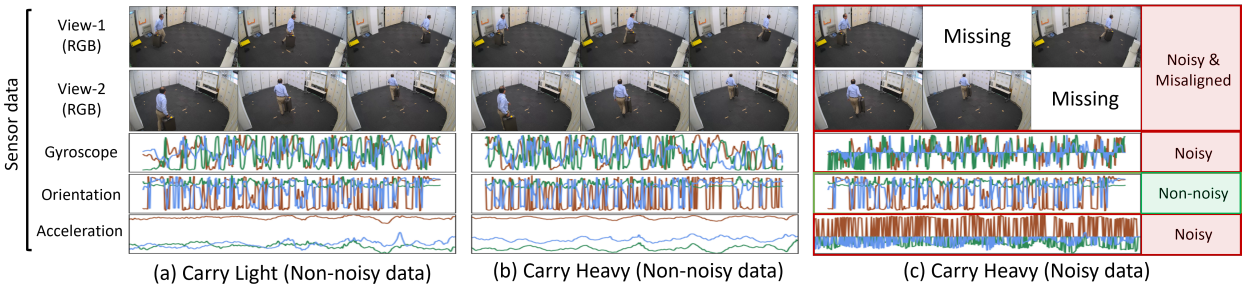


Figure 1.1: (a) Carry-Light and (b) Carry-Heavy activities have similar visual features. (a & b) However, these activities have distinct gyroscope and acceleration data. Prioritizing salient modalities (Gyroscope and Acceleration, in this case) while extracting complementary multimodal representations can help to recognize activities accurately. (c) Similarly, extracting complementary representation from noisy and non-noisy sensor data modalities can help models to robustly recognize activities. (Data samples are drawn from MMAct dataset [8]).

Moreover, multimodal data can help to develop robust models for human-AI interaction systems, such as humans interacting with robots to handover a particular object and humans interacting with an AI assistant to buy a product (e.g., Alexa, Cortana, Google Home, and Siri) [28]. Two fundamental components of these human-AI interactions (HAI) systems are to perceive human behavior (i.e., what is human doing?) and comprehend human instructions (i.e., what is human instructing?). These components enable to development of robust HAI systems and effectively improve the user experience and the usability of the HAI systems in the embodied settings.

First, perceiving human behaviors enables autonomous systems to understand human actions and act accordingly to ensure seamless interactions. Although modern autonomous systems, such as robots, are equipped with various sensors, robust human activity recognition (HAR) remains a fundamental challenge [29]. This is partly because fusing multimodal sensor data efficiently for HAR is challenging. Therefore, many researchers have focused on recognizing human activities by

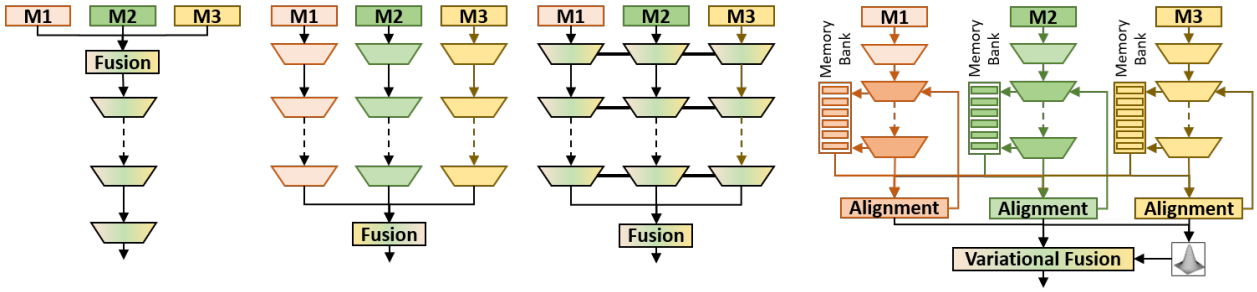
(a) Multimodal interaction (b) Human-AI assistant interaction using multimodal context

Figure 1.2: Multimodal communication forms, such as verbal utterances and non-verbal gestures (gaze and pointing gestures), ensures seamless human-AI interaction to improve user experience. (a) Human is instructing a robot using verbal instruction (*Pick up the small mug*), pointing gestures, and gaze. As the environment contains two small mugs, the robot can use pointing gestures to localize the appropriate objects. (b) A user is trying to find a product with the help of an AI assistant (e.g., Google Home). The AI assistant may need to utilize multimodal context (verbal and visual) to understand the user instruction and assist accordingly.

leveraging on a single modality, such as visual, pose or wearable sensors [30]–[34]. However, HAR models reliant on unimodal data often suffer a single-point feature representation failure. For example, visual occlusion, poor lighting, shadows, or complex background can adversely affect only visual sensor-based HAR methods. Similarly, noisy data modalities can reduce the performance of HAR methods solely depending on these sensors [9], [35].

Several approaches have been proposed to overcome the weaknesses of the unimodal methods by fusing multimodal sensor data that can provide complementary information to achieve a robust HAR [9], [35]–[39]. Distinct activities can be mistakenly classified as the same when relying on unimodal sensor data that provides similar information. For example, the activities related to carrying a light and a heavy object look similar from visual modalities; however, they have distinct physical sensor data (i.e., Gyroscope & Acceleration) (Fig. 1.1-a & b). Additionally, if a modality contains noisy data then the non-noisy data modalities can provide complementary information to accurately recognize activities using multimodal sensor data (Fig. 1.1-c). Thus, extracting salient and complementary representations from multiple data modalities can help develop a robust perception system to ensure seamless interactions with autonomous systems.

Second, humans inherently use multimodal communication forms in shared physical spaces, such as verbal and non-verbal (e.g., pointing gestures and gaze) modalities. Understanding multimodal human instructions (verbal and non-verbal) enables the development of situated human-AI interaction (HAI) systems (Fig. 1.2). To ensure seamless interactions between human and AI systems, we need to develop robust models for HAI systems to comprehensively understand embodied interactions with multimodal signals (verbal utterance and nonverbal gestures) in the shared physical space. Several tasks have been designed to comprehend multimodal interactions. For example,



(a) Early (b) Late (c) Intermediate (d) Recurrent Fusion

Figure 1.3: Comparison between various multimodal fusion approaches (Early, Late and Intermediate) and Recurrent Fusion (M1: Modality one, M2: Modality two, and M3: Modality three)

predicting the referred object's existence in the scene (scene understanding), identifying an object and its attributes (e.g., color), determining whether verbal utterance and nonverbal gestures refer to the same object (relation grounding), and comprehending from which perspective (e.g., speaker or observer) the object has been described (perspective grounding). These diverse sets of tasks with different complexity help to train robust models to comprehend multimodal instruction in embodied settings.

1.1 Challenges

Several fundamental challenges must be addressed to develop multimodal models for extracting salient and complementary representation from heterogeneous, missing, and noisy data modalities. Learning salient multimodal representations enables robust perceiving human actions and comprehension of human verbal and non-verbal interactions in embodied settings. I have planned to address these fundamental challenges in my dissertation. I described these challenges in this section.

1.1.1 Multimodal Representation Learning for Perceiving Human Behavior

Fusing Heterogeneous Multimodal Data: In multimodal representation learning, it is crucial to fuse relevant features from various heterogeneous modalities (visual, skeleton, and physical sensors) [9], [12], [28], [38], [40], [41]. Predominantly, multimodal fusion has been studied from three different perspectives: early, intermediate, and late fusion [28], [38], [42] (see Fig. 1.3). Early fusion combines unimodal raw features, which limits the capturing of distinct feature characteristics, as early-stage fusion loses the unimodal feature distribution [7], [26], [38]. Late fusion encodes modalities independently, but the absence of cross-modal interaction restricts each modality from obtaining multimodal context [7], [12], [16], [38]. Intermediate fusion allows cross-modal interactions to fuse mid-level features for extracting multimodal representation. In addition to its

advantages over other fusion approaches, intermediate fusion is also manifested in the neural multisensory system of animals [1], [43], [44].

Although several intermediate fusion approaches have been proposed [3], [7], [8], [28], [40], there remain crucial challenges in obtaining robust multimodal representation. Most importantly, heterogeneous modalities representing the same phenomenon may have disparate characteristics, making it challenging to align them. For example, visual modalities (RGB and Depth) have different feature characteristics and distributions than wearable sensor modalities (Acceleration, Gyroscope etc.). Similarly, fusing continuous representations from visual modalities to discrete representation from language-related modalities is challenge due to the disparate characteristic of data and unimodal feature encoders.

Learning Complementary Multimodal Representation: In real-world settings, some of the modalities may provide misaligned or noisy data, making it challenging to obtain a robust multimodal representation. One approach to resolve these issues is to align multimodal sensor data. In the literature, learning models have been proposed to align multimodal sensor data in two ways: explicit and implicit alignment [28]. Explicit alignment approaches temporally align the raw sensor data, whereas implicit alignment approaches align intermediate feature representations. However, explicit alignment of raw sensor data in temporal space cannot ensure alignment in representation space due to the heterogeneity of the modalities, which may lead to sup-optimal representations [26], [28]. Temporal alignments may not mitigate the impact of noise in the input data.

On the other hand, in implicit alignment approaches, one modality aligns representation independently without knowing the representations from other modalities. Thus, these alignment approaches can lead to sub-optimal multimodal representations. Moreover, state-of-the-art implicit alignment models have predominantly been used deterministic attention methods to extract salient representations from noisy heterogeneous sensor data [10], [11]. A deterministic attention approach learns a point estimate for the attention weights, limiting the multimodal learning model from aligning features and modeling uncertainty which are particularly crucial for extracting features from noisy data. However, if we can model the attention weights as a variational distribution, it can help to implicitly align unimodal representations by imposing the same prior distribution over the attention weights. Furthermore, learning an attention distribution allows the multimodal learning approaches to model uncertainty when fusing the unimodal representations and guides the unimodal learning models to extract salient representations.

Aligning and Refining Multimodal Representations: Moreover, state-of-the-art multimodal learning approaches combine the information in a feed-forward manner, restricting each modality from aligning and refining representation. In these approaches, unimodal feature encoders extract representations independently without observing feature representations from other modalities. However, if these encoders have information about other modalities, they can utilize that to *align* and iteratively *refine* the unimodal features to generate robust representations. This alignment approach can iteratively change (refine) the unimodal representations to attain aligned distributions (Fig. 1.4), which can lead to robust multimodal representations. For example, the representations

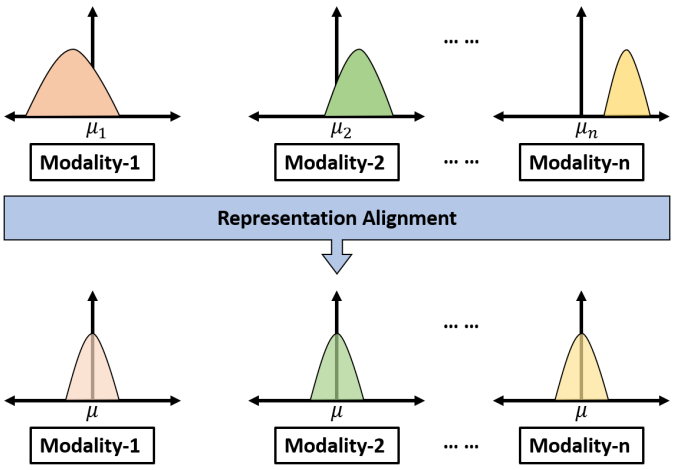


Figure 1.4: Unimodal representation distributions before (top) and after (bottom) the representation alignment process. Representation alignments process helps to attain similar distributions for all modalities to produce robust multimodal representations.

from noisy modalities can be iteratively refined to align their distributions to non-noisy modalities' distributions, which may lead the model to generate robust multimodal representations. Moreover, many state-of-the-art feed-forward fusion approaches [7], [16] utilized short-range spatial-temporal features, which hinder the performance in several learning tasks, such as activity recognition, that requires capturing long-range temporal multimodal features.

1.1.2 Comprehending Embodied Interactions using Multimodal Data

Comprehending Multimodal Embodied Interactions: Comprehending referring expressions has been generally studied in the form of the *spatial relation grounding task* [20], [21], [26], [27], [45]–[54]. This task involves identifying whether the verbal utterance of the spatial relationships between objects holds in a visual scene [20], [21]. However, the exclusion of nonverbal signals in the model makes the problem different from how people interact naturally in shared physical spaces [55]–[63].

Developing learning models to comprehend multimodal referring expressions requires a large and diverse dataset, which is time-consuming, laborious, and costly to curate in real-world settings. A few datasets have been developed to capture embodied multimodal referring expressions, which involve referring to an object using verbal utterances and nonverbal cues (pointing gesture and gaze) [24], [64]. However, these datasets have several crucial limitations. The primary limitation of existing datasets is that the nonverbal interactions are captured solely from an exocentric perspective (*exo, ego, and top* view denotes perspectives from an actor, the observer, and overhead, respectively). As comprehending embodied referring expression requires perspective-

taking, which is the awareness of the actor's and observer's point of view in shared interactions, the lack of perspective-awareness in these datasets can degrade the model's performance. Additionally, multiple views can help identify the referred object, which may be partially occluded from one view but visible from another. Moreover, in human-human interactions, learning perspective is used innately to attend to salient parts of interactions. Let's assume an actor is requesting an observer verbally to "pick up the left apple". This verbal expression can be interpreted differently from different perspectives, where the "left apple" from the exo view can be interpreted as the "right apple" from the ego view. Learning where the actor is looking and pointing can help identify the appropriate object in these scenarios. These data samples with multiple views enable the model to learn perspective-taking to ensure seamless and natural interactions in embodied settings.

Comprehending Perspective-aware Multimodal Embodied Interactions: Several recent works have attempted to address the task of comprehending referring expressions by incorporating nonverbal gestures with verbal utterances in embodied settings (known as embodied referring expression comprehension (E-REF)) [24], [64]. However, some crucial issues remain unaddressed in these recent works. Particularly, most embodied referring expression datasets only capture human interactions from an observer perspective with exo-centric views. People innately use an understanding of perspective, which can be observed in how humans interchangeably use perspectives from the speaker and the observer when referring to objects during interactions. The existing models learn to comprehend referring expressions using single verbal and visual perspectives.

Recent works studied REF and E-REF by designing two separate tasks: a relation grounding task [21], [65]–[67] and an object grounding task [20], [24], [68], [69]. In a non-embodied setting, the relation grounding task is defined as determining whether a verbal utterance appropriately describes the spatial relationships between objects in a visual scene. In an embodied setting, this relation grounding task is defined as determining whether a verbal utterance and nonverbal signals (gazes and pointing gestures) refer to the same object. The object grounding task aims to identify a referred object using a verbal utterance and nonverbal gestures. These tasks have many use-cases in real-world interactions. For example, if a person verbally describes an object but nonverbally points to another object, an AI-driven agent should identify these incoherent multimodal cues, using the relation grounding task, and request clarification. In another case, if a person points to an object and asks, "what is the object to the right of the black hat?", then the AI agent can use the object grounding task to identify the referred object. Thus, training models on these two related tasks (relations and objects grounding) and the previously mentioned perspective grounding task can enable achieve seamless human-AI interactions (HAI).

Embodied Question Answering (EQA): In the litterateur, EQA is designed in two ways. First, *embodied interactions* are defined from an agent's perspective, such as a virtual robot, where the agent perceives its environment and navigates it to answer verbal questions [70]. Second, embodied interaction refers to multimodal human expressions, where a human interacts with the environment using verbal utterances and nonverbal gestures [24], [25], [71]. Adopting the later definition, we define EQA tasks as questioning using multimodal human expressions (verbal and nonverbal gestures) in embodied settings. For example, an EQA task can involve pointing to an

object and asking "what is that object?".

We need to have a large and diverse dataset to develop robust models for EQA. Several synthetic and real-world datasets have been developed for Visual Question Answering (VQA) [22], [23], [72]–[74] and Embodied Question Answering (EQA) [70], [75]–[78]. One of the crucial drawbacks of these datasets is that these datasets have been developed in non-embodied settings, where the visual scene does not contain humans, and thus nonverbal gestures (gaze and pointing gestures) are not used for asking a question. A few embodied question answering datasets have been developed where an agent (e.g., virtual robot) finds the answer to a verbal question from the environment. However, in these datasets, humans and nonverbal gestures are not involved in the interactions. In real-world settings, humans predominantly use nonverbal gestures with verbal utterances to interact with others. Additionally, ample evolutionary evidence indicates that nonverbal gestures have been predominantly used in real-world settings compared to verbal communication forms [55]–[63], [79], [80].

Moreover, nonverbal gestures provide complementary information to seamlessly understanding a verbal utterance. Suppose a visual scene contains two balls with different colors, and we verbally ask a person to find the object's color. In that case, nonverbal signals can help disambiguate this interaction and assist that person in seamlessly determining the referred object's color. Similarly, nonverbal gestures can provide complementary information to the model for comprehending embodied interactions with multimodal signals (verbal utterance and nonverbal gestures). Thus, the lack of nonverbal interactions in the VQA and EQA datasets makes these datasets less suitable for developing and evaluating models for comprehending question-answering (QA) related interactions with multimodal signals.

1.2 Thesis Statement

Multimodal data, such as visual, verbal, and physical sensors, provide complementary information enabling autonomous systems to accurately perceive human actions and interactions in shared physical settings. While human's natural communication form involves verbal and nonverbal gestures, state-of-the-art models and datasets solely consider verbal utterances to comprehend human interactions in the shared physical system. Moreover, developing robust models to extract complementary representations from multimodal sensor data for multiple tasks is challenging due to the heterogeneous multimodal feature distributions and missing and noisy data modalities. Allowing the models to capture inter-modality interactions, recurrently fusing multimodal information, and cooperatively training related tasks enables the multimodal machine learning models to learn salient and complementary multimodal representations and improve the performance of multiple tasks. These models can ensure robust perception for autonomous systems and enable seamless interactions with humans.

1.3 Completed Work

1.3.1 Multimodal Representation Learning

To extract salient and complementary multimodal representations from heterogeneous multimodal data, we have developed several multimodal machine models [7], [10], [11], [15], [81] and conducted extensive experimental analysis on multimodal human activity datasets.

- We have developed a multimodal human activity recognition algorithm, called HAMLET: Hierarchical Multimodal Self-attention human-activity recognition algorithm [7]. HAMLET first extracts the spatial-temporal salient features from the unimodal data for each modality. HAMLET then employs a novel multimodal attention mechanism, called HAT, for disentangling and fusing the unimodal features. These fused multimodal features enable HAMLET to achieve higher human-activity recognition accuracy. The modular approach to extract spatial-temporal salient features from unimodal data allows HAMLET to incorporate pre-trained feature encoders for some modalities, such as pre-trained ImageNet models for RGB and depth modalities. This flexibility enables HAMLET to incorporate deep neural network-based transfer learning approaches. Additionally, the proposed novel multimodal fusion approach (MAT) utilizes a multi-head self-attention mechanism, which allows HAMLET to be robust in learning weights of different modalities based on their relative importance in human-activity recognition from data.

- We have developed a multimodal feature learning method for human-activity recognition, called Multi-GAT (Graphical Attention-based Hierarchical Multimodal Representation Learning Approach) [10]. Multi-GAT first extracts modality-specific salient spatial-temporal features by utilizing a unimodal attention approach. Multi-GAT then employs a multimodal mixture-of-experts model, called Multi-MoE, to disentangle and extract salient unimodal features. The Cross-GAT module, a novel message-passing based multimodal graphical attention approach, enables inter-modality feature interaction while generating complementary multimodal features. Most importantly, Cross-GAT captures cross-modal relationships to extract robust features, which aids a robot in recognizing human activities accurately. Finally, a task learning network uses the multimodal features for human-activity recognition.

- We have developed a novel Memory-Augmented Variational Attention-based Multimodal Representation learning approach, called MAVEN [81]. MAVEN aims to learn a complementary multimodal representation from heterogeneous modalities to perceive human activities accurately. To extract complementary multimodal representation, MAVEN first incorporates feature encoders to produce modality-specific spatial features. These features are populated into memory banks that are used to capture long-term spatial-temporal feature relationships. MAVEN then employs a novel Recurrent Memory-Augmented Attention-based Feature Alignment Approach (ReMATE) that iteratively refines and aligns unimodal features

by observing memory banks from other modalities. To the best of our knowledge, we are the first to propose a memory-augmented recurrent feature alignment approach for multimodal fusion. Finally, we introduce a Variational Attention-based fusion approach (VAT), which fuses the unimodal features to produce a robust multimodal representation. We incorporate a variational inference loss that helps to align unimodal features in the multimodal feature space. The final multimodal representation is then used for a given learning task (i.e, activity recognition).

1.3.2 Multitask Learning-based Guided Multimodal Representation Learning

In the literature, several multitask learning models have been proposed, which have shown promising results in learning shared representations across different tasks [82]–[86]. For example, Liu et al. [87] proposed a multitask attention model for learning task-aware shared representations. Moreover, Sun et al. [88] designed an algorithm to learn feature sharing patterns across tasks for maximizing shared representations. The overall goal of these approaches is to compress a multitask model by maximizing the shared representations among the competitive tasks.

Despite these advancements, there is still a need for more effective and efficient methods for multimodal data fusion and activity recognition using multitask models. To address this gap, we propose a novel Cooperative Multitask Learning-based Guided Multimodal Fusion Approach (MuMu) for human-activity recognition [11]. In MuMu, we have designed a multitask learning approach that involves learning two cooperative tasks: an auxiliary and a target task. First, MuMu extracts activity-group-specific features for activity-group recognition (auxiliary task). Second, the activity-group-specific features direct our Guided Multimodal Fusion Approach (GM-Fusion) to extract robust multimodal representations for recognizing activities (target task). Here, both tasks work cooperatively, where the auxiliary task guides the target task to extract complementary multimodal representations appropriately. The unique aspect of our approach is the cooperative nature of multiple-task learning, where the auxiliary task guides the target task to extract complementary multimodal representations appropriately.

1.3.3 Comprehending Embodied Referring Expressions

We have developed a novel embodied simulator, CAESAR, to generate large-scale datasets of referring expressions. To the best of our knowledge, CAESAR is the first simulator to generate multimodal referring expressions with verbal utterances and nonverbal gestures in a virtual environment. CAESAR has three novel aspects which differentiate it from other synthetic data generation systems (e.g., CLEVR [23] and Kubric [89]). First, CAESAR simulates scenarios in which verbal utterances and nonverbal cues (pointing gesture and gaze) refer to objects in an embodied setting. We have collected real human pointing gesture data using an OptiTrack motion capture system [90] and emulated the same behaviors in CAESAR by incorporating a new stochastic deictic gesture generation approach. Second, CAESAR renders multiple views from different perspectives,

such as ego-, exo-, and top-view, that can aid in training models to learn different perspectives for comprehending multimodal referring expressions. Third, taking inspiration from previous work [21], we have designed a module in CAESAR to generate contrastive samples, where the virtual human is pointing to an object while verbally describing a different object.

One of the primary goals of developing CAESAR is to democratize the data generation process so that researchers without simulator development experience can have complete control of generating a diverse dataset to train and evaluate a learning model. Similar to existing data generation systems, the development of our simulator requires extensive knowledge of motion planning and game engine. Thus, to make it accessible to everyone, we have developed a tool that enables researchers to generate diverse samples without any simulator development experience. Using this tool, we have developed two large-scale datasets, CAESAR-XL and CAESAR-L, for understanding multimodal referring expression in an embodied virtual environment.

Perspective-aware Embodied Referring Expression Comprehension: we have developed a novel perspective-aware multitask model, PATRON, for the relation and object grounding task using multimodal cues. In PATRON, we have designed two cooperative tasks, one for the perspective grounding (the auxiliary task) and another for the relation and object grounding (the target task). In the auxiliary task module, PATRON learns disentangled representations, the auxiliary task-specific and task-guidance representations, to learn perspective grounding. In the target task module, PATRON uses our proposed guided fusion approach that utilizes task-guidance representations from the auxiliary task as prior information to extract guided multimodal representations. PATRON uses a self-attention-based fusion approach to extract supplementary target task-specific representations. Finally, PATRON fuses task-guided and target task-specific disentangled representations to learn relation and object grounding.

1.3.4 Comprehending Embodied Question Answering

We have extended our embodied simulator (CAESAR) to develop a novel EQA dataset, EQA-HuMu, for training and diagnosing models in comprehending EQA. We can use our extended simulator to procedurally generate nonverbal interactions (gaze and pointing gestures) and verbal utterances in multiple embodied environments for different EQA tasks. EQA-HuMu has four novel contributions over existing EQA and VQA datasets. First, to the best of our knowledge, we are the first to use nonverbal gestures (gaze and pointing gestures) and verbal utterances to formulate a question that needs to be answered using the visual context in the embodied environment. Second, we have included multiple perspectives in the verbal utterances, which can aid in developing robust models. Third, we have captured the nonverbal interactions in the embodied setting using multiple views to reduce the model's verbal and visual perspective bias. Finally, we have designed eight new EQA tasks to appropriately understand embodied interaction with multimodal signals (verbal utterance and nonverbal gestures).

Several visual-language representation learning models have been developed for VQA tasks [26], [27], [91], [92]. Although these models work adequately for VQA tasks, these models were

designed to learn from a single visual and verbal perspective. I have developed a multimodal learning model that can align representations from multiple visual perspectives using vector quantization (VQ) to learn a unified concept. Additionally, VQ can disentangle the visual representations and enable the fusion of discrete verbal representations to produce salient multimodal representations. Moreover, this model uses a vector quantization approach to disentangle and align multiview representations. These representations are used to learn EQA tasks. We have developed baseline models by extending the existing visual-language models. We have evaluated our proposed model and baseline models on our EQA dataset to compare the performance of EQA tasks.

1.3.5 Comprehending Embodied Interactions in Real-World Environments

We have developed a diverse dataset, REMO, to comprehend human interaction in real-world settings. We have collected the dataset in diverse indoor and outdoor settings with varying environment attributes, such as lighting conditions, object arrangement, and environment appearance. We have used Azure Kinect DK devices mounted on a mobile robot to capture the embodied interactions from different angles. We installed this device on Ohmni Robotic system to capture the ego view from the robot. Moreover, we have collected gaze and ego views from a human perspective using PupilCore SmartGlass. We have collected data in two settings: constrained and unconstrained settings. In constrained settings, we explicitly give detailed instructions to the participants to use gaze, pointing gestures, and verbal utterances to describe an object. However, In unconstrained settings, we did not instruct the participants to describe an object.

Finally, we have developed a robust model to comprehend human interactions in real-world settings using the insights from training models on synthetic datasets. We have designed a language-guided multimodal representation learning model to extract salient representation for comprehending referring expressions from multiple perspectives. We have trained this model on our real-world dataset and conducted an extensive experimental analysis to investigate whether verbal and nonverbal gestures can help to comprehend embodied interactions. Finally, we have investigated whether the findings from training models on synthetic datasets align with those from training models on real-world datasets. These experimental analyses give valuable insights into developing models for comprehending embodied interactions in real-world settings.

1.4 Contributions

My dissertation has two main components: perceiving human behavior and comprehending embodied interactions using multimodal data (Fig. 1.5). First, we have developed several multimodal representation learning models addressing the issues of perceiving human behavior robustly. For example, fusing heterogeneous multimodal data (IROS-2020 [7]), learning complementary multimodal representations (IEEE RAL-2021 [10] and AAAI-2022 [11]), recurrently fusing multimodal representations (IEEE Transaction Multimedia-2022), and cooperative multitask based multimodal representation learning for robustly perceive human actions (AAAI-2023 [11]). Second, we have

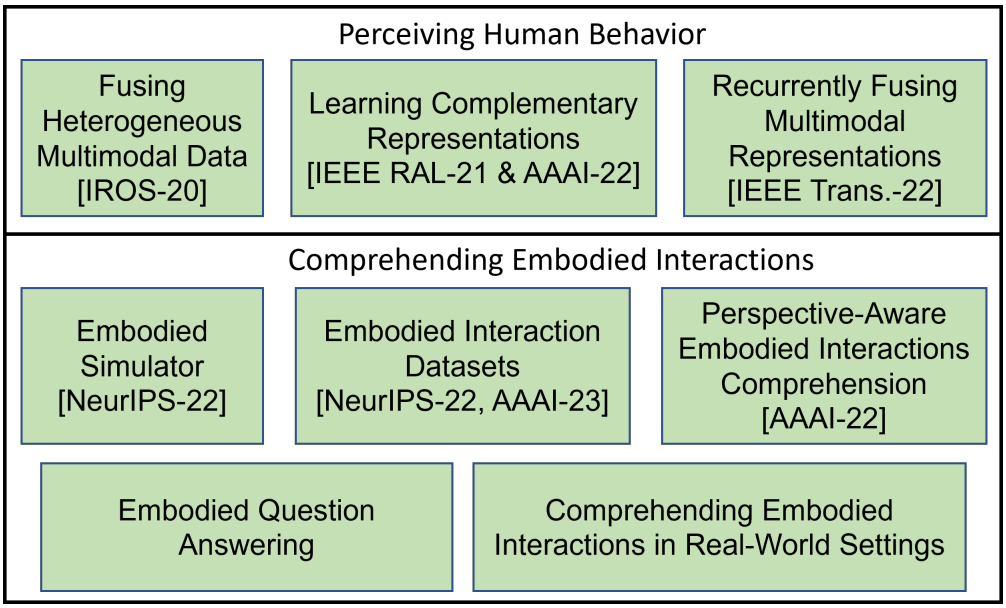


Figure 1.5: dissertation progress in addressing the challenges of perceiving human behavior and comprehending embodied interactions using multimodal data.

developed a simulator (NeurIPS-2022 [25]), datasets (NeurIPS-2022 [25] and AAAI-2023 [71]), benchmark models (NeurIPS-2022 [25]), and perspective-aware multimodal multitask representation learning model (AAAI-2023 [Under Review] [71]) to comprehend embodied interactions using multimodal signals, such as verbal utterances and nonverbal gestures (gaze and pointing gestures). Based on these works, we have developed an embodied question-answering model and datasets to ensure seamless interactions in embodied settings. In this work, we address the challenge of fusing continuous visual and discrete language representations by discretizing the visual representations using a vector-quantization approach. Moreover, We have developed embodied interaction datasets to train and evaluate models for comprehending human embodied interactions in real-world settings. Finally, we have proposed a novel reinforced residual representation-based multimodal model to comprehend embodied referring expressions from multiple perspectives.

1.5 Broader Impact

Our proposed multimodal learning models and datasets have a wide range of applicability in different domains, such as embodied question answering, visual-language-based navigation in embodied settings, and human-robot interactions. Our models can be extended to understand human verbal and nonverbal interactions in virtual and real-world settings. These models will also help to improve human interactions in virtual settings. Additionally, our developed multimodal learning models can be extended to understand the instructions from people with disabilities to ensure

improved user experience and the usability of the assistance systems. As people with disabilities use various combinations of modalities for interactions and communications, our instruction understanding approach utilizing multiple modalities (visual, audio, language, and gesture) is crucial to ensure effective assistance in the healthcare and home environment.

Additionally, our proposed learning framework can aid AI assistants (e.g., Amazon Alexa, Microsoft Cortana, Google Home, and Apple Siri) with multimodal interactions (voice and visual) in improving the user experience for various applications, such as online shopping assistants, video gaming, and personalized online learning assistants for students. For example, in the online shopping platform, multimodal human instruction and product content understanding can ensure seamless user interaction and thus strengthen product recommendations. Similarly, the online educational platform can employ the content understanding approach to gauge the students' engagement and provide personalized learning recommendations. Finally, we believe our proposed multimodal human interaction simulator can help the research community to develop and evaluate learning models for HAI systems and move this research field forward.

1.6 Publications

The following works are from my Ph.D. tenure [August 2019 - Present].

1.6.1 Journal Publications

(Peer Reviewed) [*: Equal Contribution]

5. M. S. Yasar *, **M. M. Islam***, T. Iqbal, "IMPRINT: Interactional Dynamics-aware Motion Prediction in Teams using Multimodal Context," *ACM Transactions on Human-Robot Interaction*.

4. **M. M. Islam**, M. S. Yasar, T. Iqbal, "MAVEN: A Memory Augmented Recurrent Approach for Multimodal Fusion," IEEE Transaction on Multimedia, 2022.

3. S. Samyoun*, **M. M. Islam***, T. Iqbal, J. Stankovic, "M3Sense: Affect-Agnostic Multitask Representation Learning using Multimodal Wearable Sensors," ACM Interactive Mobile Wearable Ubiquitous Technology 2022.

2. M. Rahman, M. A. Alam, **M. M. Islam**, M. M. Khan, I. Rahman, T. Iqbal, "An Adaptive Agent Specific Sub-optimal Bounding Approach for Multi-Agent Path Finding," IEEE Access Journal, 2022.

1. **M. M. Islam**, T. Iqbal, "Multi-GAT: A Graphical Attention-based Hierarchical Multimodal Representation Learning Approach for Human Activity Recognition," IEEE Robotics and Automation Letters, 2021 (Oral presentation at IEEE International Conference on Robotics and Automation 2021).

1.6.2 Conference Publications

(Peer Reviewed)

13. **M. M. Islam**, A. Gladstone, R. Islam, T. Iqbal, "EQA-MX: Embodied Question Answering using Multimodal Human Expression," Under review.

12. **M. M. Islam**, A. Gladstone, T. Groves, T. Iqbal, "SDD: A Shape Guided Diffusion Model for Generating Depth".

11. **M. M. Islam**, G. Nanduru, A. Gladstone, S. Sarker, S. Gouru, K. Du, T. Iqbal, "COBRA: Comprehending Embodied Referring Expressions from Multiple Perspectives Using Language and Visual Cues".

10. M. F. R. M. Billah, **M. M. Islam**, N. Saoda, F. Nikseresht, T. Iqbal, B. Campbell, "Scanning for Sensors: Fusing Computer Vision and BLE Advertisement Signal for Accurate Sensor Localization in AR View".

9. **M. M. Islam**, A. Gladstone, T. Iqbal, "PATRON: Perspective-aware Multitask Model for Referring Expression Grounding using Embodied Multimodal Cues," Association for the Advancement of Artificial Intelligence (AAAI) 2023.

8. R. Islam, H. Zang, M. Tomar, A. Didolkar, **M. M. Islam**, A. Goyal, S. Y. Arnob, X. Li, T. Iqbal, N. Heess, A. Lamb, "Representation Learning in Deep RL via Discrete Information Bottleneck," Artificial Intelligence and Statistics *(AISTATS) 2023*.

7. **M. M. Islam**, R. Mirzaiee, A. Gladstone, H. Green, T. Iqbal, "CAESAR: An Embodied Simulator for Generating Multimodal Referring Expression Datasets," Neural Information Processing Systems (NeurIPS) 2022 (Track on Datasets and Benchmarks).

6. **M. M. Islam**, G. Aguilar, P. Ponnusamy, C. Solomon, C. Ma, C. Guo, "A Vocabulary-Free Multilingual Neural Tokenizer for End-to-End Task Learning," Association for Computational Linguistics Workshop on Representation Learning for Natural Language Processing 2022.

5. **M. M. Islam**, T. Iqbal, "MuMu: Cooperative Multitask Learning-based Guided Multimodal Fusion," Association for the Advancement of Artificial Intelligence (AAAI), 2022 [Main Track: Oral].

4. H. Green, **M. M. Islam**, S. Ali, T. Iqbal, "Who's Laughing NAO? Examining Perceptions of Failure in a Humorous Robot Partner," ACM/IEEE International Conference on Human-Robot Interaction 2022.

3. H. Green, **M. M. Islam**, S. Ali, T. Iqbal, "iSpy a Humorous Robot: Evaluating the Perceptions of Humor Types in a Robot Partner," Association for the Advancement of Artificial Intelligence (AAAI) Spring Symposium on Putting AI in the Critical Loop: Assured Trust and Autonomy in Human-Machine Teams, 2022.

2. H. N. Green, **M. M. Islam**, S. Ali, and T. Iqbal, "Perceiving a Humorous Robot as a Social Partner," Elsevier, 2022.

1. **M. M. Islam**, T. Iqbal, "HAMLET: A Hierarchical Multimodal Attention-based Human Activity Recognition Algorithm," IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020.

Chapter 2
MULTIMODAL REPRESENTATION LEARNING FOR PERCEIVING HUMAN BEHAVIOR

2.1 Attention-based Multimodal Representation Learning

2.1.1 Proposed Modular Learning Method

In this section, we present our proposed multimodal human-activity recognition method, called HAMLET: **H**ierarchic**a**l **M**ultimoda**l** **S**elf-a**t**tention based HAR. We present the overall architecture in Fig. 2.1. In HAMLET, the multimodal features are encoded into two steps, and those features are then used for activity recognition as follows:

- At first, the Unimodal Feature Encoder module encodes the spatial-temporal features for each modality by employing a modality-specific feature encoder and a multi-head self-attention mechanism (UAT).
- In the second step, the Multimodal Feature Fusion module (MAT) fuses the extracted unimodal features by applying our proposed novel multimodal self-attention method.
- These computed multimodal features are then utilized by a fully connected neural network to calculate the probability of each activity class.

2.1.1 Unimodal Feature Encoder

The first step of HAMLET is to compute a feature representation for data from every modality. To achieve that, we have designed modality-specific feature encoders to encode data from different modalities. The main reasoning behind this type of modality-specific modular feature encoder architecture is threefold. First, each of the modalities has different feature distribution and thus needs to have a different feature encoder architecture. For example, the distribution and representation of visual data differ from the skeleton and inertial sensor data. Second, the modular architecture allows incorporating unimodal feature encoders without interrupting the performance of the encoders of other modalities. This capability enables the modality-specific transfer learning. Thus we can employ a pre-trained feature encoder to produce robust feature representation for each modality. Third, the unimodal feature encoders can be trained and executed in parallel, which reduces the computation time during the training and inference phases.

Each of the unimodal feature encoders is divided into three separate sequential sub-modules: spatial feature encoder, temporal feature encoder, and unimodal attention module (UAT). Before applying a spatial feature encoder, at first the whole sequence of data $D^m = (d_1^m, d_2^m, ..., d_T^m)$ from modality m is converted into segmented sequence $X^m = (x_1^m, x_2^m, ..., x_{S^m}^m)$ of size $B \times S^m \times E^m$,

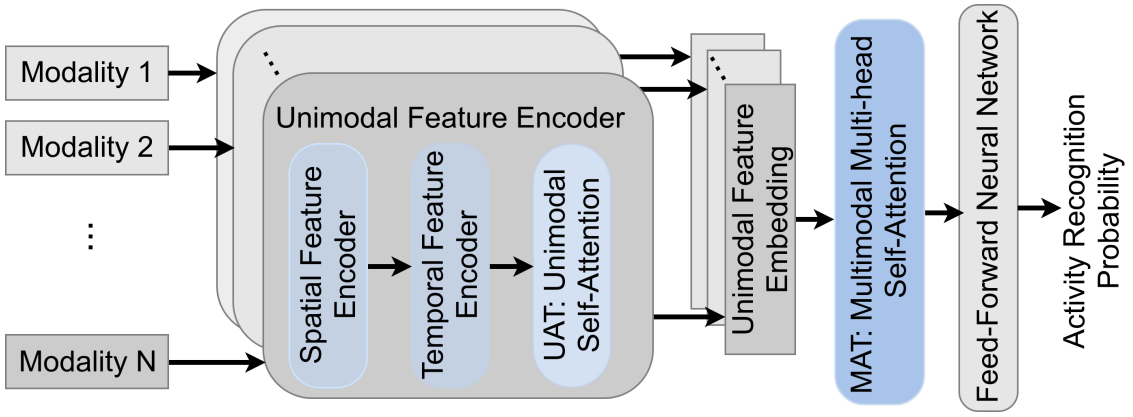


Figure 2.1: HAMLET: Hierarchical Multimodal Self-Attention based HAR.

where B is the batch size, S^m and E^m are the number of segments and feature dimension for modality m respectively. In this work, we represent the feature dimension E^m for RGB and depth modality as $(channel(C^m) \times height(H^m) \times width(W^m))$, where C^m is the number of channels in an image.

2.1.1.1.1 Spatial Feature Encoder We used a temporal pooling method to encode segment-level features instead of extracting the frame-level features, similar to [37]. We have implemented the temporal pooling for two reasons: first, as the successive frames represent similar features, it is redundant to apply spatial feature encoder on each frame, which increases the training and testing time. By Utilizing the temporal pooling, HAMLET reduces its computational time. Moreover, this polling approach is necessary to implement HAMLET on a real-time robotic system. Second, the application of recurrent neural networks for each frame is computationally expensive for a long sequence of data. We used adaptive temporal max-pool to pool the encoded segment level features.

As our proposed modular architecture allows modality-specific transfer learning, we have incorporated the available state-of-the-art pre-trained unimodal feature encoders. For example, we have incorporated ResNet50 to encode the RGB modality. We extend the convolutional co-occurrence feature learning method [93] to hierarchically encode segmented skeleton and inertial sensor data. In this work, we used two stacked 2D-CNNs architecture to encode co-occurrence features: first 2D-CNN encodes the intra-frame point-level information and second 2D-CNN extract the inter-frame features in a segment. Finally, spatial feature encoder for modality m produces a spatial feature representation F_m^S of size $(B \times S^m \times E^{S,m})$ from segmented X^m, where $E^{S,m}$ is the spatial feature embedding dimension.

2.1.1.1.2 Temporal Feature Encoder After encoding the segment level unimodal features, we employ recurrent neural networks, specifically unidirectional LSTM, to extract the temporal feature features $H^m = (h_1^m, h_2^m, ..., h_s^m)$ of size $(B \times S^m \times E^{H,m})$ from F_m^S, where $E^{H,m}$ is the LSTM

hidden feature dimension. Our choice of unidirectional LSTM over other recurrent neural network architectures (such as gated recurrent units) was based on the ability of LSTM units to capture long-term temporal relationships among the features. Besides, we need our model to detect human activities in real-time, which motivated our choice of unidirectional LSTMs over bi-directional LSTMs.

2.1.1.1.3 Unimodal Self-Attention Mechanism The spatial and temporal feature encoder sequentially encodes the long-range features. However, it cannot extract salient features by employing sparse attention to the different parts of the spatial-temporal feature sequence. Self-attention allows the feature encoder to pay attention to the sequential features sparsely and thus produce a robust unimodal feature encoding. Taking inspiration from the transformer-based multi-head self-attention methods [94], UAT combines the temporal sequential salient features for each modality. As each modality has its unique feature representation, the multi-head self-attention enables the UAT to disentangle and attend salient unimodal features.

To compute the attended modality-specific feature embedding F_m^a for modality m using unimodal multi-head self-attention method, at first we need to linearly project the spatial-temporal hidden feature embedding H^m to create query (Q_i^m), key (K_i^m) and value (V_i^m) for head i in the following way,

$$Q_i^m = H^m W_i^{Q,m} \tag{2.1}$$
$$K_i^m = H^m W_i^{K,m} \tag{2.2}$$
$$V_i^m = H^m W_i^{V,m} \tag{2.3}$$

Here, each modality m has its own projection parameters, $W_i^{Q,m} \in \mathbb{R}^{E^{H,m} \times E^K}$, $W_i^{K,m} \in \mathbb{R}^{E^{H,m} \times E^K}$, and $W_i^{V,m} \in \mathbb{R}^{E^{H,m} \times E^V}$, where E^K and E^V are projection dimensions, $E^K = E^V = E^{H,m}/h^m$, and h is the total number of heads for modality m. After that we used scaled dot-product softmax approach to compute the attention score for head i as:

$$Attn(Q_i^m, K_i^m, V_i^m) = \sigma\left(\frac{Q_i^m {K_i^m}^T}{\sqrt{d_k^m}}\right) V_i^m \tag{2.4}$$
$$head_i^m = Attn(Q_i^m, K_i^m, V_i^m) \tag{2.5}$$

After that, all the head feature representation is concatenated and projected to produce the attended feature representation, F_m^a in the following way,

$$F_m^a = [head_1^m; ...; head_h^m] W^{O,m} \tag{2.6}$$

Here, $W^{O,m}$ is the projection parameters of size $E^{H,m} \times E^H$, and the shape of F_m^a is $(B \times S^m \times E^H)$, where E^H is the attended feature embedding size. We used the same feature embedding size

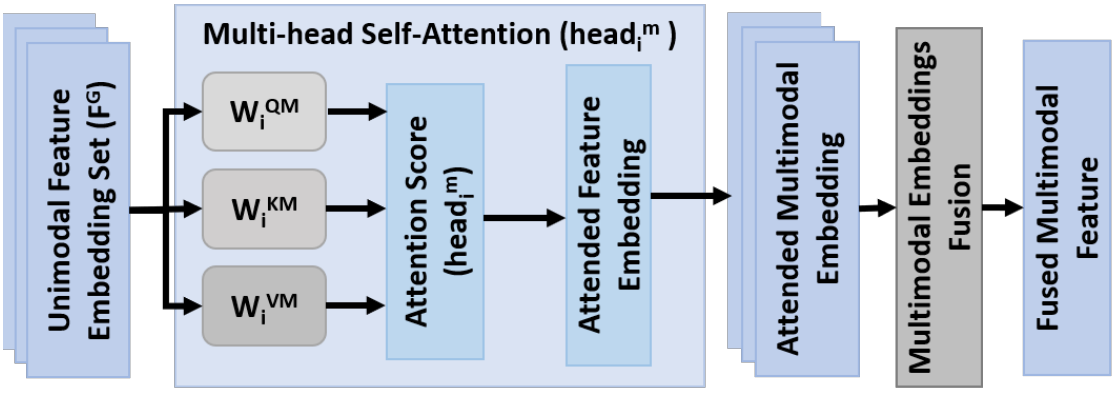


Figure 2.2: MAT : Multimodal Attention-based Feature Fusion Architecture.

E^H for all modalities to simplify the application of multimodal attention MAT for fusing all the modality-specific feature representation, which is presented in the next section 2.1.1.2. However, our proposed multimodal attention based feature fusion method can handle different unimodal feature dimensions. Finally, we fused the attended segmented sequential feature representation F_m^a to produce the local unimodal feature representation F_m of size $(B \times E^H)$. We can use different types of fusion to combine the spatio-temporal segmented feature encodings, such as sum, max, or concatenation. However, the concatenation fusion method is not a suitable approach to fuse large sequences, whereas max fusion may lose the temporal feature embedding information. As the sequential feature representations produced from the same modality, we have used the sum fusion approach to fuse attended unimodal spatial-temporal feature embedding F_m^a,

$$F_m = \sum_{s \in S^m} F_{m,s}^a \tag{2.7}$$

2.1.1 Multimodal Feature Fusion

In this work, we developed a novel multimodal feature fusion architecture based on our proposed multi-head self-attention model, MAT: **M**ultimodal **A**ten**t**ion based Feature Fusion, which is depicted in Fig. 2.2. After encoding the unimodal features using the modular feature encoders, we combine these feature embeddings F_m in an unordered multimodal feature embedding set $F^{G^u} = (F_1, F_2, ..., F_M)$ of size $(B \times M \times D^H)$, where M is the total number of modalities. After that, we fed the set of unimodal feature representations F^{G^u} into MAT, which produces the attended fused multimodal feature representation F^{G^a}.

The multimodal multi-head self-attention computation is almost similar to the self-attention method described in Section 2.1.1.1.3. However, there are two key differences. First, unlike encoding the positional information using LSTM to produce the sequential spatial-temporal feature embedding before applying the multi-head self-attention, in MAT, we combine all the modalities

Table 2.1: Performance comparison (mean top-1 accuracy) of multimodal fusion methods in HAMLET on UT-Kinect dataset [96]

Number of Heads		Fusion Method	
UAT	MAT	MAT-SUM	MAT-CONCAT
1	1	87.97	88.50
1	**2**	**93.50**	**97.45**
2	2	92.50	93.00
2	4	93.50	94.50

feature embeddings without encoding any positional information. Also, MAT and UAT modules have separate multi-head self-attention parameters. Second, after applying the multimodal attention method on the extracted unimodal features, we used two fusion approaches to fused the multimodal features:

- MAT-SUM: extracted unimodal features are summed after applying the multimodal attention

$$F^G = \sum_{m=1}^{M} F_m^{G^a} \tag{2.8}$$

- MAT-CONCAT: in this approach the attended multimodal features are concatenated

$$F^G = [F_1^{G^a}; F_2^{G^a}; ...; F_M^{G^a}] \tag{2.9}$$

2.1.1 Activity Recognition

Finally, the fused multimodal feature representation F^G is passed through a couple of fully-connected layers to compute the probability for each activity class. For aiding the learning process, we applied activation, dropout, batch normalization in different parts of the learning architecture (see the section 2.1.2.2 for the implementation details). As all the tasks of human-activity recognition, which we addressed in this work, are multiclass classification, we trained the model using cross-entropy loss function, mini-batch stochastic gradient optimization with weight decay regularization [95].

$$loss(y, \hat{y}) = \frac{1}{B} \sum_{i=1}^{B} y_i \log \hat{y}_i \tag{2.10}$$

2.1.2 Experimental Setup

2.1.2 Datasets

We evaluated the performance of our proposed multimodal HAR method, HAMLET, using three human-activity datasets: UTD-MHAD [97], UT-Kinect [96], UCSD-MIT [9].

UTD-MHAD [97] human activity dataset consists of a total of 27 human actions covering from sports, to hand gestures, to training exercises and daily activities. Eight people repeated each action for four times. After removing the corrupted sequences, this dataset contains a total of 861 data samples.

UT-Kinect [96] dataset contains a total of ten indoor daily life activities (e.g., walking, standing up, etc.) with three modalities: RGB, depth, and 3D skeleton. Each activity was performed two times by each person. Thus there were a total of 200 activity samples in this dataset.

UCSD-MIT [9] human activity dataset consists of eleven sequential activities in an automotive assembly task. Each assembly task was performed five people, and each person performed the task for five times. This dataset contains there modalities: 3D skeleton data from a motion capture system, and sEMG and IMUs data from a wearable sensor.

2.1.2 Implementation Details

Spatial-temporal feature encoder: We incorporated pre-trained ResNet50 for encoding the RGB and depth data [98]. We applied max pooling with a kernel size of five and stride of three for pooling segment level features. We extended the co-occurrence [93] feature extraction network to encode segmented skeleton and inertial sensor features. Finally, for capturing the temporal features, we used a two-layer unidirectional LSTM. We used embedding size 128 and 256 for UCSD-MIT [9] and UT-Kinect [96] spatial-temporal features embedding respectively.

Hyper-parameters and optimizer: We utilized the pre-trained ResNet architecture for encoding RGB and depth modality. However, in the case of a co-occurrence feature encoder (skeleton and inertial sensor), we applied BatchNorm-2D, ReLu activation, and Dropout layers sequentially. After encoding each unimodal features, we applied ReLu activation and Dropout. Finally, in MAT, after fusing the multimodal features, we used BatchNorm-1D, ReLu activation, and Dropout sequentially. We varied the dropout probability between $0.2 - 0.4$ in different layers. In multi-head self-attention for both unimodal and multimodal feature encoders, we varied the number of heads from one to eight. We train the learning model using Adam optimizer with weight decay regularization option [95] and cosine annealing warm restarts [99] with an initial learning rate set to $3e^{-4}$.

Training environment: We implemented all the parts of the learning model using Pytorch-1.4 deep learning framework [100]. We trained our model in different types of GPU-based computing environments (GPUs: P100, V100, K80, and RTX6000).

2.1.2 State-of-the-art Methods and Baselines

We designed two baseline HAR methods and reproduce a state-of-art HAR method to evaluate the impact of attention method in encoding and fusing multimodal features:

- **Baseline-1 (NSA)** does not use the attention mechanism for encoding unimodal or fusing multimodal features.

Table 2.2: Performance comparison (mean top-1 accuracy) of multimodal HAR methods on UT-Kinect dataset [96]

Method	Fusion Type	Top-1 Accuracy (%)
NSA	SUM	54.34
	CONCAT	52.31
USA	SUM	55.82
	CONCAT	54.34
KEYLESS [37] (2018)	CONCAT	94.50
HAMLET	MAT-SUM	95.56
	MAT-CONCAT	**97.45**

- **Baseline-2 (USA)** only applies multi-head self-attention to encode unimodal features but fuses the multimodal embedding without applying attention. This baseline method is similar to the self-attention based multimodal HAR proposed in [36].
- **Keyless Attention** [37] employed an attention mechanism to encode the modality-specific features. However, it did not utilize attention methods to fuse the multimodal features, instead those were concatenated.

2.1.2 Evaluation metrics

To evaluate the accuracy of HAMLET, the Keyless Attention model [37], the NSA, and the USA algorithms, we performed leave-one-actor-out cross-validation across all the trials for each person on each dataset. Similar to the original evaluation schemes, we reported activity recognition accuracy for the UT-Kinect [96] and the UTD-MHAD datasets [97], and F1-score for the UCSD-MIT dataset [9]. To evaluate HAMLET, the Keyless attention method, and baseline methods on UT-Kinect and UTD-MHAD datasets, we used RGB and skeleton data. We leveraged skeleton, IMUs, and sEMG modalities on the UCSD-MIT dataset.

2.1.3 Experimental Results and Discussion

2.1.3 Multimodal Attention-based Fusion Approaches

We first evaluated the accuracy of two multimodal attention-based feature fusion approaches of HAMLET: MAT-SUM and MAT-CONCAT. We also varied the number of heads used in UAT and MAT steps to determine the optimal configuration of these values.

Results: We evaluated UAT and MAT attention methods as well as the fusion approaches (MAT-SUM and MAT-CONCAT) on the UT-Kinect dataset [96], presented in Table 2.1. We used the RGB and skeleton modalities and reported top-1 accuracy by following the original evaluation

Table 2.3: Performance comparison (mean top-1 accuracy) of multimodal fusion methods on UTD-MHAD dataset [97]

Method	Year	Top-1 Accuracy (%)
Kinect & Inertial [97]	2015	79.10
DMM-MFF [101]	2015	88.40
DCNN [102]	2016	91.2
JDM-CNN [103]	2017	88.10
S^2DDI [104]	2017	89.04
SOS [105]	2018	86.97
MCRL [106]	2018	93.02
PoseMap [107]	2018	94.51
HAMLET (MAT-CONCAT)	-	**95.12**

scheme. The results suggest that the MAT-CONCAT fusion method showed the highest top-1 accuracy (97.45%), with one and two heads in UAT and MAT methods, respectively.

Discussion: The results suggest the concatenation-based fusion approach (MAT-CONCAT) performed better than the summation-based fusion approach (MAT-SUM). Because the MAT-CONCAT allows MAT to disentangle and apply attention mechanisms on the unimodal features to generate robust multimodal features for activity classification. On the other hand, the sum-based fusion method merged the unimodal features into a single representation, which makes it difficult for MAT to disentangle and apply appropriate attention to unimodal features.

The results from Table 2.1 also indicate an improvement in activity recognition accuracy with the increment of the number of heads in the MAT when keeping the number of heads fixed in the UAT. However, this relationship does not hold when the number of heads was changed in the UAT. As a large number of heads reduce the size of feature embedding, increasing the number of heads in the UAT may result in an inadequate feature representation. Thus, based on the size of the features used in this work, the results suggest that one head in the UAT and two heads in the MAT methods display the best accuracy. Thus, we utilized these values for further evaluations.

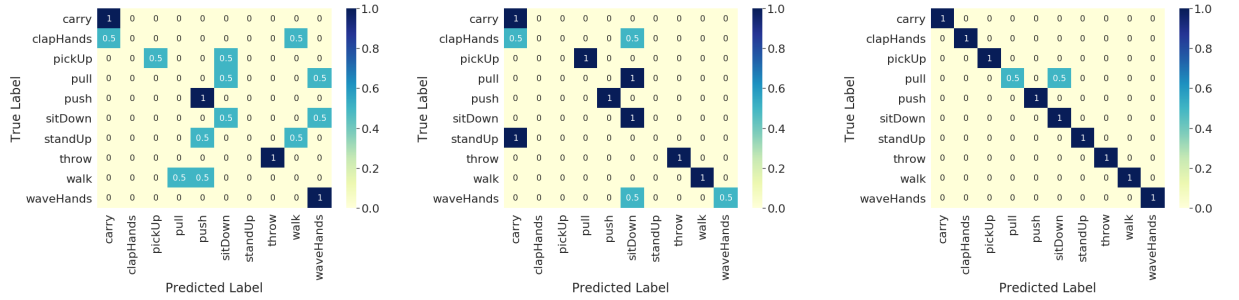
2.1.3 Comparison with Multimodal HAR Methods

As HAMLET takes a multimodal approach, it is reasonable to evaluate the accuracy against the state-of-the-art multimodal approaches. Thus, we compare the performance of HAMLET with two baseline methods (the USA and the NSA, see Sec. 2.1.2.3) and several state-of-the-art multimodal approaches. We presented the results in Tables 2.2 (UT-Kinect), 2.3 (UTD-MHAD) & 2.4 (UCSD-MIT).

Results: In the UT-Kinect dataset, RGB and skeleton modalities have been used to train the learning models. Following the original evaluation scheme, we report the top-1 accuracy in Ta-

Table 2.4: Performance comparison (mean F1-scores in %) of multimodal HAR methods on UCSD-MIT dataset [9]

Method	Fusion Type	F1-Score (%)
NSA	SUM	59.61
	CONCAT	45.10
USA	SUM	60.78
	CONCAT	69.85
KEYLESS [37] (2018)	CONCAT	74.40
Best of UCSD-MIT[9] (2019)	Early Fusion	59.0
HAMLET	**MAT-SUM**	**81.52**
	MAT-CONCAT	76.86



(a) Without attention (b) Unimodal attention (c) Unimodal and multimodal attention

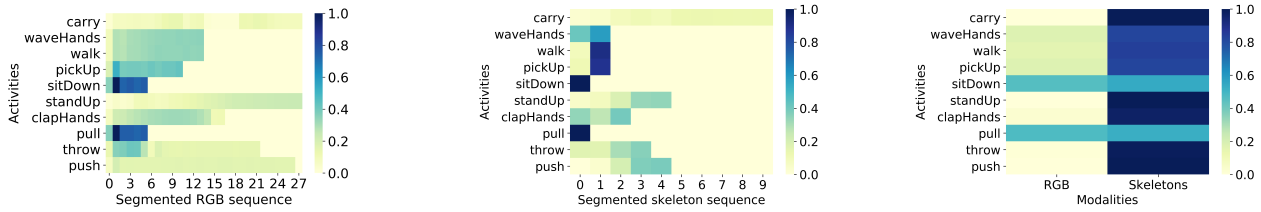
Figure 2.3: Comparative impact of multimodal and unimodal attention in HAMLET for different activities on UT-Kinect dataset.

ble 2.2. The results indicate that HAMLET achieved the highest 97.45% top-1 accuracy across all other methods.

We also evaluate the performance of HAMLET on the UTD-MHAD [97] dataset. We train and test HAMLET on RGB and Skeleton data and report the top-1 accuracy while using MAT-CONCAT in Table 2.3. The results suggest that HAMLET outperformed all the evaluated state-of-the-art baselines and achieved the highest accuracy of 95.12%.

For the UCSD-MIT dataset, all the learning methods are trained on the skeleton, inertial, and sEMG data. All the training models have been used late or intermediate fusion except for the results presented from [9], which used an early feature fusion approach. In Table 2.4, the results suggest that HAMLET with MAT-SUM fusion method outperformed the baselines and state-of-the-art works by achieving the highest 81.52% F1-score (in %).

Discussion: HAMLET outperformed all other evaluated baselines across all datasets and metrics tested. The results on the UTD-MHAD dataset suggest that HAMLET outperformed all the



(a) RGB sequence embedding attention (b) Skeleton sequence embedding attention (c) Multimodal fusion attention

Figure 2.4: Multimodal and unimodal attention visualization for different activities on UT-Kinect Dataset.

state-of-the-art multimodal HAR methods. These methods didn't leverage the attention-based approaches to dynamically weighting the unimodal features to generate multimodal features. The results also suggest that, the other attention-based approaches, such as USA and Keyless [37], also showed better performance compared to the non-attention based approaches on UT-Kinect (Table 2.2) and UCSD-MIT (Table 2.2) datasets. The overall results support that our proposed approach is robust in finding appropriate multimodal features, hence it has achieved the highest HAR accuracies.

The results indicate that the MAT-CONCAT approach achieved higher accuracy on the UT-Kinect dataset; however, the MAT-SUM approach delivered higher accuracy on the UCSD-MIT dataset. One explanation behind this variation is that the modalities (skeleton, sEMG, and IMUs) in the UCSD-MIT dataset represent similar physical body features, thus summing up the feature vectors work well. However, as the UT-Kinect dataset modalities have different characteristics, the visual (RGB) and the physical body (skeleton) features, MAT-CONCAT works better than MAT-SUM.

Finally, the overall results suggest that HAMLET achieved the mean F-1 score of 81.52% on the UCSD-MIT dataset, which is lower compared to the highest accuracy on other datasets (please note that the top-1 accuracies were presented for other datasets). The main reason behind this performance degradation in UCSD-MIT is that this dataset contains missing data, especially sEMG, and IMUs data are missing in many instances. However, in the presence of the missing information, HAMLET showed the best performance compared to all other approaches.

2.1.3 Combined Impact of Unimodal and Multimodal Attention

We evaluated the comparative importance of unimodal and multimodal attention mechanism (presented in Fig. 2.3). We can observe that the incorporation of unimodal attention (Fig. 2.3-b) can help to reduce the miss-classification error in comparison to the non-attention based feature learning method (Fig. 2.3-a). This is because unimodal attention can able to extract the sparse salient spatio-temporal features. We also can observe an improved accuracy in activity classification when the multimodal attention based unimodal feature fusion approach was incorporated (Fig. 2.3-c vs.

a, b). The results indicate that HAMLET can reduce the number of miss-classification, especially in the cases of similar activities, such as sitDown and pickUp, which is depicted in the confusion matrix in Fig. 2.3-c.

2.1.3 Visualizing Impact of Multimodal Attention

We visualize the attention map of the unimodal and multimodal feature encoders to gauge the impact of attention in local (unimodal) and global (multimodal) feature representation in Fig 2.4. We used the data of the eighth performer from the UT-Kinect dataset [96] as a sample data to produce the attention map for different activities, as shown in Fig. 2.4, where we observe that the unimodal attention is able to detect salient segments of RGB (Fig 2.4-a) and skeleton (Fig 2.4-b) modalities. For example, the unimodal attention method focuses on the beginning parts of the *sitDown* and the *pull* activities, as these activities have distinguishable actions in the beginning parts of the activity. On the other hand, the unimodal attention method needs to pay attention to the full sequence to differentiate the *carry* and the *push* activities, as a specific part of these activities are not more informative than the other parts.

Moreover, we evaluate the impact of MAT by observing the multimodal attention map in Fig. 2.4-c, which represents the relative attention given to unimodal features. For example, the *pickUp* and *sitDown* may involve similar skeleton joints movements, and thus if we concentrate only on the skeleton data, it may be challenging to differentiate between these two activities. However, if we incorporate the complementary modalities, such as RGB and skeleton, it may be easier to differentiate between similar activities. Thus, MAT pays equal attention to the RGB and skeleton data while recognizing the *sitDown* activity, whereas solely pay attention to the skeleton data while identifying the *pickUp* activity (Fig. 2.4-c).

2.1.4 Limitations

Although our proposed learning model, HAMLET, outperforms state-of-the-art multimodal HAR approaches, HAMLET does not allow the inter-modality interaction to extract complementary multimodal features. HAMLET uses a self-attention approach that learns weights of different modalities' representations and fuses those representations by multiplying those weights. Additionally, HAMLET fuses those representations at the penultimate task learning layers. Moreover, many state-of-the-art methods do not allow inter-modality interactions and may not learn complementary multimodal features. Consequently, many of these approaches fail to perform well on noisy data. Our experimental results also suggest that the performance of HAR approaches degrades in the presence of noisy data. To address those issues, we have developed, Multi-GAT, Graphical Attention-based Hierarchical Multimodal Representation Learning Approach [10]. We present this graphical attention multimodal learning methods in the next section.

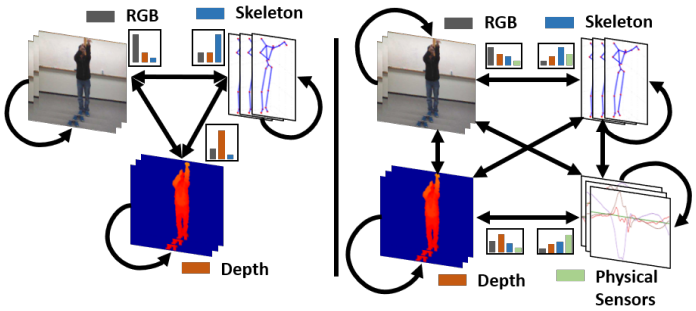


Figure 2.5: Example scenarios of multimodal graphical attention approach applied to a basketball-shooting activity. Here, RGB modality pays greater attention to visual modalities (RGB and Depth), whereas Skeleton modality pays greater attention to non-visual modalities (Physical Sensors) for extracting complementary multimodal features.

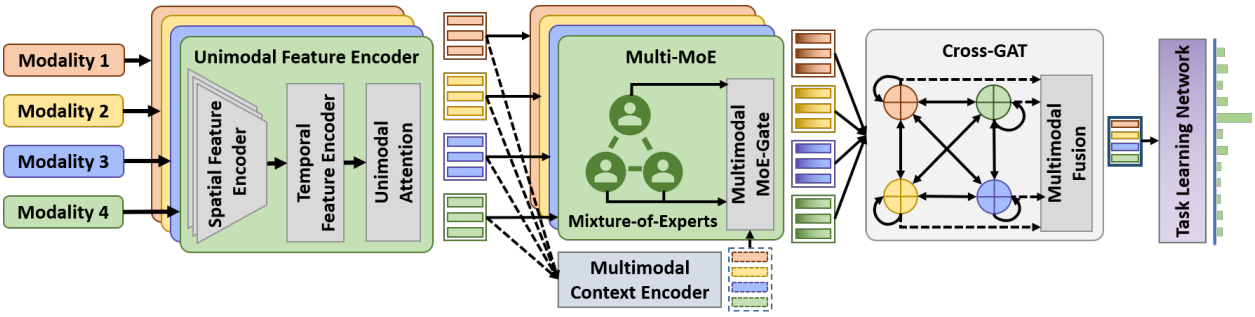
2.2 Graphical Attention-based Hierarchical Multimodal Representation Learning

Self-attention-based multimodal learning models, such as HAMLET [7] and Keyless [37], do not allow the modalities to share information and calibrate their representations to extract complementary multimodal representation. As a result, these models can not extract complementary multimodal representations, specifically in the presence of noisy sensor data. To address these issues, we have developed, Multi-GAT, Graphical Attention-based Hierarchical Multimodal Representation Learning Approach [10].

2.2.1 Proposed Multimodal Learning Approach

In this section, we present our proposed multimodal feature learning method for HAR, called Multi-GAT: Graphical Attention-based Hierarchical Multimodal Representation Learning approach (see Fig. 2.6). Multi-GAT consists of four sequential learning modules:

- **Unimodal Feature Encoder:** Modality-specific salient features are encoded by using spatial-temporal feature encoders and a unimodal attention module (Section 2.2.1.2).
- **Multimodal Mixture-of-Experts (Multi-MoE) Model:** Multimodal mixture-of-experts (MoE) model factors and extracts salient unimodal features utilizing a conditional attention method (Section 2.2.1.3).
- **Cross-Modal Graphical Attention (Cross-GAT) Approach:** A novel message-passing based graphical attention approach captures the cross-modal relationships to extract complementary multimodal features (Section 2.2.1.4).



(a) Unimodal Feature Encoder (b) Multimodal Mixture-of-Experts (c) Cross-Modal Graphical Attention

Figure 2.6: Multi-GAT: Graphical Attention-based Hierarchical Multimodal Representation Learning Framework for HAR. (a) First, unimodal feature encoders extract modality-specific spatial-temporal features independently. (b) Second, Multimodal Mixture-of-Experts (Multi-MoE) model disentangles and extracts salient unimodal features by employing the multimodal context. (c) Third, Cross-Modal Graphical Attention (Cross-GAT) is employed to capture inter-modality relationships for producing complementary multimodal features. Finally, multimodal features are used in the task learning network (HAR).

- **Task Learning Network:** A task learning network uses multimodal feature representation for HAR (Section 2.2.1.5).

Consider an example scenario where a robot has three sensor modalities (RGB, depth, and skeleton) and is applying Multi-GAT for HAR (see Fig. 2.5). In this scenario, Multi-GAT will first employ Unimodal Feature Encoder (Section 2.2.1.2) and Multi-MoE (Section 2.2.1.3) model for each of the three modalities independently to extract modality-specific salient features. If we only consider RGB modality, the unimodal feature encoder will extract salient spatial-temporal features from the RGB modality. Multi-MoE will then use the Self-MoEAT (Section 2.2.1.3.1) module to factor RGB features to produce a set of experts. Multi-MoE will also produce multimodal context from the extracted unimodal features from all the three modalities. MMoE-Gate (Section 2.2.1.3.2) will then utilize the multimodal context to pool the salient features from the RGB experts and produce the encoded feature for RGB modality.

Multi-GAT will then employ Cross-GAT (Section 2.2.1.4) to extract multimodal features from the encoded RGB, depth, and skeleton features. If we only consider RGB modality, then RGB modality will utilize conditional attention to determine the directional relationship from RGB to all other modalities. Cross-GAT will then use this relationship to extract the complementary features for RGB modality from all the other modalities. The same procedure will be applied for depth and skeleton modalities, which resembles the complete cross-modal graphical attention approach. Finally, the extracted complementary features of RGB, depth, and skeleton modalities will be fused to produce the multimodal features, which will then be used in the robot for HAR.

2.2.1 Pre-processing of Data Modalities

Multi-GAT uses a separate data pre-processing approach for each modality m. First, unimodal raw data is split to produce a sequence of segmented features $X_m = (x_{m,1}, ..., x_{m,S_m})$ of size $(B \times S_m \times D_m^r)$, where B is the batch size, S_m is the segment size, and D_m^r is the raw feature dimension of the modality $m \in M$ (r stands for raw feature).

2.2.1 Unimodal Feature Encoder

Multi-GAT employs unimodal feature encoder to extract salient modality-specific features, as each modality has unique feature characteristics and distributions. This approach allows Multi-GAT to leverage transfer learning by utilizing pre-trained unimodal encoders. Moreover, the results from our previous work (see Islam and Iqbal [7] for detail) suggest that this modality-specific feature encoding approach helps to leverage intermediate fusion and allows inter-modality interactions at the abstract features space.

In Multi-GAT, each unimodal feature encoder consists of two sequential learning modules: Spatial-Temporal Feature Encoder and Unimodal Attention Module.

2.2.1.2.1 Spatial-Temporal Feature Encoder In Multi-GAT, we adopt the spatial-temporal feature encoder architecture, similar to the one used in [7]. Each unimodal feature sequence X_m is encoded to extract the spatial-temporal features in two sequential steps. First, the modality-specific spatial encoder is used for encoding each unimodal feature sequence to produce $X_m^s = (x_{m,1}^s, ..., x_{m,S_m}^s)$ of size $(B \times S_m \times D_m^s)$, where D_m^s is the spatial feature dimension (s stands for spatial). Also, we pool the segment-level spatial features. This pooling mechanism reduces the feature overlaps, noise, and computational complexity [7]. Each unimodal encoder has separate learning architecture to encode features. For example, we used a ResNet model to encode visual modalities (RGB and depth) and co-occurrence learning architecture [93] to encode skeleton and physical sensors modalities.

Second, Multi-GAT employs a long-short-term-memory (LSTM) recurrent neural network (RNN) on the extracted spatial features to produce the spatial-temporal features $X_m^t = (x_{m,1}^t, ..., x_{m,S_m}^t)$ of size $(B \times S_m \times D_m^f)$, where D_m^f is the feature dimension of modality m (t and f stand for temporal and feature, respectively). We used LSTM over the other variations of RNN, as LSTM can capture the long-term feature dependencies which is crucial for HAR.

2.2.1.2.2 Unimodal Attention Module Although spatial-temporal feature encoders can encode unimodal features, it may not sparsely extract the salient sequential features. Recently, the attention approach has been widely adopted in the literature to extract salient sequential features. In Multi-GAT, we leverage unimodal self-attention approach to sparsely and adaptively extract the salient modality-specific features X_m^a (a stands or attention) from the encoded spatial-temporal feature sequence X_m^t in the following way,

$$X_m^a = \sum_{i=1}^{S_m} \alpha_{m,i} X_{m,i}^t \tag{2.11}$$

Here the attention weight $\alpha_{m,i}$ is calculated as follows,

$$\beta_{m,i} = W_m^{t^T} X_{m,i}^t \tag{2.12}$$

$$\alpha_{m,i} = \frac{exp(\beta_{m,i})}{\sum_i^{S_m} exp(\beta_{m,i})} \tag{2.13}$$

Here W_m^t is the modality-specific learnable parameters. We can represent this self-attention approach as,

$$X_m^a = SelfAttn(X_{m,i}^t, W_m^t) \tag{2.14}$$

2.2.1 Multimodal Mixture-of-Experts (Multi-MoE) Model

In this section, we present the second learning module of Multi-GAT, called Multi-MoE: Multimodal Mixture-of-Experts (Fig. 2.7). In the previous step of Multi-GAT, unimodal feature encoder may not disentangle and extract the unimodal salient features which can complement the multimodal feature. In Multi-GAT, we adopt the Mixture-of-Experts (MoE) model [108] with two key extensions to extract the complementary multimodal features. First, Self Mixture-of-Experts Attention (Self-MoEAT) independently factors unimodal features, which represent the experts set. Second, Multimodal MoE Gate (MMoE-Gate) employs a conditional gating method for pooling the salient features from unimodal experts conditioned on the multimodal context.

2.2.1.3.1 Self Mixture-of-Experts Attention (Self-MoEAT) First, each unimodal feature embedding X_m^a is factored using modality-specific set of experts network F_m^e to create N_m^e experts set (e stands for experts) for modality m,

$$E_{m,i} = F_{m,i}^e(X_m^a) \quad , i \in N_m^e \tag{2.15}$$

Self-MoEAT allows the interaction among the intra-modality experts through a query-key-value based conditional attention approach for disentangling the unimodal feature. The conditional attention weights are calculated over modality-specific experts features $E_m = \{E_{m,1}, E_{m,2}, ..., E_{m,F_m^e}\}$ condition on each expert i of modality m. At first each expert's feature is projected into query (Q), key (K) and value (V) vectors as follows,

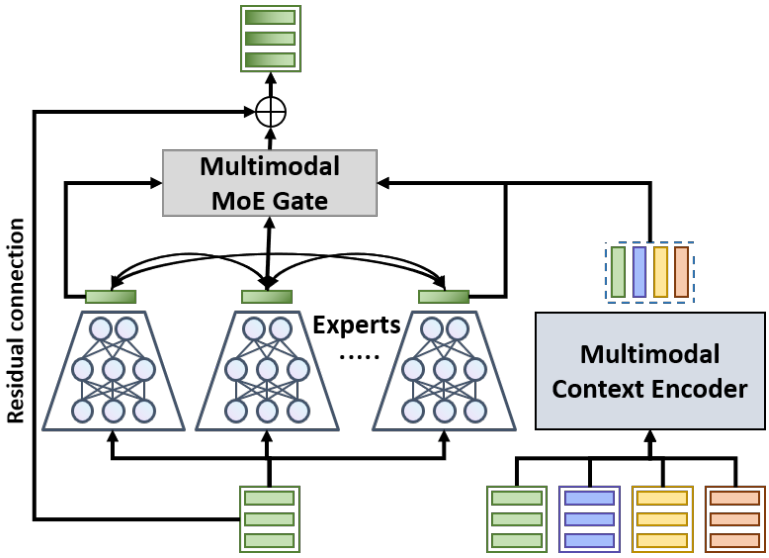


Figure 2.7: Multimodal Mixture-of-Experts (Multi-MoE).

$$Q^e_{m,i} = E_{m,i}W^Q_{m,i}; K^e_{m,i} = E_{m,i}W^K_{m,i}; V^e_{m,i} = E_{m,i}W^V_{m,i} \tag{2.16}$$

Here, $W^Q_{m,i} \in \mathbb{R}^{D^f_m \times D^e_m}$, $W^K_{m,i} \in \mathbb{R}^{D^f_m \times D^e_m}$, and $W^V_{m,i} \in \mathbb{R}^{D^f_m \times D^e_m}$ are the learnable parameters, where D^e_m is the feature dimension of i^{th} expert. The conditional attention scores are calculated to produce the fused feature representation for each expert i in the following way,

$$\alpha^e_{m,i} = \sigma\left(\frac{Q^e_{m,i}K^{e^T}_m}{\sqrt{D^e_m}}\right)V^e_m \quad , i \in N^e_m \tag{2.17}$$

$$E^a_{m,i} = \alpha^e_{m,i}W^e_m \quad , i \in N^e_m \tag{2.18}$$

Here $W^e_m \in \mathbb{R}^{D^e_m \times D^f_m}$ are the experts feature projection parameters. We represent this conditional attention-based feature extraction approach with learnable parameter set $W_{m,i} = (W^Q_{m,i}, W^K_{m,i}, W^V_{m,i})$ as follows,

$$E^a_{m,i} = CondiAttn(Q^e_{m,i}, K^e_m, V^e_m, E_m, W_{m,i}) \tag{2.19}$$

2.2.1.3.2 Multimodal MoE Gate (MMoE-Gate) In the second step, Multi-MoE uses MMoE-Gate to sparsely pool the salient features from unimodal experts. Multi-MoE employs self-attention

(Eq. 2.14) to create multimodal context E^c (c stands for multimodal context) from the encoded unimodal features $X^a = (X^a_1, X^a_2, ..., X^a_M)$:

$$E^c \quad = \quad SelfAttn(X^a, W^c) \tag{2.20}$$

MMoE-Gate leverages the conditional attention (Eq. 2.19) with the multimodal context E^c for sparsely and adaptively gating salient features from unimodal experts $E^a_m = (E^a_{m,1}, E^a_{m,2}, ..., E^a_{m,N^e_m})$:

$$E^c_m = CondiAttn(E^c, K^c_m, V^c_m, E^a_m, W^g_m) \tag{2.21}$$

Here, learnable parameter set W^g_m is used to produce keys (K^c_m) and values (V^c_m). Finally, MMoE-Gate uses a residual connection from the unimodal spatial-temporal features to produce the fused unimodal feature representation,

$$E^u_m = E^c_m + X^a_m \quad , (u : unimodal) \tag{2.22}$$

This residual connection limits the deviation of unimodal feature space, due to the MMoE-Gate.

2.2.1 Cross-Modal Graphical Attention (Cross-GAT)

In Multi-GAT, Unimodal Feature Encoder and Multi-MoE module extract the salient modality-specific features independently. However, these modules do not allow inter-modality interactions to distill complementary multimodal features. We propose a novel message-passing based cross-modal graphical attention method, called Cross-GAT, for inter-modality interaction (Fig. 2.8). Multi-GAT uses Cross-GAT to capture inter-modality relationship for pooling complementary multimodal features.

In Cross-GAT, the unimodal feature set $E^u = (E^u_1, E^u_2, ..., E^u_M)$ is represented by a complete directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Here, each node $m \in \mathcal{V}$ represents modality-specific features and edge set $\mathcal{E}$ denotes the directed inter-modality relationships. Each modality m extracts features from the neighboring modalities $\mathcal{N}_m$ based on the relations. Apart from the data-driven approach to appropriately determine the inter-modality relationship, Cross-GAT also allows incorporating the domain expert's knowledge in creating a multimodal relationship graph for determining the inter-modality relationships. Although domain expert knowledge can be useful in some situations, it can often be biased and may not represent the appropriate relationships among the modalities. Therefore, in this implementation of Cross-GAT, we leverage a data-driven approach to determine the relationships among the modalities for extracting complementary multimodal representation.

In Cross-GAT, $\mathcal{N}_m$ initially contains all modalities except m. Each modality m interacts with its neighboring modalities $\mathcal{N}_m$ using conditional attention to determine inter-modality relationships (edge-weights) for extracting the complementary features (message). At first, each unimodal feature is projected to produce query, key, and value vectors:

$$\begin{aligned} Q_m^x &= E_m^u W_m^{Q^x}; K_m^x = E_m^u W_m^{K^x}; V_m^x = E_m^u W_m^{V^x} && (2.23) \\ K_m^{\mathcal{N}} &= \{K_i^x\}, V_m^{\mathcal{N}} = \{V_i^x\}, i \in \mathcal{N}_m && (2.24) \end{aligned}$$

Here, $(W_m^{Q^x}, W_m^{K^x}, W_m^{V^x}) \in \mathbb{R}^{3 \times D_m^f \times D_m^x}$ are the learnable parameters and D_m^x is the node feature dimension in graphical attention. Each modality m adapts its feature representation by extracting the complementary features from its neighbor modalities using the conditional attention.

$$\alpha_m^x = \sigma\left(\frac{Q_m^x K^{\mathcal{N}^T}}{\sqrt{D_m^x}}\right) V_m^{\mathcal{N}} \quad , m \in M \tag{2.25}$$

$$Msg(m) = \alpha_m^x W_m^x \quad , m \in M \tag{2.26}$$

Here $W_m^x \in \mathbb{R}^{D_m^x \times D^f}$ are the learnable feature projection parameters and D^f is the shared multimodal feature dimension. This attention α_m^x represents the directional relations from modality m to other modalities. The conditionally attentive message $Msg(m)$ is used to produce cross-modal complementary feature representation for modality m:

$$E_m^x = Msg(m) + E_m^u \quad , m \in M \tag{2.27}$$

Finally, Cross-GAT concatenates and projects the cross-modal complementary features $E^x = (E_1^x, E_2^x, ..., E_M^x)$ to produce the fused multimodal feature representation,

$$E^f = [E_1^x; E_2^x; ...; E_M^x] W^O \tag{2.28}$$

Here, $W^O \in \mathbb{R}^{D^f \times D^O}$ are learnable parameters.

2.2.1 Task Learning

The task learning network uses fused multimodal feature representation E^f. In this work, we used these multimodal features for a human-activity recognition task. The multimodal features are passed through a task learning network F^t to classify the activities. We used a cross-entropy loss function with stochastic gradient to train the network,

$$loss(y, \hat{y}) = \frac{1}{B} \sum_{i=1}^{B} y_i \log \hat{y}_i \quad , y = F^t(E^f) \tag{2.29}$$

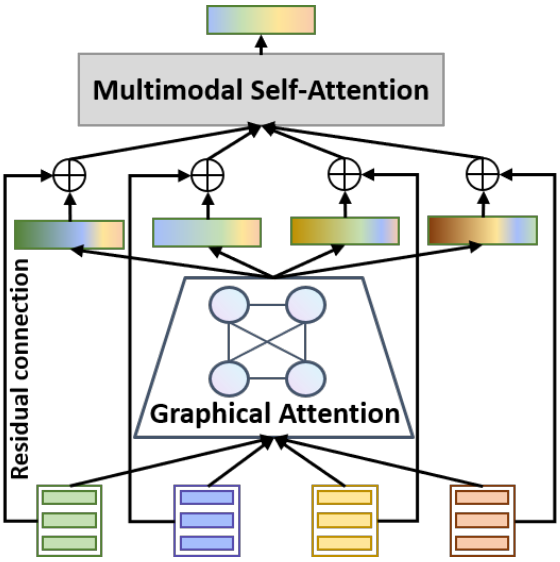


Figure 2.8: Cross-Modal Graphical Attention (Cross-GAT).

2.2.2 Experimental Setup

2.2.2 Datasets

We evaluated our approach by comparing its performance to several contemporary HAR methods on two multimodal HAR datasets, MMAct [8] and UTD-MHAD [97]. MMAct dataset consists of 37 daily life activities recorded with seven modalities, where twenty subjects performed each activity for five times, resulting in $37K$ video clips. Many of these demonstrations have visually occluded data. We used four viewpoints of RGB videos, acceleration, gyroscope, and orientation data for our analysis. UTD-MHAD consists of 27 activities performed by eight persons for four times, resulting in 861 daily activities clips recorded with four modalities: RGB, depth, skeleton, and physical sensors.

2.2.2 Learning Architecture Implementation

We segmented the visual modalities (RGB and depth) with the window size 1 and stride 3. For skeleton and physical sensors, we used window size 5 and stride 5. We utilized the ResNet-50 to extract spatial features for visual modalities and the Co-occurrence method [93] for the skeleton and physical sensors. Each modality is encoded with the feature embedding size of 256. As an increased number of experts increases the computation cost, we decided to use two experts in each modality for our evaluations. We applied BatchNorm, ReLU-activation, and dropout (probability varies in between $0.2-0.5$) sequentially after each layer and module. We used two fully connected layers with ReLU activation after the first layer to represent the task learning network F^t.

Table 2.5: Cross-subject performance comparison (F1-Score) of multimodal learning methods on MMAct dataset

Method	F1-Score (%)
SMD [109]	63.89
Student [8]	64.44
Multi-Teachers [8]	62.67
MMD [8]	64.33
MMAD [8]	66.45
HAMLET [7]	69.35
Keyless [37]	71.83
Multi-GAT	**75.24**

We used Pytorch deep learning framework to implement the evaluated approaches. We used Adam optimizer with weight decay regularization [95] and cosine annealing warm restarts with an initial learning rate set to $3e^{-4}$ to train the evaluated methods. We trained the models in a distributed manner on two RTX-6000 GPUs.

2.2.3 Experimental Results and Discussion

2.2.3 Comparison with Multimodal HAR Methods

Results: We evaluated the performance of Multi-GAT by applying on two multimodal HAR datasets: MMAct [8] and UTD-MHAD [97]. For the MMAct dataset, we followed the train-test splits and F1-Score evaluation metric for cross-subject and cross-session based evaluation settings originally proposed by Kong et al.[8]. The results suggest that Multi-GAT outperformed all other state-of-the-art methods by achieving 75.24% and 91.48% F1-Score values in cross-subject (Table 2.5) and cross-session (Table 2.6) based evaluation settings, respectively. These performances of Multi-GAT on MMAct dataset posit 3.41% and 7.59% improvement on F1-score over benchmark multimodal learning methods in cross-subject and cross-session evaluation settings, respectively.

For the UTD-MHAD, we followed a leave-one-subject-out evaluation approach and top-1 accuracy metric by following the evaluation criteria proposed in the original paper [97]. The experimental results on UTD-MHAD (Table 2.7) suggest that Multi-GAT outperformed all other state-of-the-art approaches by achieving 97.56% top-1 accuracy with four modalities (RGB, depth, skeleton, and physical sensor).

Discussion: Multi-GAT outperformed all other evaluated baselines across all datasets and metrics tested. The results from Tables 2.5, 2.6 & 2.7 suggest that similar to Multi-GAT, the other attention-based approaches (HAMLET, MMD and Keyless) exhibit better performance compared to the non-attention based approaches (PoseMap and TSN). The attention-based approaches ex-

Table 2.6: Cross-session performance comparison (F1-Score) of multimodal learning methods on MMAct dataset

Method	F1-Score (%)
SVM+HOG [110]	46.52
TSN (RGB) [111]	69.20
TSN (Optical-Flow) [111]	72.57
MMAD [8]	74.58
TSN (Fusion) [111]	77.09
MMAD(Fusion) [8]	78.82
Keyless [37]	81.11
HAMLET [7]	83.89
Multi-GAT [10]	**91.48**

hibit such improved performance because they allow to attend and fuse the salient features from heterogeneous modalities adaptively. Although previous attention-based approaches achieve performance gain by extracting salient features, those approaches may not effectively capture the cross-modal relation, which is crucial for producing robust complementary multimodal features. The experimental results support that multimodal mixture-of-experts module (Multi-MoE) and graphical attention approach (Cross-GAT) aid Multi-GAT to disentangle and extract cross-modal relationships while producing robust multimodal representation. Thus, Multi-GAT has achieved the highest HAR accuracy.

Although Multi-GAT outperformed all the evaluated methods in cross-subject and cross-session evaluation settings on the MMAct dataset, there is a performance gap between these two evaluation settings. This performance gap can also be observed for all other state-of-the-art approaches as the multimodal learning approaches are not generalized enough for the unseen subjects. Moreover, all the approaches performed relatively worse in the cross-subject evaluation setting than in the cross-session evaluation setting. Thus, there are scopes for developing generalized multimodal representation learning approaches to improve the HAR performance for unseen subjects in the future.

2.2.3 Impact of Modalities in Multimodal Learning

Results: To investigate the performance of Multi-GAT in various modality conditions, we compared its accuracy with two state-of-the-art multimodal learning methods, HAMLET [7] and Keyless [37]. We conducted this study on the UTD-MHAD dataset and varied the types of modalities (RGB, depth, skeleton, and physical sensors). We presented the results in Table 2.8, which suggest that Multi-GAT outperformed the other methods for all the variations of the modality types and numbers tested.

Table 2.7: Performance comparison (top-1 accuracy) of multimodal learning methods on UTD-MHAD dataset.

Method	Accuracy (%)
MHAD [97]	79.10
SOS [105]	86.97
S^2DDI [104]	89.04
DCNN [102]	91.20
Keyless [37]	92.67
MCRL [106]	93.02
PoseMap [107]	94.51
HAMLET [7]	95.12
Multi-GAT [10]	**97.56**

Discussion: The results from Table 2.8 suggest that the HAR accuracies of Multi-GAT increase with the addition of a modality. This may indicate that Multi-GAT is robust in computing salient multimodal features as it computes the relationships among various modalities using the Cross-GAT module, which ensures that the extracted features contain complementary information captured from multiple modalities. The performance degradation of HAMLET and the Keyless method in these conditions suggest that the accuracy of these methods is susceptible to variation in modalities.

2.2.3 Impact of Noisy Modalities

Results: We evaluated the robustness of Multi-GAT in the presence of noisy data by comparing its performance with two other state-of-the-art multimodal learning approaches: HAMLET and Keyless on the UTD-MHAD dataset. We randomly dropped raw features from three heterogeneous modalities (RGB, skeleton, and physical sensors) with a 50% probability of introducing noise. As we are interested in investigating the impact of noise on heterogeneous modalities, and as the data from RGB and depth represent the visual modalities, we did not include the depth modality in this experimental evaluation.

We conducted this study on two evaluation settings: first, imputing noise during both training and testing phases; second, training the learning models without imputing noise and introducing noise only in the testing phase. The results in Table 2.9 indicate that Multi-GAT outperformed the other methods in the presence of noisy data.

Furthermore, the MMAct dataset contains visually occluded and non-occluded data in a real-world setting. We conducted the experimental evaluations on both cross-subject and cross-session evaluation settings of the MMAct dataset, presented in Tables 2.5 and 2.6. The results also suggest that Multi-GAT outperformed all the evaluated baselines in both evaluation settings on the MMAct

Table 2.8: Performance comparison (Accuracy %) of the impact of modality changes on UTD-MHAD dataset. R: RGB, D: Depth, S: Skeleton, P: Physical Sensors.

Learning Methods	Modality Combinations		
	R+S	R+S+P	R+D+S+P
Keyless	90.20	92.67	83.87
HAMLET	95.12	91.16	90.09
Multi-GAT	**96.27**	**96.75**	**97.56**

Table 2.9: Performance comparison (Accuracy %) of noisy modalities' impact on UTD-MHAD dataset. R: RGB, S: Skeleton, P: Physical Sensors (Depth is not used)

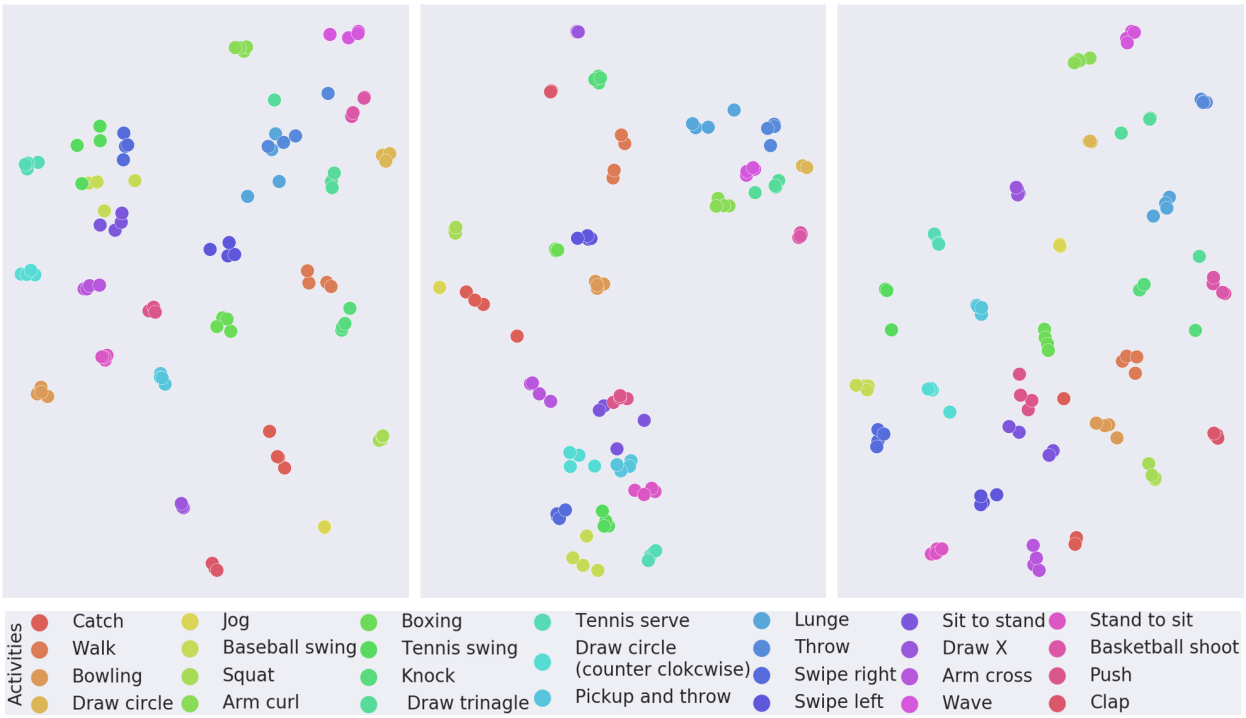
Evaluation Settings	Learning Methods	No Noise	Noisy Modalities	
			S+P	R+S+P
Train & Test (w/wo Noise)	Keyless	92.67	88.03	85.48
	HAMLET	91.16	87.31	84.04
	Multi-GAT	**96.75**	**90.23**	**87.54**
Test (w/wo Noise)	Keyless	92.67	60.04	61.52
	HAMLET	91.16	70.59	71.42
	Multi-GAT	**96.75**	**77.88**	**80.02**

dataset.

Discussion: The results in Table 2.9, along with the results in Tables 2.5 and 2.6, indicate that Multi-GAT outperformed other methods in the presence of noisy data on the UTD-MHAD and MMAct datasets, and provide evidence of the robustness of our method in noisy environments. Although HAMLET and the Keyless utilize attention mechanisms to extract unimodal and multimodal features, none of these methods allow cross-modal interaction while generating salient multimodal features. On the other hand, the Cross-GAT module of Multi-GAT can capture cross-modal relationships, which enables extracting robust complementary multimodal features in the presence of noisy data. Therefore, Multi-GAT achieved the highest HAR accuracies, even in the presence of noisy data. The performance of Multi-GAT and all the evaluated methods, however, degrade with more noisy modalities. Moreover, the accuracy drops when we train the model without noise and introduce noise during testing. Thus, extracting complementary multimodal features that can perform robustly with unseen noisy data will be an exciting future research direction.

2.2.3 Ablation Study of Learning Modules

Results: We conducted an ablation study to assess the importance of various learning modules in Multi-GAT. We performed this study on the UTD-MHAD dataset with RGB, skeleton, and



(a) Non-attention (b) Unimodal attention (c) Cross-GAT

Figure 2.9: Comparative impact of cross-modal graphical attention in Multi-GAT to learn multi-modal representation for HAR on UTD-MHAD dataset with RGB, Skeleton and Physical Sensor modalities (t-SNE embeddings).

physical sensor modalities and present the results in Table 2.10. We also evaluated the impact of cross-modal graphical attention in Multi-GAT by visualizing the feature embeddings in the t-distributed stochastic neighbor embedding (t-SNE) visualizations (Fig. 2.9).

Discussion: In Table 2.10, the results suggest that the utilization of modality-specific attention method (MA) for extracting unimodal salient features aids in improving the HAR accuracy. However, MA alone does not allow the inter-modality interaction, closely resembling two state-of-the-art works, HAMLET and Keyless. Our proposed novel Multi-MoE (ME) and Cross-GAT (CG) learning modules allow inter-modality interaction for capturing the cross-modal relations. This inter-modality interaction and extraction of cross-modal relationships enable Multi-GAT to compute complementary multimodal features, which improves HAR accuracy by 6% compared to the non-attention based approach. In Multi-GAT, residual connections help to prevent the vanishing and exploding gradients issues.

The t-SNE visualization in Fig. 2.9-a indicates that when unimodal and multimodal features are extracted without applying the attention mechanism (Non-attention), the feature learning method can not produce distinguishable representations. Although incorporating the unimodal attention

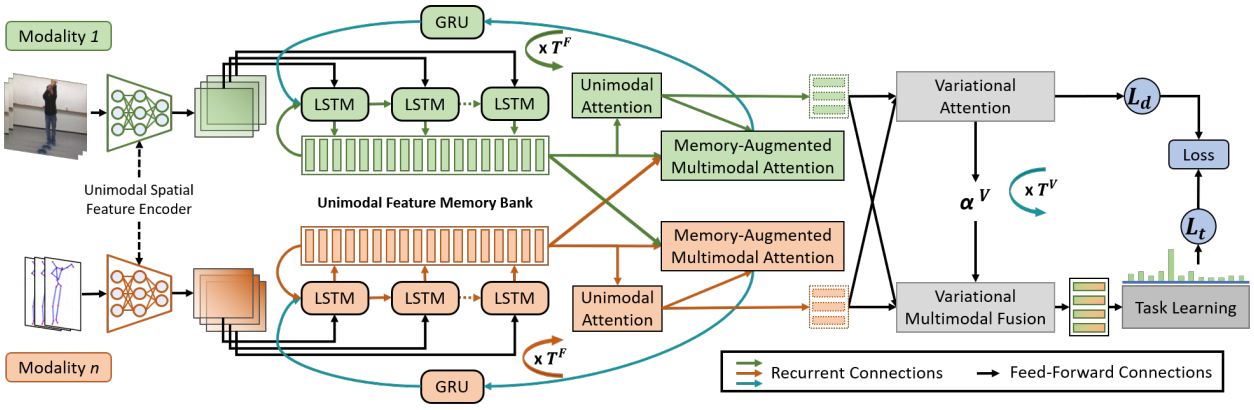
Table 2.10: Ablation study of Multi-GAT learning modules on UTD-MHAD dataset. Here, MA: Modality-Specific Attention, R: Residual Connection, ME: Multi-MoE, MC: Multimodal Context, CG: Cross-GAT

Learning Modules	Top-1 Accuracy (%)
Non-Attention	90.44
MA	92.30
MA+ME+R	93.25
MA+ME+R+MC	95.04
MA+CG+R	95.09
MA+ME+R+MC+CG+R	**96.75**

method can cluster the multimodal features better than the non-attention mechanism (in Fig. 2.9-b), that mechanism alone can not produce well-separated clusters of multimodal features. The t-SNE plot in Fig. 2.9-c indicates that our proposed cross-modal graphical attention approach helps the Multi-GAT to produce distinguishable and well-spaced clusters of multimodal feature representations. The reasoning behind this well-spaced clustered multimodal feature extraction is that Multi-GAT utilizes multimodal graphical attention, which allows inter-modality interaction for capturing the cross-modal relationships while extracting the distinguishable multimodal features.

2.2.4 Limitations

Although Multi-GAT achieved improved HAR performance compared to our other multimodal learning model, HAMLET, specifically in the presence of missing and noisy data modalities, there are two primary limitations of Multi-GAT. First, Multi-GAT uses a mixture-of-experts and graphical attention approaches to extract multimodal representations, which are resource-intensive learning modules. Graphical attention approach creates a complete graph of interactions among the data modalities, which computationally expensive. For example, four data modalities create 16 interactions among the modalities, and each interaction uses a multi-head attention module to calibrate each modality's representations. Additionally, Multi-GAT use mixture-of-experts modules on each unimodal representations which create added computational burden. Second, Multi-GAT uses feed-forward fusion approach, which does not allow unimodal feature encoders to calibrate unimodal representations. As a result, the unimodal encoders can not update their representations and align with the other multimodal representations. This bottleneck limits the learning model to extract salient multimodal representations.



(a) Unimodal Feature Encoder (b) ReMATE Learning (c) VAT (d) Task

Figure 2.10: MAVEN : Memory-Augmented Recurrent Approach for Multimodal Fusion . (a) MAVEN employs Unimodal Feature Encoders to encode modality-specific features. (b) Recurrent Memory-Augmented Attention-based Feature Alignment Approach (ReMATE) iteratively aligns unimodal features by leveraging memory banks. (c) Multimodal Variational Attention-based Fusion Approach (VAT) fuses multimodal features. (d) A task learning network uses fused multimodal representations to determine the outcome. MAVEN is trained end-to-end for HAR using joint task learning and variational inference losses.

2.3 Recurrent Multimodal Fusion

State-of-the-art multimodal learning approaches combine information in a feed-forward manner, which prevents each modality from aligning and refining representation. In these approaches, unimodal feature encoders extract features independently without observing features from other modalities. However, if these encoders have information about other modalities, they can utilize that information to *align* and iteratively *refine* the unimodal features to generate robust representations from the noisy sensor data. Prior works in machine learning [112], [113] and neuroscience on human perception [114], [115] suggest that recurrent information processing aids task learning over feed-forward learning approaches. To address these issue, we have developed a recurrent multimodal fusion approach, MAVEN, to recurrently calibrate and fuse unimodal representations and extract complementary multimodal representations.

2.3.1 Proposed Multimodal Learning Approach

In this section, we describe our proposed approach, MAVEN: Memory-Augmented Recurrent Approach for Multimodal Fusion. The four main components of MAVEN are depicted in Fig. 2.10 and are as follows:

- **Unimodal Feature Encoder**: MAVEN extracts spatial features from each modality to capture the distinct feature characteristics (Section 2.3.1.1).
- **Recurrent Memory-Augmented Attention-based Feature Alignment Approach (ReMATE)**: ReMATE iteratively aligns and refines the modality-specific features by leveraging the unimodal feature memory banks (Section 2.3.1.2.1 and Algorithm 1).
- **Multimodal Variational Attention-based Fusion Approach (VAT)**: VAT aligns unimodal features in the multimodal feature space to produce fused multimodal feature representation (Section 2.3.1.2.2).
- **Task Learning Network**: MAVEN utilizes the multimodal features for task-specific learning (Section 2.3.1.3).

2.3.1 Unimodal Feature Encoder

In MAVEN, the raw feature sequence of each modality m is pre-processed and segmented to produce, $X_m^R = (x_{(m,1)}, x_{(m,2)}, ..., x_{(m,N_m^R)})$ of size $(B \times N_m^R \times F_m^R)$. Here, B, N_m^R, and F_m^R are the batch size, sequence length, and raw feature dimension, respectively (R stands for raw feature). In the subsequent sections, we use the notations' superscript to denote the type of variables and subscripts for indexing.

MAVENr MAVEN employs state-of-the-art unimodal encoders to extract spatial features from each segment, $X_m^S = (x_{(m,1)}, ..., x_{(m,N_m^S)})$ of size $(B \times N_m^S \times F_m^S)$. Here, N_m^S and F_m^S are the spatial feature sequence length and dimension, respectively, of modality m (S represents spatial features). The reasoning behind the modality-specific spatial feature encoding approach is twofold. First, as each modality has its unique feature characteristics, it is best to implement modality-specific encoders to capture the salient representation appropriately. Moreover, interaction among the unimodal features at early layers may prevent the encoders from capturing modality-specific feature characteristics [7], [12], [26], [42]. Second, this unimodal feature learning approach allows the use of transfer learning by leveraging the state-of-the-art feature encoders.

2.3.1 Multimodal Feature Alignment and Fusion

Ideally, multiple modalities should have aligned features that capture the same phenomena. Having aligned representations is key to providing a robust multimodal representation for a given downstream task. However, these modalities may represent different semantic concepts due to the heterogeneous feature characteristics. Thus, MAVEN iteratively aligns the unimodal features prior to the fusion for producing a robust multimodal representation. This unimodal feature alignment is also observed in the animals' multisensory cognition systems [1], [116].

Furthermore, noisy or incomplete data from a modality may not provide relevant information compared to other modalities. For example, in low light environments, visual modalities may not

contain the relevant salient information compared to physical sensors. This discrepancy among unimodal representations can lead to negative knowledge transfer in the multimodal feature space, whereby a lack of information from one modality may weaken the representation of other modalities. In these scenarios, aligning the features can help the weak representation of some modalities to recalibrate based on the information from other modalities.

MAVEN hierarchically aligns and fuses the multimodal features in two steps. First, MAVEN employs Recurrent Memory-Augmented Attention-based Feature Alignment Approach (ReMATE) to align unimodal features (Section 2.3.1.2.1). Second, MAVEN leverages our proposed Multimodal Variational Attention-based Fusion Approach (VAT) to fuse the converged salient unimodal features to produce a robust multimodal representation (Section 2.3.1.2.2).

2.3.1.2.1 Recurrent Memory-Augmented Attention-based Feature Alignment Approach (ReMATE) As each modality has unique characteristics, such as feature distribution and sampling rate, MAVEN utilizes ReMATE to incrementally align the unimodal features. MAVEN recurrently applies the following three sub-learning modules of ReMATE: Unimodal Feature Memory Bank, Unimodal Attention, and Memory-Augmented Multimodal Attention, to align unimodal features for extracting relevant multimodal information. The procedure of ReMATE is summarized in Algorithm 1 and depicted in Fig. 2.10.

Unimodal Feature Memory Bank: ReMATE produces a memory bank (β) of spatial-temporal features, $X_m^\beta = f^R(X_m^S, X_{m'}^C)$ of size $(B \times N_m^\beta \times F_m^\beta)$ for each modality m. Here, N_m^β and F_m^β are the memory bank lengths and feature dimensions of each memory bank entry, respectively. ReMATE leverages recurrent neural network, f^R, to produce memory bank from spatial features X_m^S, previously extracted by unimodal feature encoders (Section 2.3.1.1). Although we can use several learning architectures to design f^R, such as transformers, GRU or LSTM, we have used LSTM to reduce the complexity and capture long-range spatial-temporal features. At the first iteration of ReMATE, as there is no prior multimodal contextual information, $X_{m'}^C$ is set to NULL. In subsequent iterations, X_m^β is refined using multimodal contextual information, $X_{m'}^C$, in the following way,

$$X_m^\beta = f^R(X_m^\beta, X_{m'}^C) \tag{2.30}$$

Here, $X_{m'}^C$ contains the multimodal contextual information (C) of all modalities except modality m. ReMATE employs the Memory-Augmented Multimodal Attention module to extract $X_{m'}^C$, described later in this section. ReMATE uses the multimodal context to align and refine modality-specific features and memory banks based on the observed information from other modalities. At the first iteration of ReMATE, as $X_{m'}^C$ is empty, unimodal memory bank X_m^β contains only modality-specific spatial-temporal features, which is the output from f^R at each timestamp. Thus, the sequence of spatial-temporal features is stored in the memory bank using a write operation. At subsequent iterations, f^R reads the contents from the memory bank, X_m^β, and uses multimodal

context $X^C_{m'}$ to iteratively refine X^β_m and align the spatial-temporal unimodal latent features with the other modalities (Algorithm 1, Line 5). These refined spatial-temporal features replace the old contents of memory banks.

ReMATE utilizes unimodal attention to extract the salient modality-specific features from the memory bank. The memory banks can store long-term temporal context, which helps to capture the relationship among modalities. The long-term relationship helps the modality-specific feature encoder to iteratively re-calibrate the spatial-temporal features in a fine-grained manner, thus obtaining an aligned representation by observing information in memory banks of other modalities. This feature alignment approach also helps a noisy modality to obtain relevant salient representation by aligning its features to the representations from other less-noisy modalities. Moreover, the aligned unimodal features help the variational attention module (Section 2.3.1.2.2) to produce a fused complementary multimodal representation.

Unimodal Attention: In the second step, ReMATE employs a unimodal attention approach to extract salient modality-specific features, $X^U_m \in \mathbb{R}^{B \times F^U_m}$ (Algorithm 1, Lines 6-7). Here, F^U_m is the unimodal (U) feature embedding size. The unimodal attention approach attends each entry of the feature memory bank X^β_m to extract salient representation X^U_m of modality m. The attention weight $\alpha_{(m,i)}$ of each memory bank entry i can be calculated in the following way,

$$\gamma_{(m,i)} = W^{U^T}_m X^\beta_{(m,i)} \tag{2.31}$$

$$\alpha_{(m,i)} = \frac{exp(\gamma_{(m,i)})}{\sum_i^{N^M_m} exp(\gamma_{(m,i)})} \tag{2.32}$$

Here, $W^{U^T}_m$ is a learnable parameter. Subsequently, $\alpha_{(m,i)}$ is used to fuse memory bank features and extract salient unimodal representation,

$$X^U_m = \sum_{i=1}^{N^\beta_m} \alpha_{(m,i)} X^\beta_{(m,i)} \quad , m \in M \tag{2.33}$$

In ReMATE, we determine the attention weights by leveraging a lightweight 1D-CNN with a filter size of 1.

Memory-Augmented Multimodal Attention Module: ReMATE utilizes unimodal representation X^U_m to extract the multimodal context $X^C_{m'}$ from memory banks of all modalities except modality m. $X^C_{m'}$ is used to iteratively align and refine unimodal features and memory banks.

First, ReMATE projects X^U_m and produces a query vector (Q^β_m) for each modality m (Algorithm 1, Line 8),

$$Q^\beta_m = X^U_m W^{Q^B}_m \quad , \quad m \in M \tag{2.34}$$

Algorithm 1: ReMATE(X^S, M, T^F)

Input: X^S: spatial feature, M: modalities, T^F: total ReMATE iterations
Output: X^U: salient unimodal features

1 $X^C_{m'} \leftarrow \emptyset$ ▷ Multimodal context set to empty
2 $X^\beta_m \leftarrow f^R(X^S_m), \forall m \in M$ ▷ Initialize memory bank
3 **for** $k \leftarrow 1$ *to* T^F **do**
4 **Repeat steps 4-9 for each modality** $m \in M$
5 $X^\beta_m \leftarrow f^R(X^\beta_m, X^C_{m'})$ ▷ Memory bank (Eq. 2.30)
6 Calculate attention $\alpha_{(m,i)}$ for each entry of X^β_m
7 ▷ (Eqs. 2.31 & 2.32)
8 $X^U_m = \sum_{i=1}^{N^\beta_m} \alpha_{(m,i)} X^\beta_{(m,i)}$ ▷ Unimodal feature (Eq. 2.33)
9 Project query feature Q^β_m from X^U_m ▷ (Eq. 2.34)
10 Project key ($K^\beta_{m'}$) and value ($V^\beta_{m'}$) features from other modalities memory bank ▷ (Eqs. 2.35 & 2.36)
11 $X^C_{m'} = \sigma\left(\frac{Q^\beta_m K^{\beta^T}_{m'}}{\sqrt{F^\beta_m}}\right) V^\beta_{m'}$ ▷ Multimodal context (Eq. 2.37)
12 $X^C_{m'} = GRU(X^C_{m'}, X^{C'}_{m'})$ ▷ (Eq. 2.38)
13 **end**
14 $X^U \leftarrow (X^U_m), \forall m \in M$
15 **return** salient unimodal features X^U

Here, $W^{Q^\beta}_m$ is a learnable parameter for query projection. Similarly, ReMATE projects the memory bank X^β_m to produce key (K^β_m) and value (V^β_m) feature vectors for each modality m in the following way (Algorithm 1, Line 9),

$$K^\beta_m = X^U_m W^{K^\beta}_m; V^\beta_m = X^U_m W^{V^\beta}_m \quad , m \in M \tag{2.35}$$

$$K^\beta_{m'} = \{K^\beta_i\}, V^\beta_{m'} = \{V^\beta_i\} \quad , i \in \{M \setminus m\}, m \in M \tag{2.36}$$

Here, $W^{K^\beta}_m$ and $W^{V^\beta}_m$ are learnable parameters for key and vector projections. ReMATE then uses the query feature Q^β_m to extract the multimodal context, $X^C_{m'}$, for modality m (Algorithm 1, Line 10). Additionally, we have used GRU to capture the correlation of $X^C_{m'}$ in the subsequent iterations.

$$X^{C'}_{m'} = \sigma\left(\frac{Q^\beta_m K^{\beta^T}_{m'}}{\sqrt{F^\beta_m}}\right) V^\beta_{m'} \quad , m \in M \tag{2.37}$$

$$X^{C}_{m'} = GRU(X^{C}_{m'}, X^{C'}_{m'}) \tag{2.38}$$

Finally, ReMATE sends $X^{C}_{m'}$ to the first module (Unimodal Feature Memory Bank, Eq. 2.30) T^F times for aligning and refining the unimodal feature and memory bank of modality m. These steps are summarized in Algorithm 1.

The primary reason for incorporating our proposed recurrent feature alignment approach in the multimodal learning model, MAVEN, is to extract aligned representations by incrementally aligning modality-specific representations. ReMATE aligns modality-specific representations by observing representations from other modalities. The information from other modalities can help a modality refine its own feature representation. On the other hand, many state-of-the-art multimodal learning approaches use feed-forward feature fusion, where the modalities cannot align their representations before fusion by observing other modalities' representations. To the best of our knowledge, we are the first to propose a recurrent feature representations alignment approach, where each modality can iteratively refine the unimodal representation by observing the other modalities.

2.3.1.2.2 Multimodal Variational Attention-based Fusion Approach (VAT) In MAVEN, we propose a multimodal variational attention approach (VAT) to fuse the aligned unimodal features for extracting salient multimodal representation. In VAT, we consider multimodal attention as a random variable following a prior distribution, in line with prior work on sequence-to-sequence learning [117], [118]. The parameters of this random variable are obtained using amortized variational inference [119], which approximates the *evidence lower bound* (ELBO) over the attention weights.

We model the attention weights as the posterior distribution over the concatenated multimodal representation, $X^{U^C} = (X^U_1; X^U_2; ...; X^U_M)$:

$$\alpha^V_m \sim q_{\theta_m}(\alpha^V_m | X^{U^C}) \tag{2.39}$$

Here, α^V_m represents variational (V) attention weights, sampled from the posterior distribution $q(\cdot|X^{U^C})$ for each modality m. The posterior distribution can be matched to a prior distribution, using Kullback-Leibler (KL) divergence loss. The sampled attention weights help to fuse features by capturing relationships among the modalities.

Finally, MAVEN uses this multimodal variational attention weights to fuse features, $X^U = (X^U_1, X^U_2, ..., X^U_M)$, for producing the multimodal representation, $X^{F'}$,

$$X^{F'}_m = \alpha^V_m X^U_m \tag{2.40}$$

$$X^F_m = [X^{F'}_1; X^{F'}_2; ....; X^{F'}_M]W^O \tag{2.41}$$

Moreover, MAVEN samples the multimodal attention weights T^V times to tighten the decision boundaries. Thus, MAVEN produces a set of multimodal representation $X^M = (X_1^F, X_2^F, ..., X_{T^V}^F)$ of size $(T^V \times F^M)$, where F^M is the dimension of the multimodal feature.

Although several approaches have been proposed in the literature to fuse the features by employing concatenation or deterministic attention-based fusion approach, these approaches can not ensure the alignment among the unimodal feature representations in the multimodal representation space. Moreover, due to the absence of explicit feature alignment, these deterministic attention models suffer in aligning and fusing relevant features from noisy data modalities. In these scenarios, our proposed variational attention-based feature fusion can help to extract robust multimodal representations for two reasons. First, VAT learns a distribution over attention weights for each unimodal representation. This results in their implicit alignment as VAT imposes the same prior distribution over the attention weights. The variational attention approach has been widely studied to align the latent representations in sequence-to-sequence based learning models [117], [118]. However, to the best of our knowledge, we are the first to effectively incorporate the variational attention approach for multimodal latent representation alignment and fusion. Second, learning a distribution allows VAT to model uncertainty when fusing the unimodal features. VAT learns to sample attention weights from different parts of the distribution space for noisy and non-noisy signals to fuse multimodal features and thus produce robust representation.

2.3.1 Task Learning

The task learning network utilizes multimodal features to make decisions for various tasks. In this work, we use the multimodal features for the task of activity recognition. Each multimodal feature representation is passed through a task learning network f_t to produce the outputs of task t,

$$y = \sum_{i=1}^{T^V} f_t(X_i^M) \tag{2.42}$$

Learning Losses: To train the learning architecture for a particular task t, we design a combined training loss. First, we use a cross-entropy loss to ensure MAVEN learns the task-specific multimodal features,

$$L_t(y, \hat{y}) = \frac{1}{B} \sum_{i=1}^{B} y_i \log \hat{y}_i \tag{2.43}$$

Here, B is the batch size. Second, we include KL-divergence loss between the posterior distribution of the attention weights and a prior distribution. This KL-loss enforces a prior on the attention weights of each modality, thus ensuring that the unimodal latent representations align in the multimodal feature space.

$$L_d = \sum_{m \in M} D_{KL} \left[q_{\theta_m}(\alpha^V | X^{U^C}) || p(\alpha^V) \right] \tag{2.44}$$

We model $p(\alpha^V)$ as a Normal distribution, $\mathcal{N}(\mu_m^a, \sigma_m^a)$, using the reparameterization trick [119], where μ_m^a and σ_m^a are obtained by a recognition neural network, with parameters θ_m. The reparameterization trick introduces a random variable $\epsilon \sim \mathcal{N}(0, 1)$, which is multiplied by the μ_m^a and σ_m^a in order to sample from the Normal distribution. The random variable ϵ allows us to model stochasticity, whereas the recognition network allows us to train end-to-end and backpropagate through a distribution.

Finally, the task-specific loss and KL-divergence-based distribution alignment loss are combinedly utilized to train the learning architecture for the target task.

$$L(y, \hat{y}) = L_t(y, \hat{y}) + \gamma^{KL} L_d \tag{2.45}$$

Here, γ^{KL} is the weight of KL-divergence loss.

2.3.2 Experimental Setup

We evaluated MAVEN by comparing its performance to several state-of-the-art multimodal fusion approaches for the task of human activity recognition. Furthermore, we conducted experiments with noisy data to evaluate the robustness of MAVEN for HAR.

2.3.2 Human Activity Datasets

We evaluated the performance of MAVEN by applying on three widely used Human Activity Recognition datasets: UTD-MHAD [97], MMAct [8], and UCSD-MIT[9].

UTD-MHAD consists of 27 human actions covering sports, hand gestures, training exercises and daily activities. Each action was performed by eight people and repeated over four trials. The dataset is comprised of skeleton, depth, inertial, and RGB data modalities. In our experiments, we used the RGB, skeleton and inertial modalities.

MMAct dataset consists of 37 daily life activities, with each activity performed by 20 people and repeated over five trials. The dataset is comprised of seven modalities, spanning from RGB data to acceleration and gyroscope. We used the two available viewpoints of RGB videos, acceleration, gyroscope, and orientation data in our experiments. MMAct dataset contains visually occluded data samples, which helps us to evaluate the multimodal learning approaches in extracting complementary multimodal features for activity recognition.

UCSD-MIT dataset contains nine automotive and block assembly activities from 2 groups. The gross activity-group contains four activities (e.g., walking, receiving part, and attaching part), and the fine activity-group contains five activities (e.g., palmar grab, pincer grab, and ulnar pinch

grab). Five people performed each activity five times. UCSD-MIT dataset uses data from three modalities: 3D joint positions from motion capture system, sEMG, and IMUs data from wearable sensors. In our experimental evaluations, we have used all the available modalities.

2.3.2 Implementation Details

Learning Architecture: We segmented the visual modalities (RGB and depth) with the window size = 1 and stride = 3. For skeleton and physical sensors, we segmented using a window size = 5 and stride = 5. We set the maximum raw sensor feature sequence length before segmentation to 100. We employed modality-specific spatial feature encoders to extract unimodal features. We utilized the ResNet-50 learning model [98], which is pre-trained on ImageNet, to extract the RGB features. The extracted ResNet features from both the RGB and Depth modalities are then passed through a fully connected neural network to produce feature embeddings of size 128. We leveraged a two-layer Convolutional Neural Network (CNN) based co-occurrence learning model [93] to encode the physical sensor modalities (skeleton, accelerometer, gyroscope, and orientation). In this two-layer CNN, we used 64 and 32 channels with (1×1) and (3×3) kernel sizes, respectively. We also applied BatchNorm after each convolutional layer. The extracted CNN features are then passed through a fully connected neural network to produce 128 sized feature embeddings. After extracting the modality-specific features, we applied BatchNorm, ReLu and Dropout (50% features dropout probability) sequentially in each unimodal feature encoder.

The extracted unimodal features are passed through a modality-specific LSTM layer (f^R) with the hidden feature dimension of 128 to capture temporal features and produce a spatial-temporal memory bank for each modality. The encoded spatial-temporal features from memory banks are passed through a self-attention module to extract salient unimodal features, which are then fused (summed) to produce a spatial-temporal representation. We used 1-D convolutional neural network to implement the self-attention module. The spatial-temporal features are then passed through ReLu and Dropout (50% dropout probability) layers sequentially in each unimodal feature encoder. We recurrently applied ReMATE, as stated in Algorithm 1, to iteratively refine unimodal features and memory banks. We have empirically found that recurrent iteration 3 provides the best trade-off between computational time and performance. Thus, we have used 3 recurrent iteration in all the experimental evaluations.

Finally, we employed VAT to combine the unimodal features. We used a fully connected network, which has an input dimension of $(128 \times total\_number\_of\_modalities)$ and an output dimension of 128, to produce fused multimodal features. The extracted multimodal features are then passed through BatchNorm, ReLu, and Dropout (50% dropout probability) layers. Finally, these multimodal features are fed into a fully connected neural network with 128 and $total\_number\_of\_activities$ as input and output dimensions, respectively, to produce activity recognition probabilities.

Training details: We trained the model end-to-end by employing our proposed combined loss (Eq. 2.43). We set the KL-divergence loss weight, $\gamma^{KL} = 0.3$. We used Pytorch [100](version:

1.6) and Pytorch-Lightning [120] (version: 0.9.1rc3) as deep learning framework. We utilized Adam optimizer with weight decay regularization and cosine annealing warm restarts with an initial learning rate set to $3e^{-4}$ to train the evaluated methods [99]. In cosine annealing warm restarts learning scheduler, we set the cycle length (T_0) and cycle multiplier (T_{mult}) to 100 and 2, respectively, to train the learning model on the UTD-MHAD dataset. For the MMAct dataset, we set the cycle length (T_0) and cycle multiplier (T_{mult}) to 30 and 2, respectively.

We trained each evaluated models for 80, 210, and 210 epochs on MMAct, UTD-MHAD, and UCSD-MIT datasets, respectively. The batch size used in all our experiments was 2. We held-out 10% validation data from the training dataset of UTD-MHAD and MMAct datasets to select the best performing model, which is then evaluated on the test datasets. We used Pytorch-Lightning wrapper implementation of Pytorch distributed-data-parallel library to train the models in a distributed manner. We also fix the random seed in Pytorch-Lightning deep learning framework to ensure reproducibility of the training process and the experiment results. To learn more about the detailed implementation, we highly encourage the reader to look at our submitted source code and documentations.

2.3.3 Experimental Results and Discussion

2.3.3 Comparison with Multimodal HAR Methods

Results: We evaluated the performance of MAVEN by applying it on three multimodal HAR datasets, UTD-MHAD [97], MMAct [8], UCSD-MIT [9]. For state-of-the-art HAR methods, we followed the original implementation details to evaluate the performance or reported results from the original paper when available.

For MMAct dataset [8], we followed the original subject and session-based evaluation settings and reported the F1-score. The experimental results suggest that MAVEN outperformed all state-of-the-art multimodal HAR approaches in both subject and session-based evaluation settings (Tables 2.11 & 6.9). MAVEN achieved the highest F1-score of 76.76% (Table 2.11) and 95.38% (Table 6.9) in the subject and session-based evaluations, respectively.

For our experiments on the UTD-MHAD dataset [97], we followed the leave-one-subject-out cross-validation evaluation approach and reported average top-1 accuracy (Table 2.13). The results indicate that MAVEN outperformed all the evaluated multimodal HAR approaches by achieving 96.45% top-1 accuracy (Table 2.13).

For UCSD-MIT dataset [9], we followed the leave-one-subject-out cross-validation evaluation setting and reported average top-1 accuracy (Table 2.14). The results indicate that MAVEN outperformed all the evaluated multimodal HAR approaches by achieving 63.68% top-1 accuracy.

Discussion: The experimental results indicate that MAVEN outperformed all the evaluated baselines across all the datasets and metrics tested for human activity recognition (Tables 2.11, 6.9, 2.13 & 2.14). MAVEN achieved 4.93% and 11.49% performance improvement over state-of-the-art results on cross-subject and cross-session based evaluations of MMAct dataset, respectively.

Table 2.11: Cross-subject performance comparison (F1-Score) of multimodal learning methods on MMAct dataset [8] in cross-subject evaluation setting.

Method	F1-Score (%)
SMD [109]	63.89
Student [8]	64.44
Multi-Teachers [8]	62.67
MMD [8]	64.33
MMAD [8]	66.45
HAMLET [7]	69.35
Keyless [37]	71.83
Multi-GAT [10]	75.24
MuMu [11]	76.28
MAVEN [81]	**76.76**

Table 2.12: Cross-session performance comparison (F1-Score (%)) of multimodal learning methods on MMAct dataset [8] in cross-session evaluation setting.

Method	F1-Score (%)
SVM+HOG [110]	46.52
TSN (RGB) [111]	69.20
TSN (Optical-Flow) [111]	72.57
MMAD [8]	74.58
TSN (Fusion) [111]	77.09
MMAD (Fusion) [8]	78.82
Keyless [37]	81.11
HAMLET [7]	83.89
MuMu [11]	87.50
Multi-GAT [10]	91.48
MAVEN [81]	**95.38**

Moreover, MAVEN shows performance gain on leave-one-subject-out cross-validation evaluation on UTD-MHAD and UCSD-MIT datasets. Although MAVEN outperforms all the approaches on the UCSD-MIT dataset, the performance of all the evaluated approaches degrades on the challenging UCSD-MIT dataset with the leave-one-subject-out evaluation setting compared to the cross-subject evaluations on UTD-MHAD and MMAct datasets. The reasoning behind this performance degradation is that UCSD-MIT contains only wearable sensors data, which varies considerably across subjects. These performance improvements posit the generalized multimodal representation learning capabilities of MAVEN.

Table 2.13: Performance comparison (top-1 accuracy) of multimodal learning methods on UTD-MHAD dataset [97] in leave-one-subject-out evaluation setting.

Method	Accuracy (%)
MHAD [97]	79.10
SOS [105]	86.97
JDM-CNN [103]	88.10
DMM-MFF [101]	88.40
S^2DDI [104]	89.04
DCNN [102]	91.20
Keyless [37]	92.67
MCRL [106]	93.02
PoseMap [107]	94.51
HAMLET [7]	95.12
Multi-GAT [10]	97.56
MuMu [11]	97.60
MAVEN [81]	**97.81**

Table 2.14: Performance comparison (F1-Score) of multimodal learning methods on UCSD-MIT dataset [9] in leave-one-subject-out evaluation setting.

Learning Methods	Merge Types	F1-Score (%)
Non-Attention	SUM	52.35
	CONCAT	50.92
HAMLET [7]	SUM	50.04
	CONCAT	48.26
Keyless [37]	SUM	51.68
	CONCAT	54.48
Best of UCSD-MIT [9]	Early Fusion	51.00
Multi-GAT [10]	-	56.77
MuMu [11]	-	61.34
MAVEN [81]	-	**63.68**

The results imply that attention-based modality-specific feature extraction approaches, such as MAVEN, Keyless [37], and HAMLET [7], show better performance compared to other evaluated feature fusion approaches (Tables 2.11-2.14). MAVEN differentiates from state-of-the-art approaches with regards to feature alignment and fusion. In general, state-of-the-art multimodal learning approaches fuse features in a feed-forward manner, restricting each modality from refining its representation when observing other modalities' features.

The experimental results (Table I-IV) suggest that MAVEN outperformed state-of-the-art models which use early [9], late [8], [37], [111] or intermediate [7], [10], [11] fusion approaches. The primary difference between MAVEN and state-of-the-art fusion approaches is that these fusion approaches do not allow cross-modal interaction to re-calibrate the unimodal representations for multimodal fusion. For example, Keyless [37] extracts salient unimodal representations, which are concatenated to produce multimodal representations. This late fusion approach does not allow cross-modal interaction to re-calibrate the unimodal representations. Although intermediate fusion approaches allow cross-modal interactions, these approaches incrementally fuse the representations and do not allow unimodal feature encoders to re-calibrate the representations after observing the complete representations from other modalities. Additionally, although early fusion approaches have been used in some learning models, these approaches are not applicable when the feature distributions across modalities are considerably different, such as fusing representations from visual and wearable sensor data modalities. For this reason, we have compared the performance of MAVEN with the early fusion approach [9] on UCSD-MIT dataset. The results in Table IV suggest that MAVEN outperformed the early fusion approach, which does not allow cross-modal interaction to recalibrate and align unimodal representations.

In contrast to state-of-the-art multimodal learning models, MAVEN employs recurrent feature alignment, which allows to refine and align the unimodal representations by observing the representations of other modalities. Moreover, MAVEN uses variational attention-based fusion approach (VAT) to align multimodal representation to extract complementary representations for robust activity recognition. The performance gains of MAVEN suggest the importance of incorporating our proposed ReMATE and VAT modules to align and fuse salient multimodal representation for HAR. Additionally, the performance improvement of MAVEN on visually occluded data samples of the MMAct dataset suggests that our proposed multimodal fusion approach can extract complementary multimodal representations to recognize activities accurately.

2.3.3 Evaluation on Noisy Data

It is often unrealistic in real-world settings to assume that a learning model can get only non-noisy data from all modalities for training and inference. To investigate how various algorithms perform in the presence of noisy data, we randomly injected noise in the data of various modalities. We randomly choose a modality with 50% probability and randomly drop 20% or 50% of the raw features. We evaluated the performance of MAVEN on noisy data from the MMAct dataset in cross-subject evaluation setting.

We compare the performance of MAVEN against Keyless [37], HAMLET [7], and Non-Attention baseline approach. Non-Attention baseline approach extracts unimodal features without utilizing attention and feature alignment mechanism and concatenates features to obtain multimodal representation. These baselines use feed-forward architecture whereas MAVEN uses our proposed recurrent feature alignment based multimodal representation learning approach. The results in Table 2.15 indicate that MAVEN shows robust performance even in the presence of noisy

Table 2.15: Performance comparison (F1-Score (%)) on noisy data modalities of MMAct dataset in cross-subject evaluation setting.

Method	Noisy Level		
	None	20%	50%
Non-Attention	69.90	59.81	59.37
Keyless [37]	71.83	69.51	67.24
HAMLET [7]	69.35	69.94	69.00
MAVEN without VAT	73.83	72.50	72.17
MAVEN [81]	**76.76**	**74.15**	**72.27**

data compared to other evaluated baselines.

Results and Discussion: The results in Table 2.15 suggest that MAVEN achieved better performance on noisy data compared to other multimodal HAR approaches. Our proposed recurrent feature alignment approach (ReMATE) aligns the representations from noisy modalities to non-noisy modalities to obtain salient representations. Moreover, our proposed variational attention-based fusion approach (VAT) forces each modality to follow a prior distribution in the multimodal feature representations space, which helps the non-noisy modalities to restrict the noisy modalities from deviating the multimodal representation. For example, if the visual modality provides noisy data, ReMATE and VAT use other less-noisy modalities, such as physical sensors, to guide visual modalities for obtaining relevant and robust features. Furthermore, the results suggest that our recurrent feature alignment-based multimodal learning approach, without variational attention-based fusion, can outperform the state-of-the-art learning models on noisy data that uses a feed-forward-based multimodal learning model. Thus, this performance improvement shows the importance of our recurrent feature alignment technique to extract complementary multimodal representation from noisy sensor data over the feed-forward learning models.

2.3.3 Ablation Studies

We conducted several ablation studies to assess the impact of different learning modules and recurrent learning parameters of MAVEN. We conducted the following ablations studies: quantitative ablation study of MAVEN's learning components, the impact of memory bank, significance and qualitative analysis of MAVEN by applying on the MMAct dataset [8].

2.3.3.3.1 Quantitative Ablation Study of Various Components of MAVEN We first ablated different learning modules such as Unimodal Attention (UA), recurrent feature alignment (ReMATE), and variational attention-based multimodal fusion (VAT), to investigate their impact on the overall performance. First, we evaluate the efficacy of VAT. To this end, we developed a transformer [94] style deterministic self-attention models for multimodal fusion, called DMA. We de-

Table 2.16: Ablation study of MAVEN learning modules on MMAct dataset in cross-subject evaluation setting. UA: Unimodal Attention, DMA: Deterministic Multimodal Attention.

Learning Modules	F1-score (%)
Non-Attention	69.90
UA	71.83
UA+DMA(H-1, L-1)	69.35
UA+DMA(H-2, L-2)	72.84
UA+DMA(H-4, L-2)	70.95
Transformer	71.18
Transformer+DMA(H-1, L-1)	71.95
UA+ReMATE ($T^F = 3$)	73.83
UA+ReMATE ($T^F = 3$)+DMA(H-1, L-1)	74.72
UA+ReMATE ($T^F = 3$)+DMA(H-2, L-2)	73.39
UA+ReMATE ($T^F = 10$)	75.56
MAVEN: UA+ReMATE ($T^F = 3$)+VAT	**76.76**

veloped different DMA models by varying the number of heads and layers in the attention model. For example, $DMA(H-2, L-2)$ denotes DMA attention model with two heads and two layers of attention model. For the DMA-based learning model, we have extracted features from each modality independently using self-attention based unimodal feature encoder. The extracted unimodal features are then fused using a deterministic self-attention model. Second, we developed another model using Transformer model [94], where we replaced the LSTM model with this Transformer model to design temporal encoder f^R. After extracting the unimodal representations using the Transformer model, we concatenated unimodal representations for activity recognition. Additionally, we developed another Transformer-based learning model where the extracted unimodal features using the Transformer model are fused using DMA. The experimental results of these ablation studies are presented in Table 2.16 (ablation study of MAVEN learning modules).

Third, we evaluate the efficacy of our recurrent feature alignment (ReMATE) approach, by varying the recurrent alignment iterations of ReMATE in between 1 and 10. The experimental results of this ablation study are presented in Table 2.17.

Results and Discussion: The results in Table 2.16 suggest that using unimodal attention (UA) to extract modality-specific features results in a performance improvement over a Non-Attention approach. When we use a deterministic multimodal attention model with one head and one layer of self-attention model $(DMA(H-1, L-1))$ for feature fusion, along with UA, we noticed a marginal performance degradation compared to when only UA was used. Although increasing the number of heads and attention layer from one to two in the DMA model leads to performance improvement, increasing the number of heads from two to four degrades the learning model's performance. During training time, we noticed that DMA with four heads and two layers of attention

Table 2.17: Impact of recurrent iteration of ReMATE in MAVEN (without VAT) on cross-subject evaluation of MMAct dataset.

Recurrent Iterations	F1-score (%)
1	70.45
3	73.83
7	73.56
10	75.56

model leads to the model overfitting, which is the primary reason for the performance degradation. Thus, increasing the multimodal attention model parameters (head and layer in the self-attention model) does not guarantee performance improvement.

Next, we investigated the impact of our proposed recurrent feature alignment approach, ReMATE, by extending UA with ReMATE. From the results in Table VI, we can observe a performance improvement over both UA and UA+DMA. This performance improvement can be attributed to ReMATE's ability to align unimodal feature representations by observing the other modalities' memory banks, effectively reducing the misalignment brought about by the heterogeneous feature characteristics of different modalities. We have attained further performance gain by incorporating UA+DMA with ReMATE. Interestingly, we have noticed significant performance improvement by only increasing the recurrent feature alignment iteration in ReMATE, even without using any attention-based multimodal fusion models, such as DMA or VAT (Table VI). However, increasing the number of heads and self-attention layer in the DMA model does not help to improve the performance of the learning model. Thus, the results suggest that recurrent feature alignment plays a vital role in fusing heterogeneous multimodal representations.

Additionally, we investigated whether a Transformer based learning model can result in a performance improvement compared to our proposed recurrent fusion approach. The results in Table VI suggest that the Transformer model helps to improve the performance over the unimodal attention and deterministic multimodal attention-based feed-forward fusion baselines. However, the self-attention based Transformer model failed to outperform MAVEN in recognizing activities accurately. Additionally, we observed that using the DMA model to fuse the extracted unimodal representations from the Transformer models slightly increased the performance of Transformer models, which concatenates the extracted representations. As the Transformer model uses a self-attention approach to extract salient representation, additional attention layer (DMA) to fuse multimodal representations can not help to effectively improve the performance. In contrast, using only our proposed recurrent feature alignment approach, ReMATE, without our variation attention-based fusion approach, VAT, outperforms Transformer-based learning models. This performance improvement provides additional evidence regarding the importance of our proposed recurrent alignment approach (ReMATE) and the variational attention-based multimodal fusion approach (VAT).

Table 2.18: Ablation study of MAVEN memory length on MMAct dataset in cross-subject evaluation setting.

Memory Bank Length	F1-score (%)
10	65.85
20	73.61
40	74.89
60	74.92
100	76.76

Additionally, the experimental analysis in Table 2.17 suggests that increasing the an increment in the iteration of ReMATE results in improved performance. However, increasing the recurrent iterations results in memory and computation overhead brought about by each iteration. Apart from aligning representations using ReMATE, we also developed a lightweight variational attention-based multimodal fusion approach, VAT, which allows explicit alignment of latent feature distribution from unimodal representations. We incorporated VAT to align and fuse unimodal feature representations, which requires less recurrent iterations to achieve improved performance. The experimental results in Table 2.16 suggest that ReMATE with VAT helps MAVEN to obtain robust multimodal representation, achieving the best performance.

2.3.3.3.2 Impact of Memory Banks We investigated the impact of the length of memory banks in ReMATE on the activity recognition performance of MAVEN. We varied the memory bank length between 10 and 100. We conducted this experimental analysis on the MMAct dataset in cross-subject evaluation setting. The results of this experimental evaluation are presented in Table 2.18.

Results and Discussion: The results in Table 2.18 suggest that the increased memory bank size helps ReMATE to align unimodal features better and obtain a robust multimodal representation. We achieved the highest F1-Score of activity recognition with a memory bank length of 100. This further supports our reasoning for using memory banks to refine and temporally align cross-modal feature representations. In general, state-of-the-art learning approaches fuse multimodal representations in a feed-forward manner without an explicit feature representation alignment. These approaches do not incorporate a memory bank to align and fuse multimodal features by observing the representation from other modalities. On the other hand, ReMATE utilizes a long-term spatial-temporal memory bank to refine and temporally align cross-modal features in a fine-grained manner iteratively. This recurrent multimodal representation alignment results in improved performance compared to the state-of-the-art feed-forward-based multimodal fusion approaches.

2.3.3.3.3 Significance Analysis Following the procedure proposed by Dror, Shlomov, and Reichart [121], we conducted a significance analysis to assess the importance of MAVEN over the

Table 2.19: Significance analysis of multimodal learning models on MMAct Dataset in cross-subject evaluation setting.

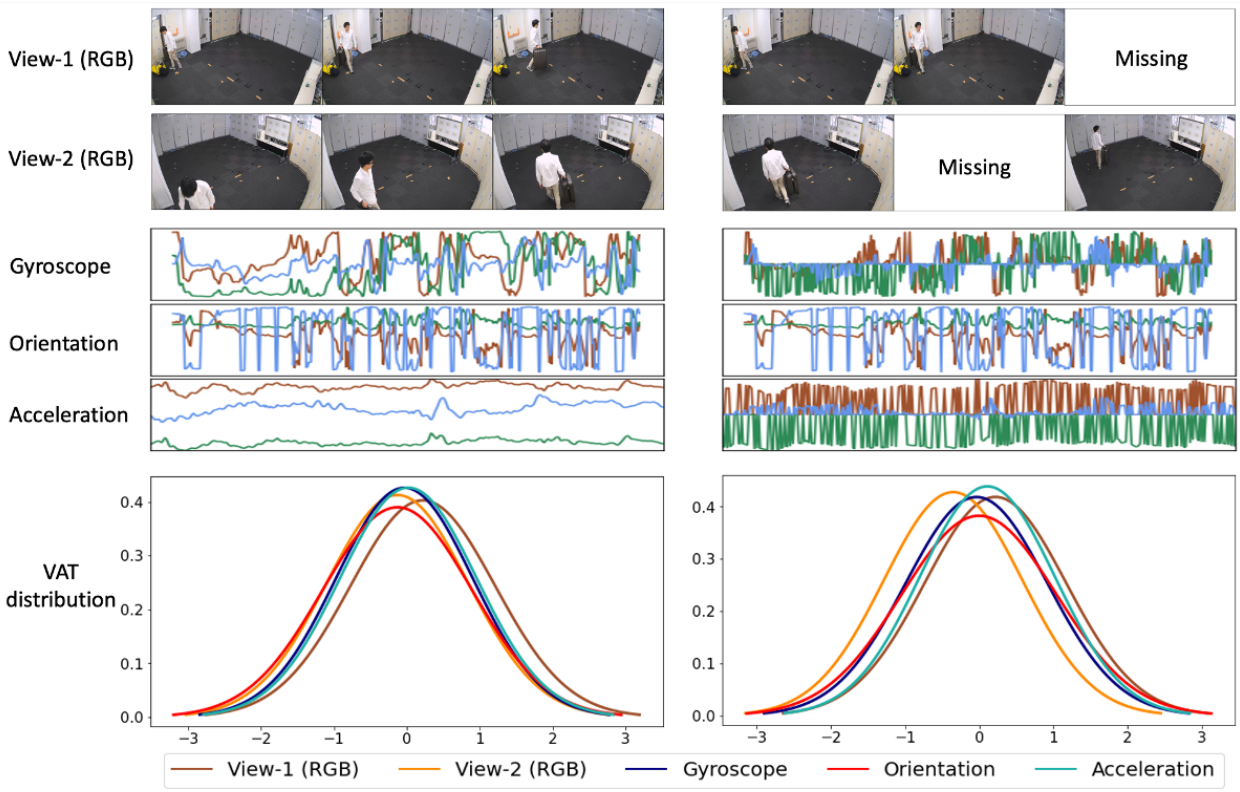
Learning Models	Average F1-Score (%)	Standard Deviation	Significant Over [§]
B1: Non-Attention	68.48	1.26	None
B2: Keyless [37]	70.52	0.98	B1 & B3
B3: HAMLET [7]	69.19	0.72	B1
B4: Multi-GAT [10]	74.66	0.51	B1-B3
B5: MuMu [11]	75.97	0.29	B1-B4
MAVEN	**76.29**	**0.47**	B1-B5

[§] We conduct significance analysis at $\alpha = 0.05$ (Following Dror et al. [121])

state-of-the-art learning models (HAMLET [7], Keyless [37], Multi-GAT [10], and MuMu [11]). We conducted this significance analysis at level $\alpha = 0.05$ on the MMAct dataset in cross-subject evaluation settings. We trained and evaluated all the multimodal learning approaches five times for the significance analysis. The results of the significance analysis are presented in Table 2.19.

Results and Discussion: In this anylsis, all the evaluated baselines use feed-forward multimodal fusion approaches to extract representation for activity recognition, whereas MAVEN uses recurrent representation alignment-based multimodal fusion approach. The results of the significance analysis in Table 2.19 suggest that MAVEN significantly outperformed $(p < 0.05)$ all the evaluated baselines. The primary difference between the baseline models and MAVEN is that these baseline models do not align unimodal representations before fusion. Moreover, all the evaluated baselines use deterministic attention approaches to fuse multimodal representations. On the other hand, MAVEN recurrently aligns unimodal feature representation before fusion which helps to improve the activity recognition performance significantly $(p < 0.05)$. Furthermore, MAVEN samples different sets of attention weights from a variational attention distribution to fuse multimodal representation compared to the deterministic point estimate of the attention weights, which is a common approach in the state-of-the-art multimodal learning models. Thus, our proposed recurrent feature alignment with variational attention-based multimodal fusion can help to extract complementary multimodal representation to attain robust activity recognition performance.

2.3.3.3.4 Qualitative Analysis We conducted a qualitative analysis of the alignment capabilities of MAVEN. For these experiments, we randomly misaligned one or more of the modalities by dropping a sub-sequence of their input and shifting the input sequence in the temporal domain. The results in Fig. 2.11 illustrates MAVEN can align and refine multimodal features by attaining aligned unimodal attention distributions. Finally, we assessed the impact of ReMATE and VAT by visualizing t-SNE embeddings (Fig. 2.12). We compared the unimodal and multimodal embeddings of four learning models: (a) MAVEN without ReMATE and VAT; (b) MAVEN without VAT;



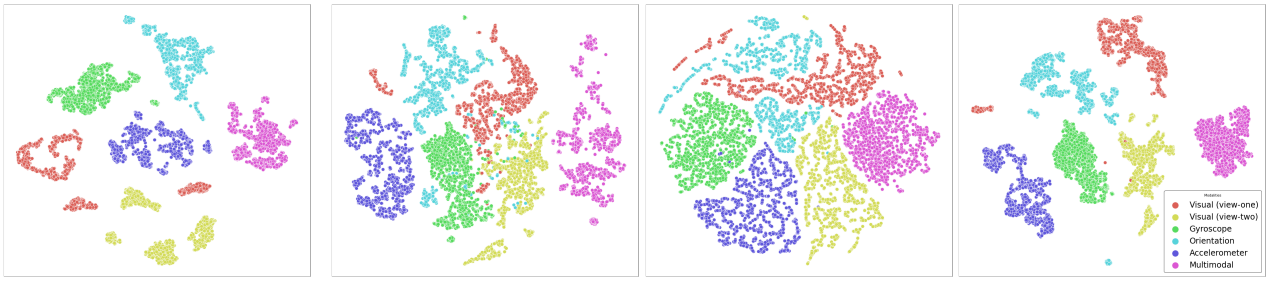
(a) Carry Light (Non-noisy data) (b) Carry Light (Noisy data)

Figure 2.11: Qualitative analysis of MAVEN's feature alignment. (a) *Carry Light* with aligned input, (b) *Carry Light* with misaligned and noisy data for all modalities, except orientation. MAVEN can align variational attention distributions for each modality, even if they are misaligned or noisy.

(c) MAVEN without ReMATE; (d) MAVEN. Additionally, we evaluated MAVEN's ability to align and refine the representation from misaligned input modalities (Fig. 2.11).

Results and Discussion: The results in Fig. 2.11(b) show that MAVEN ensures implicit feature alignment, which enforces each modality to have aligned variational attention distributions. This aligned distribution helps MAVEN to extract complementary multimodal features from noisy and misaligned data effectively. Thus, the results demonstrate MAVEN's robustness to noisy and misaligned inputs, which implies its effectiveness of MAVEN in real-world settings.

Additionally, we evaluated MAVEN's ability to align and refine the representation from different input modalities. For this, we explicitly misaligned the input modalities by performing a translation on the temporal dimension as well as dropping part of the input. The results in Fig. 2.11 illustrates the distribution of attention weights for each modality for the case of an aligned input sample (top row) as well as a misaligned input sample (bottom row). As can be observed, the



(a) MAVEN without ReMATE and VAT (b) MAVEN without VAT (c) MAVEN without ReMATE (d) MAVEN

Figure 2.12: Comparative impact of recurrent feature alignment and variational attention in MAVEN to learn robust unimodal and multimodal representations (t-SNE embeddings).

ReMATE and VAT modules of MAVEN ensures implicit alignment in the feature space, which enforces each modality to have overlapping distributions. The results further demonstrate MAVEN's robustness to misaligned inputs and underlines the efficacy of the memory banks in ReMATE to align features.

Finally, in Fig. 2.12, we visualized the manifolds of unimodal and multimodal representations obtained using MAVEN without ReMATE and VAT (Fig. 2.12(a)) and MAVEN (Fig. 2.12(b)). In Fig. 2.12(a), one can observe fractured manifolds for the visual modalities (both views), leading to sparse multimodal representation in the absence of ReMATE. Although, as shown in Fig. 2.12(b), incorporating ReMATE in MAVEN helps to cluster the unimodal and multimodal representations, the clusters are overlapped (Fig. 2.12(b)). Similarly, although incorporating only VAT in MAVEN can cluster the representations, the clusters are not well-separated (Fig. 2.12(c)). However, as shown in Fig. 2.12(d), employing ReMATE and VAT for multimodal fusion in MAVEN can better align the unimodal features, resulting in a clustered manifold that suggests that the latent space is compact and well-spaced [122]. When sampling the attention weights from these clustered manifolds for multimodal fusion, we observed that the manifold for multimodal feature space is also clustered. These findings further validate the benefit of ReMATE for aligning unimodal features and VAT for subsequent fusion in multimodal space.

2.3.3 Model Complexity Analysis

The number of parameters of MAVEN with five modalities (acceleration, gyroscope, orientation, and two visual modalities) is $25.67M$. On the other hand, the size of the best performing evaluated baseline learning models MuMu, HAMLET, and Keyless are $25.08M$, $24.97M$, and $24.90M$, respectively. The unimodal feature encoders of MAVEN, MuMu, HAMLET, and Keyless have similar model architectures, which include spatial-temporal feature encoder and self-attention model to extract salient unimodal representations. The primary differences between MAVEN and the evaluated baseline learning models are a representation alignment module (ReMATE) and multimodal

fusion models (VAT). Although baseline learning models use a multimodal fusion approach, such as multimodal attention-based fusion approach (HAMLET), these models do not use a representation alignment module. However, the number of parameters of ReMATE and VAT modules in MAVEN are $758.9K$ and $6.41K$, respectively. Thus, our proposed learning modules (ReMATE and VAT) do not increase the number of parameters of MAVEN considerably compared to the state-of-the-art multimodal learning models.

Additionally, MAVEN, with three recurrent iterations in ReMATE, takes approximately $1.20s$ to process each batch of size 2 in $RTX - 6000$ GPU. If we increase the number of iterations in ReMATE, it slightly increases the processing time. Although increasing the recurrent iterations result in improved accuracy, the experimental results suggest that the gains drop beyond recurrent iterations of more than 3. Additionally, MAVEN's use of separate unimodal feature encoders processing data independently in the distributed computing environment, which further reduces the computational time. On the other hand, MuMu, HAMLET, and Keyless models take approximately $1.21s$, $1.15s$, and $1.14s$ to execute a batch of size 2, respectively. Therefore, despite MuMu using a recurrent feature alignment approach with multiple alignment iterations, it takes a similar amount of time to process the multimodal sensor data compared to the feedforward learning models.

2.3.4 Findings

In this paper, we have presented a recurrent feature representation alignment and variational attention-based fusion approach to extract complementary multimodal features. We conducted extensive experiments to evaluate the efficacy of our proposed recurrent representation alignment-based multimodal learning approach, MAVEN, over the state-of-the-art multimodal fusion approaches. Our findings from the experimental analysis are four-folds.

First, we found that most of the state-of-the-art learning models use feed-forward multimodal fusion approaches, where the unimodal feature representations are not calibrated to align these unimodal representations before fusion. However, aligning unimodal representations before fusion can help to produce robust multimodal representation. Our experimental analysis suggests that the feed-forward fusion approaches can not ensure robust performance, specifically in the presence of noisy sensor data (Table 2.15).

Second, the experimental evaluations suggest that our recurrent feature alignment-based multimodal representation learning approach (ReMATE) helps MAVEN to extract robust representations and improve the activity recognition performance (Table 2.16 - 2.18). ReMATE allows each unimodal feature encoder to calibrate their representations by observing the representations from other modalities. MAVEN iteratively uses ReMATE to align unimodal representations. Moreover, our recurrent unimodal representation alignment approach aids in extracting complementary multimodal features, even from noisy sensor data. However, this recurrent feature alignment can not be incorporated in the feed-forward learning models to extract robust multimodal representation. To the best of our knowledge, we are the first to propose a recurrent latent representations alignment approach to extract robust multimodal representation.

Third, the experimental analysis suggests that our variational attention-based multimodal fusion approach, VAT, outperforms the deterministic attention-based multimodal fusion approaches. Specifically, our qualitative experimental analysis in Fig. 2.11 suggests that MAVEN can align the representations of the noisy modalities to the non-noisy modalities' representations to extract robust multimodal features. For example, in Fig. 2.11 we can observe that VAT can attain similar multimodal attention weights distribution in both noisy and non-noisy scenarios. The reasoning behind the distribution alignment is that our VAT explicitly aligns the latent representations from both noisy and non-noisy sensor modalities prior to fusing them by employing a KL-divergence based distribution alignment loss (Eq. 2.44). On the other hand, deterministic attention-based multimodal fusion approaches calculate a point estimate of attention weights to fuse multimodal features without explicitly aligning latent representations. Thus, these deterministic approaches can not ensure robust performance in the presence of noisy sensor data.

Finally, according to Dror, Shlomo, and Reichart [121] it is not suitable to compare the significance of various learning models using only point metric estimates, such as accuracy and F1-Score. However, most of the multimodal learning approach uses only point estimate to compare the performance. Following the procedure presented by Dror, Shlomo, and Reichart [121], we conducted the significance analysis to evaluate the efficacy of our proposed multimodal learning approach, MAVEN. Our significance analysis suggests that MAVEN significantly outperforms $(p < 0.05)$ the feed-forward fusion approaches. Thus, our extensive experimental analysis posits the significance of our recurrent feature alignment-based multimodal fusion approach to extract complementary multimodal representation over the state-of-the-art learning models.

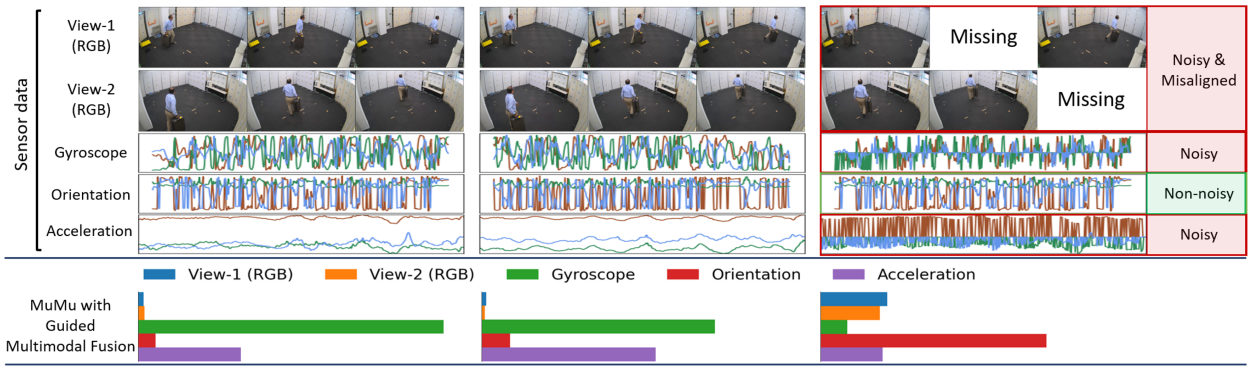
Chapter 3
MULTIMODAL AND MULTITASK MODEL FOR PERCEIVING HUMAN BEHAVIOR

Understanding human activity ensures effective human-autonomous-system collaboration in various settings, from autonomous vehicles to assistive living to manufacturing [123]–[129]. For example, accurate activity recognition could aid collaborative robots in assisting a worker by bringing tools or autonomous vehicles in requesting to take over the controls from a distracted driver to ensure safety [130], [131]. Human activity recognition (HAR) has been extensively studied by utilizing unimodal sensor data, such as visual [132]–[134], skeleton [31], [32], [135], [136], and wearable sensors [30], [137]. However, unimodal HAR methods struggle to recognize activity in various real-world scenarios for multiple reasons. First, distinct activities can be mistakenly classified as the same when relying on visual sensors [8]. For example, the activities related to carrying a light and a heavy object look similar from visual modalities; however, they have distinct physical sensor data (i.e., Gyroscope & Acceleration) (Fig. 3.1-a & b). Second, HAR algorithms relying on unimodal sensor data may fail to recognize activities when the sensor data is noisy (Fig. 3.1-c). Thus, in these cases, using multiple modalities can compensate for the weaknesses of any particular modality in recognizing an activity.

Several multimodal learning approaches have been proposed to accurately recognize human activities by fusing data from multiple sensors, such as visual, motion capture, and wearable sensors [7], [8], [12], [16], [17], [36], [138], [139]. Although these approaches work adequately in many scenarios, some crucial challenges remain in achieving robust recognition performance, particularly when data from multiple sensors are missing or misaligned.

First, disparate activity-groups require different modalities to accurately recognize activities (an activity-group consists of a set of activities, that exhibit similar characteristics). For example, Kubota et al. [9] found that data from the motion capture system helps to recognize gross-motion activities involving arm and leg movements (e.g., walking). Moreover, they found that data from wearable sensors helps to recognize fine-grained motion activities involving hand or finger movements (e.g., grasping). Thus, if a learning model can exploit the characteristics of activity-groups while extracting the multimodal representations, then that model can extract robust representation to improve HAR performance. Moreover, in many existing datasets, activities are grouped into major categories based on shared characteristics [8], [9], [97], [140]. For example, [8] grouped daily human activities into three groups: complex (e.g., carrying, talking), simple (e.g., kicking, jumping), and desk (e.g., using PC). Surprisingly, apart from grouping the activities, these labels of auxiliary activity-groups have not been utilized in extracting multimodal representations.

Second, most existing multimodal learning approaches assume non-noisy and time-aligned multimodal sensor data during training and testing phases. These assumptions limit the applicability of the existing multimodal learning approaches in real-world settings, as the presence of



(a) Carry-Light (Non-noisy data) (b) Carry-Heavy (Non-noisy data) (c) Carry-Heavy (Noisy data)

Figure 3.1: (a) Carry-Light and (b) Carry-Heavy activities have similar visual features. (a & b) However, these activities have distinct gyroscope and acceleration data. (a & b: bottom-row) Our proposed method, MuMu, utilizes a guided multimodal fusion approach to appropriately prioritize salient modalities (Gyroscope and Acceleration, in this case) while extracting multimodal representations. (c) MuMu can adaptively adjust attention weights when data is noisy. For example, MuMu pays more attention to the non-noisy data (Orientation) than the noisy data (Gyroscope and Acceleration) or misaligned data (View-1 & 2). (Data samples are drawn from MMAct dataset [8]).

misaligned and noisy sensor data is not uncommon due to occlusion and sensor noises (Fig. 3.1-c). Thus, we need to develop and evaluate the multimodal learning approaches in the presence of noisy and misaligned sensor data to ensure their applicability in real-world settings.

To address the aforementioned challenges, we propose a novel Cooperative Multitask Learning-based Guided Multimodal Fusion Approach (MuMu) for HAR. In MuMu, we have designed a multitask learning approach that involves learning two cooperative tasks: an auxiliary and a target task. First, MuMu extracts activity-group-specific features for activity-group recognition (auxiliary task). Second, the activity-group-specific features direct our Guided Multimodal Fusion Approach (GM-Fusion) to extract robust multimodal representations for recognizing activities (target task). Here, both tasks work cooperatively, where the auxiliary task guides the target task to extract complementary multimodal representations appropriately.

We compared the performance of MuMu to several state-of-the-art HAR algorithms on three multimodal activity datasets (MMAct [8], UTD-MHAD [97] and UCSD-MIT [9]). The results from our extensive experimental evaluations suggest that MuMu outperforms all the state-of-the-art approaches in all evaluation conditions. MuMu achieved an improvement of 4.45% and 3.61% (F1-score) on the MMAct dataset for the cross-subject and cross-session evaluation conditions, compared to the state-of-the-art approaches, respectively. Additionally, MuMu achieved an improvement of 6.86% and 2.48% (top-1 accuracy) on the UCSD-MIT and the UTD-MHAD datasets for leave-one-subject-out evaluation settings, compared to the state-of-the-art approaches, respec-

tively. Furthermore, our qualitative analysis of multimodal attention weights suggests that our proposed guided multimodal fusion approach can appropriately prioritize the modalities while extracting complementary representations, even in the presence of noisy and misaligned sensor data (Fig. 3.1 & 3.4). Moreover, our extensive ablation study suggests that our proposed approach significantly outperforms the baseline multimodal learning approaches ($p < 0.05$), which do not use guided fusion.

3.1 Cooperative Multitask Learning-Based Guided Multimodal Fusion

3.1.1 Problem Formulation

We define a cooperative multitask learning problem, which involves learning the auxiliary and the target tasks cooperatively for multimodal fusion. Similar to the multi-class activity recognition problem, we aim to recognize a set of K activities, $A = (A_1, \ldots, A_K)$, by extracting multimodal representations (X^c) from M heterogeneous modalities, $X^r = (X^r_1, \ldots, X^r_M)$ (r stands for raw feature). We have termed this activity recognition ($A_i \in A$) as the *target task*.

Activity datasets defined activity-group in various ways. For example, UCSD-MIT uses human motion to define activity-group (gross & fine), whereas the MMAct dataset uses the complexity of the activities (complex, simple & desk). As different activity-groups share disparate characteristics, they require different modalities for recognizing activities [9]. Thus, we divide the activity set A into N activity-groups (G), where $G = (G_1, \ldots, G_N)$. Here, each activity-group (G_i), consists of J_i unique activities that share similar characteristics, where $G_i = (A^i_1, \ldots, A^i_{J_i})$, and $A^i_j \in A$. We have termed the activity-group recognition ($G_i \in G$) as the *auxiliary task*.

3.1.2 Approach Overview

Our proposed Cooperative Multitask Learning-based Guided Multimodal Fusion Approach (MuMu) consists of three learning modules (Fig. 3.2):

- **Unimodal Feature Encoder (UFE)** encodes modality-specific spatial-temporal features.
- **Auxiliary Task Learning (ATL)** Module extracts activity-group-specific multimodal representations.
- **Target Task Learning (TTL)** Module utilizes the activity-group-specific features from the auxiliary task as prior information to appropriately fuse and extract multimodal representations for activity recognition.

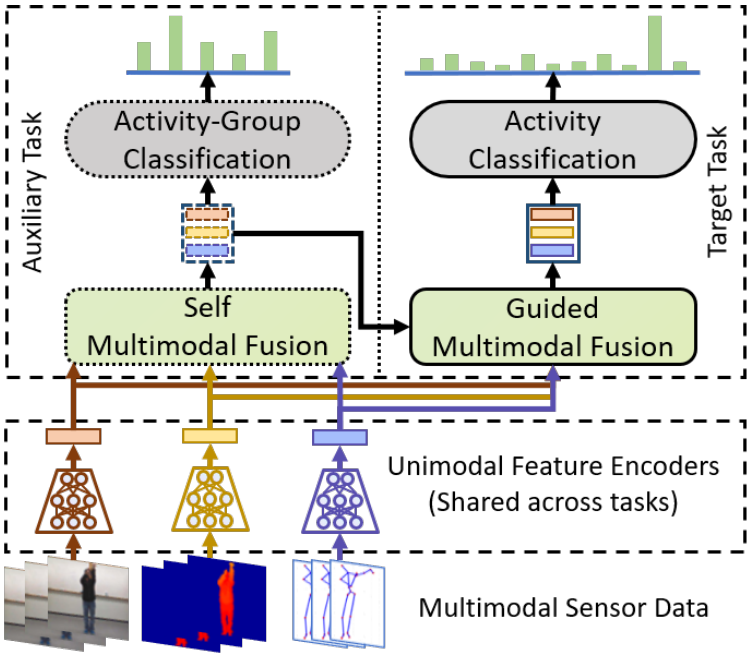


Figure 3.2: MuMu: Cooperative Multitask Learning-based Guided Multimodal Fusion Approach. The Unimodal Feature Encoder encodes unimodal spatial-temporal features. The Auxiliary Task module fuses the unimodal features to extract the activity-group-specific features. The activity-group features guide the Target Task module to fuse and extract complementary multimodal representations by employing a Guided Multimodal Fusion Approach. We have designed a multitask learning loss for end-to-end training.

3.1.3 UFE: Unimodal Feature Encoder

We have adopted the Unimodal Feature Encoder (UFE) architecture from the work by HAMLET model [7]. In our implementation, UFE independently encodes salient unimodal features of each modality $m \in M$ in four steps. First, UFE segments the raw unimodal features and produces $X^r_m = (x^r_{m,1}, x^r_{m,2}, \ldots, x^r_{m,S_m}) \in \mathbb{R}^{B \times S_m \times D^r_m}$, where B is the batch size, S_m is the segment size, and D^r_m is the raw feature dimension of the modality m. Second, UFE encodes the spatial features of each segment of modality $m \in M$. Third, UFE utilizes an LSTM, a variant of recurrent neural network, to encode unimodal spatial-temporal features. Fourth, a self-attention approach has been employed to extract salient unimodal features, $X^u = (x^u_1, x^u_2, \ldots, , x^u_M) \in \mathbb{R}^{B \times M \times D^u}$, from the extracted spatial-temporal features (D^u is the unimodal (u) feature embedding size). Instead of utilizing a resource intensive multi-head self-attention approach [94], which was used by HAMLET model [7], in this work, we have adopted a lightweight self-attention model from Keyless model [37]. MuMu uses the unimodal features, X^u, in the subsequent learning modules to produce robust multimodal representations.

3.1.4 ATL: Auxiliary Task Learning Module

In the auxiliary task learning step, MuMu fuses the unimodal features to extract activity-group-specific multimodal representation for classifying the activity-groups in two steps:

3.1.4 Self Multimodal Fusion Approach (SM-Fusion):

We have designed a Self Multimodal Fusion Approach (SM-Fusion) for extracting activity-group-specific salient features. SM-Fusion assigns attention weight (α_m) to each modality for fusing unimodal features, X^u, and extracting multimodal auxiliary representation, X^{aux}. The attention weight, α_m, is calculated in the following way,

$$\gamma_m = (W^{aux})^T X^u_m \tag{3.1}$$

$$\alpha_m = \frac{exp(\gamma_m)}{\sum_{m \in M} exp(\gamma_m)} \tag{3.2}$$

Here, W^{aux} is a learnable parameter. We have utilized a 1D-CNN with a filter size of 1 to calculate α_m. Finally, this weight is used to fuse the unimodal features and extract multimodal auxiliary representation, X^{aux}:

$$X^{aux} = \sum_{m \in M} \alpha_m X^u_m \tag{3.3}$$

3.1.4 Activity-Group Classification:

The auxiliary representation, X^{aux}, is passed through a auxiliary task learning network, F^{aux}, to classify the activity-group:

$$y^{aux} = F^{aux}(X^{aux}) \tag{3.4}$$

3.1.5 TTL: Target Task Learning Module

In MuMu, we have designed a target task to extract multimodal representations and classify activities in two steps. First, MuMu uses activity-group features from the auxiliary task to direct our proposed Guided Multimodal Fusion Approach (GM-Fusion) to extract multimodal representations. Because activity-group features can help to prioritize the salient modalities to extract multimodal representations appropriately. Second, MuMu uses fused representations to classify the activities. In MuMu, the auxiliary and the target tasks work cooperatively to extract complementary multimodal representations for recognizing activities accurately.

3.1.5 Guided Multimodal Fusion Approach (GM-Fusion):

GM-Fusion uses the extracted activity-group-specific features from auxiliary task as prior information, X^{aux}, to extract multimodal representations for activity recognition.

First, GM-Fusion projects the extracted unimodal features, X^u, to produce unimodal key (K^u) and value (V^u) feature vectors in the following way:

$$K^u = X^u W^K; V^u = X^u W^V \tag{3.5}$$

Here, W^K and W^V are learnable parameters. These unimodal key and value vectors are used to extract the multimodal representation. Second, GM-Fusion projects multimodal auxiliary representation, X^{aux}, to produce auxiliary query feature vector (Q^{aux}).

$$Q^{aux} = X^{aux} W^Q \tag{3.6}$$

Here, W^Q is a learnable parameter. This auxiliary query feature vector (Q^{aux}) is used as a prior to extract complementary multimodal representation, X^c, by utilizing the unimodal key (K^u) and value (V^u) feature vectors:

$$X^{c'} = \sigma\left(\frac{Q^{aux} K^{u^T}}{\sqrt{D^u}}\right) V^u \tag{3.7}$$

$$X^c = W^o X^{c'} \tag{3.8}$$

Here, W^o is a learnable projection parameter.

3.1.5 Activity Classification:

Multimodal representation, X^c, is concatenated with activity-group-specific features, X^{aux}, for activity classification. X^c is passed through a target task learning network, F^t, to classify the activities:

$$X^f = W^f[X^c; X^{aux}] \tag{3.9}$$

$$y^t = F^t(X^f) \tag{3.10}$$

Here, W^f is a learnable projection parameter.

3.1.6 Multitask Learning Loss

We have designed a multitask learning loss for end-to-end training of MuMu. This loss is used to train the auxiliary and the target tasks jointly. First, we use cross-entropy auxiliary loss, L^{aux}, to

train the auxiliary task for activity-group classification. L^{aux} enforces the auxiliary task branch to learn the activity-group-specific multimodal representations.

$$L^{aux}(y^{aux}, \hat{y}^{aux}) = \frac{1}{B} \sum_{i=1}^{B} y_i^{aux} \log \hat{y}_i^{aux} \tag{3.11}$$

Second, we calculate the cross-entropy loss, L^t, to train the target task for activity classification. This loss ensures that the target task learns the robust multimodal representations for activity recognition.

$$L^t(y^t, \hat{y}^t) = \frac{1}{B} \sum_{i=1}^{B} y_i^t \log \hat{y}_i^t \tag{3.12}$$

Finally, the auxiliary and target task losses are combined for end-to-end training of MuMu:

$$loss = L^t(y^t, \hat{y}^t) + \beta^{aux} L^{aux}(y^{aux}, \hat{y}^{aux}) \tag{3.13}$$

Here, β^{aux} is the weight of auxiliary task learning loss.

3.2 Experimental Setup

3.2.1 Datasets

We evaluated the performance of our proposed approach, MuMu, by applying it on three multimodal activity datasets: UCSD-MIT [9], UTD-MHD [97] and MMAct [8].

MMAct dataset contains 37 activities which are categorized into 3 groups: 16 complex activities (e.g., carrying, exiting), 12 simple activities (e.g., kicking, talking on the phone, jumping), 9 desk activities (e.g., using PCs, sitting). Twenty people performed each activity five times, resulting in $37k$ data samples. The MMAct dataset uses data from seven modalities: four RGB views, acceleration, gyroscope, and orientation. We used data from two opposing RGB views, acceleration, gyroscope, and orientation modalities to train and test. MMAct dataset contains visually occluded data samples, which allows evaluating the effectiveness of HAR approaches for real-world settings.

UCSD-MIT dataset contains nine automotive and block assembly activities from 2 groups. The gross activity-group contains four activities (e.g., walking, receiving part, and attaching part), and the fine activity-group contains five activities (e.g., palmar grab, pincer grab, and ulnar pinch grab). Five people performed each activity five times. UCSD-MIT dataset uses data from three modalities: 3D joint positions from motion capture system, sEMG, and IMUs data from wearable sensors.

UTD-MHAD contains 27 activities which are categorized into 4 groups: 9 hand gesture (e.g., draw X, draw circle), 9 sports (e.g., bowling, tennis serve), 5 daily (e.g., door knock, sit to stand), and 4 training exercises (e.g., lunge, squat). Eight people performed each activity five times. UTD-MHAD uses data from four modalities: RGB, depth, skeleton, and physical sensors.

Table 3.1: Cross-subject performance comparison (F1-Score) of multimodal learning methods on MMAct dataset

Method	F1-Score (%)
SMD [109]	63.89
Student [8]	64.44
Multi-Teachers [8]	62.67
MMD [8]	64.33
MMAD [8]	66.45
HAMLET [7]	69.35
Keyless [37]	71.83
MuMu [11]	**76.28**

Table 3.2: Cross-session performance comparison (F1-Score) of multimodal learning methods on MMAct dataset

Method	F1-Score (%)
SVM+HOG [110]	46.52
TSN (RGB) [111]	69.20
TSN (Optical-Flow) [111]	72.57
MMAD [8]	74.58
TSN (Fusion) [111]	77.09
MMAD (Fusion) [8]	78.82
Keyless [37]	81.11
HAMLET [7]	83.89
MuMu [11]	**87.50**

3.2.2 Learning Architecture Implementation

We segmented the data from visual modalities (RGB and depth) with a window size of 1 and a stride of 3. For the data from other sensor modalities, we used a window size of 5 and a stride of 5. To encode segmented spatial features, we used ResNet-50 model [98] for data from visual modalities (RGB and depth) and Co-occurrence approach [93] for data from other sensors modalities (sEMG, Acceleration, Gyroscope, and Orientation). The unimodal feature of each modality is encoded to 128 sized feature embedding. We used two fully connected layers with Re-LU activation after the first layer for activity-group classification in auxiliary task learning. We used similar task learning architecture for the activity classification in target task learning.

Table 3.3: Performance comparison (F1-Score) of multimodal learning methods on UCSD-MIT dataset [9].

Learning Methods	Merge Types	F1-Score (%)
Non-Attention	SUM	52.35
	CONCAT	50.92
HAMLET [7]	SUM	50.04
	CONCAT	48.26
Keyless [37]	SUM	51.68
	CONCAT	54.48
MuMu [11]	-	**61.34**

3.3 Experimental Results and Discussion

3.3.1 Comparison with Multimodal Approaches

Results: We evaluated MuMu's performance by comparing it against the state-of-the-art HAR approaches on three datasets: MMAct, UTD-MHAD, and UCSD-MIT. For MMAct dataset, we followed originally proposed cross-subject and cross-session evaluation settings and reported F1-scores (Tables 3.1 & 3.2). The results suggest that MuMu outperforms state-of-the-art approaches on both cross-subject and cross-session evaluation settings with improvements of 4.45% and 3.61% in F1-score, respectively. For UTD-MHAD and UCSD-MIT datasets, we followed leave-one-subject-out cross-validation and reported top-1 accuracies (Tables 3.4 & 3.3). The results suggest that MuMu outperforms the best performing baselines with improvements of 6.86% and 2.48% in top-1 accuracy on UCSD-MIT and UTD-MHAD datasets, respectively.

Discussion: The experimental results on these activity datasets (Tables 3.1, 3.2, 3.4 & 3.3) suggest that MuMu outperforms all the state-of-the-art approaches in all evaluation conditions. Moreover, the results indicate that attention-based HAR methods (i.e., MuMu, Keyless [37] and HAMLET [7]) outperform Non-Attention-based methods (i.e., PoseMap [107] and TSN [111]). Unlike MuMu, the other attention-based methods do not consider the activity-group-specific information to extract multimodal representations. In our implementation, MuMu utilizes the activity-group-specific information to extract complementary multimodal representations by utilizing our proposed Guided Multimodal Fusion approach (GM-Fusion). GM-Fusion allows the prioritization of different modalities based on the activity-group information extracted by the auxiliary task learning module. Thus, the experimental results posit that incorporating activity-group information allows the extraction of complementary multimodal representations effectively to improve the HAR accuracy.

Although state-of-the-art multimodal HAR approaches show comparatively better performance on cross-session evaluation settings (Tables 3.2 & 3.4), the performance degrades on challenging

Table 3.4: Performance comparison (top-1 accuracy) of multimodal learning methods on UTD-MHAD dataset.

Method	Accuracy (%)
MHAD [97]	79.10
SOS [105]	86.97
S^2DDI [104]	89.04
DCNN [102]	91.20
Keyless [37]	92.67
MCRL [106]	93.02
PoseMap [107]	94.51
HAMLET [7]	95.12
MuMu	**97.60**

cross-subject evaluation conditions for all evaluated baselines (Tables 3.1 & 3.3). The performance degrades because MMAct and UCSD-MIT datasets contain data samples that enforce the utilization of the wearable sensors to recognize activities accurately, where the wearable sensor data vary considerably across subjects (see Fig. 3.1). To address this challenge, MuMu utilizes activity-group features to guide GM-Fusion to extract salient multimodal representations for recognizing activities accurately. On the other hand, state-of-the-art approaches fused unimodal features without considering activity-group information. Additionally, in the cross-subject evaluation conditions, MuMu outperforms the F1-score of state-of-the-art approaches on MMAct and UCSD-MIT datasets with an improvement of 4.45% and 6.86%, respectively. These performance improvements indicate that MuMu can generate robust multimodal representation by prioritizing the salient modalities than other approaches.

3.3.2 Impact of Supplementary Modalities

To investigate whether additional modalities help to improve the performance of learning models, we evaluated the performance of MuMu and two baseline approaches (Keyless [37]) and HAMLET [7]) with various combinations of modalities. We conducted this study on the UTD-MHAD dataset with RGB, Depth, Skeleton, Physical sensors modalities. The experimental results suggest that MuMu outperformed the evaluated baselines on all the combinations of modalities tested (see Table 3.5).

Results & Discussion: In Table 3.5, the results suggest that incorporating additional modalities helps MuMu to improve the HAR accuracy. However, additional modalities do not always improve the performance of two baselines. For example, incorporating the depth modality degrades the accuracy of the baseline methods, whereas the HAR accuracy of MuMu improves slightly with this additional modality.

Table 3.5: Performance comparison (Accuracy %) of the impact of modality changes on UTD-MHAD dataset. R: RGB, D: Depth, S: Skeleton, P: Physical Sensors.

Learning Methods	Modality Combinations		
	R+S	R+S+P	R+D+S+P
Keyless [37]	90.20	92.67	83.87
HAMLET [7]	95.12	91.16	90.09
MuMu	**96.10**	**97.44**	**97.60**

The performance of the baselines degrades, as additional modalities may not provide salient information to recognize a set of activities accurately. For example, visual modality may not provide salient information for gesture recognition (e.g., wave, swipe), whereas physical sensors can help recognize those activities accurately. The baseline methods either concatenated or used a self-attention approach to fuse unimodal features without considering the characteristics of activity-group, which results in performance degradation with supplementary modalities. However, MuMu uses activity-group information from the auxiliary task to guide the target task for prioritizing and fusing the additional modalities to extract complementary multimodal representations for recognizing activities accurately. Therefore, it is essential to prioritize the salient modalities for extracting robust representation to recognize activities accurately.

3.3.3 Impact of Noisy Modalities

We conducted both quantitative and qualitative experiments to evaluate the performance of MuMu and three baselines (Non-Attention, HAMLET, and Keyless) in the presence of noisy and misaligned sensor data. We developed the Non-Attention method for evaluation purposes, where we extract unimodal features using CNN+LSTM model without using an attention mechanism. The extracted unimodal features are concatenated to classify activities.

We conducted this study in cross-subject evaluation setting on MMAct dataset with two visual modalities (RGB View 1 & 2) and three non-visual modalities (Gyroscope, Orientation & Acceleration). We randomly selected either visual or non-visual modalities with 50% probability and then dropped raw features to introduce noise. The quantitative and qualitative experimental results are presented in Table 3.6 and Fig 3.4, respectively.

Results & Discussion: The experimental results suggest that MuMu outperforms the evaluated baselines in the presence of noisy data (Table 3.6). In MuMu, our proposed Guided Multimodal Fusion Approach (GM-Fusion) appropriately prioritizes the modalities and extracts the robust multimodal representation from noisy sensor data for accurate activity recognition. However, the baseline multimodal learning approaches either use Non-Attention or self-attention based multimodal fusion, which may not effectively extract complementary multimodal representations.

Additionally, the qualitative results of multimodal attention visualization (Fig. 3.4-Bottom row)

Table 3.6: Performance comparison (F1-Score %) of the impact of noisy data on MMAct dataset. Visual: RGB (View 1 & 2), Non-visual: Gyroscope, Orientation & Acceleration.

Learning Methods	No Noisy Modality	Noisy Modalities	
		Visual	Non-Visual
Non-Attention	68.29	66.30	66.02
HAMLET [7]	69.35	64.10	67.57
Keyless [37]	71.83	67.94	68.29
MuMu	**76.28**	**74.22**	**73.78**

indicate the same phenomenon that MuMu can prioritize the salient modalities to extract complementary multimodal representations from noisy and misaligned sensor data. For example, although the gyroscope and acceleration data provide distinctive features for carry-heavy activity, MuMu adjusts the multimodal attention weights when we introduce noise in those modalities (Fig. 3.4-Bottom row), by paying more attention to the non-noisy modality (Orientation) and less attention to noisy modalities (Gyroscope and Acceleration), which contribute to better HAR performance on noisy data (Table 3.6). In Fig. 3.4-Center row, it can be observed that HAMLET, which uses a self-attention based fusion approach, increased the attention weight to the noisy sensor data (i.e., Acceleration in Fig 3.4-Right) compared to the attention weight assigned on the non-noisy data samples (Fig 3.4-Left). These qualitative results indicate that self-attention based fusion may not appropriately prioritize the noisy sensor data to extract robust multimodal representations (Fig. 3.4-Center row), which also reflects in the quantitative results in Table 3.6.

3.3.4 Ablation Study and Significance Analysis

To investigate the importance of various modules of MuMu, we developed three single-task-based baseline models by removing the auxiliary task learning branch in MuMu (Fig. 3.2). The Non-Attention model (B1) does not employ any attention approach in extracting unimodal or fusing multimodal features. The Unimodal Attention model (B2) employs an attention approach to extract unimodal features and concatenate multimodal features (similar to Keyless [37]). The Unimodal + Multimodal Attention model (B3) uses an attention approach to extract unimodal and fuse multimodal features (similar to HAMLET [7]). We trained and tested all these baselines and MuMu five times with different initialization of the learning parameters. Additionally, we conducted the significance analysis at level $\alpha = 0.05$ by following the procedure proposed by [121]. We conducted this experimental analysis on MMAct dataset in cross-subject evaluation setting.

Results and Discussion: The experimental results in Table 3.7 suggest that the baseline B3, which uses an attention approach to prioritize the modalities, fails to outperform B2 significantly. Here, B2 uses the attention approach only to extract unimodal and concatenate the multimodal features. These results indicate that how a multimodal learning approach fuses the information is

Table 3.7: Ablation study of MuMu components on MMAct Dataset.

Model Type	Learning Models	Average F1-Score (%)	Standard Deviation	Significant Over §
Single Task	B1	68.48	1.26	None
	B2 †	70.52	0.98	B1 & B3
	B3 †	69.19	0.72	B1
Multitask	**MuMu** *	**75.97**	**0.29**	B1, B2 & B3

B1: Non-Attention, B2: Unimodal Attention, B3: Uni + Multimodal Attention
† Self-Attention based Multimodal Fusion, * Guided Multimodal Fusion
§ We conduct the significance analysis at $\alpha = 0.05$ (Following Dror et al. (2019))

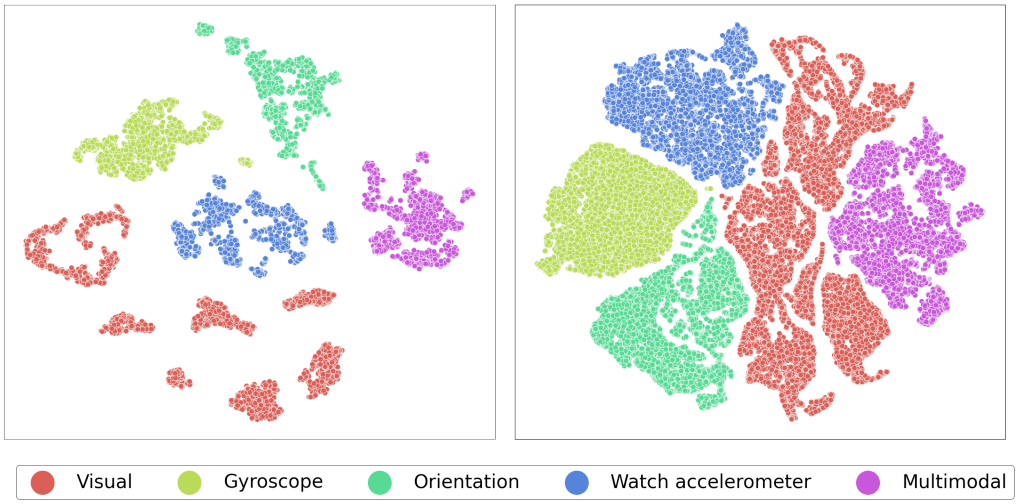


Figure 3.3: The t-SNE visualization of unimodal and multimodal representations. (Left) HAMLET with Self-Attention based Fusion, (Right) MuMu with Guided Multimodal Fusion.

crucial in improving the HAR performance.

Moreover, the experimental results in Table 3.7 indicate that MuMu significantly outperforms all the baseline models and improves the HAR accuracy. The primary difference between MuMu and the baseline models is that MuMu uses activity-group features to guide the target task for extracting multimodal representations. Thus, this experimental analysis indicates that MuMu, with the help of our guided multimodal fusion approach, can appropriately fuse multimodal features to improve the HAR accuracy significantly.

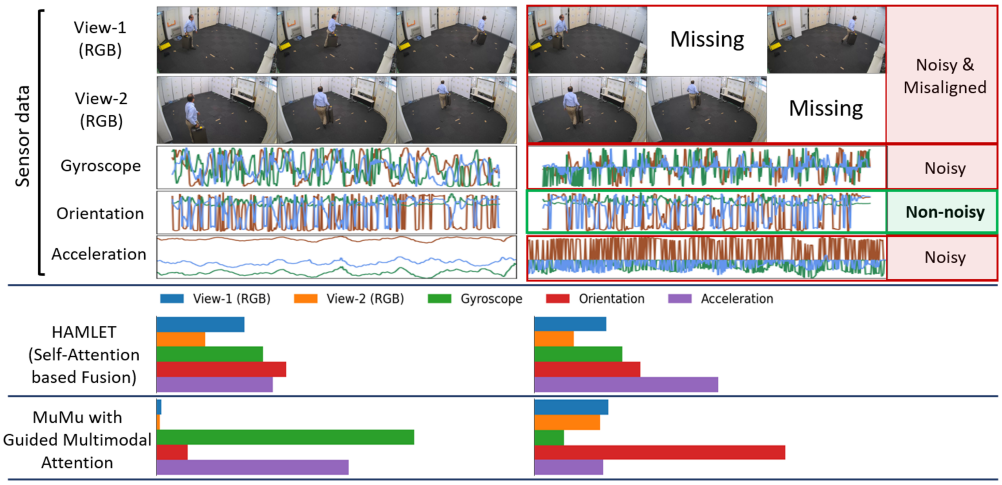


Figure 3.4: Comparative impact of guided multimodal attention in MuMu to extract complementary multimodal representations from noisy sensor data (Multimodal attention weights visualization).

3.3.5 Qualitative Analysis

We conducted two qualitative analyses to evaluate the effectiveness of our guided multimodal fusion approach. First, we visualized the attention weights to evaluate whether MuMu can prioritize the salient modalities (Fig. 3.1 & 3.4). Second, we visualized t-SNE embeddings of unimodal and multimodal representations obtained using MuMu (Fig. 3.3-Right) and HAMLET with self-attention based fusion [7] (Fig. 3.3-Left). We conducted these studies on the MMAct dataset in cross-subject evaluation setting.

Attention Visualization: Our experimental analysis (Fig. 3.1 & 3.4) suggests that appropriately prioritizing the relevant modalities aids in improved HAR performance. The results in Fig. 3.1-a & b indicate that MuMu can appropriately prioritize the salient modalities (Gyroscope and Acceleration) in extracting complementary representations to distinguish visually similar activities (i.e., carry-light and carry-heavy). Additionally, when the data from these modalities are noisy, MuMu adjusts the attention weights to the non-noisy modalities (i.e., visual and orientation) to extract complementary multimodal representations (Fig. 3.4). These results indicate that MuMu can adjust attention weights based on the extracted unimodal features to produce complementary representations. On the other hand, the self-attention based multimodal fusion approach can not appropriately prioritize the relevant modalities (Fig. 3.4), which results in performance degradation (Table 3.7).

Feature Visualization (t-SNE):

In Fig. 3.3-(Right), t-SNE visualization of unimodal and multimodal features suggests that

MuMu with guided multimodal fusion approach can cluster the unimodal and multimodal features. Therefore, the results indicate that MuMu can extract salient unimodal features to produce complementary multimodal representations. On the other hand, in Fig. 3.3-(Left), the results suggest that HAMLET with self-attention based multimodal fusion [7] introduces sparse clusters, which may indicate self-attention based multimodal fusion approach may not extract salient multimodal features for activity recognition (Fig. 3.3 & 3.4).

3.4 Broader Impact

The broader impact of our work lies in its potential to significantly advance the field of human activity recognition (HAR) and its many applications. We proposed a robust model to the challenges posed by the heterogeneous characteristics of data from multimodal sensors and disparate human activities.

In healthcare, accurate HAR can be crucial for patient monitoring, early detection of health issues, and personalized treatment plans. For instance, MuMu could be used to monitor the activities of elderly individuals living alone, providing valuable data for fall detection systems or for tracking the progression of conditions like Parkinson's disease. In rehabilitation, it could help track a patient's recovery and adapt exercises to their current capabilities.

Moreover, in the realm of sports and fitness, MuMu could be used to analyze athletes' performance, providing detailed feedback that could help them improve their techniques or prevent injuries. For everyday fitness enthusiasts, it could be used in wearable devices to provide more accurate activity tracking and personalized workout recommendations.

In Human-Computer Interaction (HCI), accurate HAR can enable more intuitive and responsive interfaces. For example, MuMu could be used in virtual or augmented reality systems to allow users to interact with the virtual environment through natural movements. It could also be used in smart home systems to automate user activity responses.

Furthermore, in surveillance and security, MuMu could help to detect unusual or suspicious activities, enhancing the capabilities of security systems and contributing to safer public spaces. Additionally, in industrial settings, MuMu could be used to monitor workers' activities, helping to ensure safety protocols are followed and identifying areas where efficiency could be improved.

In addition to these direct applications, our work could also stimulate further research in multimodal learning and fusion techniques, potentially leading to new methods and applications. However, it's important to note that using HAR technologies also raises important ethical and privacy considerations, which must be carefully addressed as these technologies are developed and deployed.

3.5 Limitations

While our proposed MuMu approach has demonstrated promising results in human activity recognition (HAR) using multimodal sensor data, several limitations to our study can be addressed in

future work to ensure robust performance in diverse settings.

Our approach relies heavily on the quality of the sensor data. While we have shown that MuMu is robust to noisy and misaligned sensor data, its performance may be significantly affected by extreme noise or severe misalignment, which were not fully explored in this study. For example, if the sensors are being used outdoors, they may be exposed to harsh environmental conditions such as heavy rain, snow, or high winds. These conditions can introduce extreme noise into the sensor data. For example, a visual sensor might have difficulty accurately capturing images in a heavy rainstorm due to water droplets on the lens or rapid changes in lighting conditions. Moreover, If the sensors themselves malfunction, they may produce extremely noisy data. This could be due to hardware issues, software bugs, or problems with the power supply. For example, a wearable sensor might start producing erratic data if its battery is running low or if it's experiencing a hardware failure. Furthermore, sensors can sometimes pick up interference from other electronic devices, which can introduce extreme noise into the data. For example, a wearable sensor might pick up electromagnetic interference from a nearby smartphone or Wi-Fi router. Additionally, if the person wearing the sensors is making abrupt or unusual movements, this could introduce extreme noise into the data. For example, if the person falls down or starts dancing suddenly, the sensor data might become very noisy and difficult to interpret.

Three datasets used in this study were chosen for their diversity and relevance to the field of human activity recognition (HAR). However, they may not encompass the full spectrum of human activities or the variety of contexts in which these activities occur. The datasets may not include all types of human activities. For instance, they might focus on common activities like walking, running, or sitting, but not include less common or more complex activities like rock climbing or performing manual labor. As a result, it's unclear how well MuMu would perform when applied to these untested activities.

Heterogeneity of Sensor Data: Although MuMu is designed to handle heterogeneous sensor data, the current study mainly focused on visual, non-visual, and wearable sensors. The performance of MuMu when applied to other types of sensors, such as audio data.

Despite these limitations, our work represents a significant step forward in the field of human activity recognition (HAR) using multimodal sensor data. The proposed MuMu approach introduces a novel cooperative multitask learning-based guided multimodal fusion technique, which has demonstrated superior performance compared to existing state-of-the-art methods on three diverse activity datasets. Future work should aim to address these limitations and further improve the performance and applicability of the MuMu approach.

Chapter 4
MULTIMODAL REFERRING EXPRESSION DATASETS AND BENCHMARKS

Natural communication forms of humans are inherently multimodal with verbal and nonverbal (gestures and gaze) signals [24], [55]–[57], [141]. People use multimodal cues with their interaction partners for the joint focus of attention on salient objects and events, specifically when they share a physical space in an environment [24], [57], [141]–[145]. As humans use multimodal communication forms for interactions, we need AI-driven agents interacting with us to understand multimodal referring expressions to generate seamless interactions [24], [126], [128], [146].

Comprehending referring expressions has been generally studied in the form of the *spatial relation grounding task* [20], [21], [26], [27], [45]–[54]. This task involves identifying whether the verbal utterance of the spatial relationships between objects holds in the visual scene [20], [21]. However, the exclusion of nonverbal signals in the model makes the problem different from how people interact naturally in shared physical spaces, as people start to use multimodal signals very early in their developmental phase [55]–[63]. To address this gap, in this work, we have designed an *embodied spatial relation grounding task*, which involves identifying whether a person is verbally and nonverbally (pointing gesture and gaze) referring to the same objects in the visual scene. This task can help develop learning frameworks to understand multimodal referring expressions in embodied settings.

A few datasets have been developed to capture embodied multimodal referring expressions, which involve referring an object using verbal utterances and nonverbal cues (pointing gesture and gaze) [24], [64]. However, these datasets have several crucial limitations. The primary limitation of existing datasets is that the nonverbal interactions are captured solely from an exocentric perspective (*exo, ego, and top* view denotes perspectives from an actor, the observer, and overhead, respectively (Fig. 4.1)). As comprehending embodied referring expression requires perspective-taking, which is the awareness of the actor's and observer's point of view in shared interactions, the lack of perspective-awareness in these datasets can degrade the model's performance. Additionally, multiple views can help identify the referred object, which may be partially occluded from one view but visible from another. Moreover, in human-human interactions, learning perspective is used innately to attend to salient parts of interactions. Let's assume an actor is requesting an observer verbally to "pick up the left apple" (Fig. 4.1). This verbal expression can be interpreted differently from different perspectives, where the "left apple" from the exo view can be interpreted as the "right apple" from the ego view. Learning where the actor is looking and pointing can help identify the appropriate object in these scenarios. These data samples with multiple views enable the model to learn perspective-taking to ensure seamless and natural interactions in embodied settings.

Additionally, contrastive verbal and nonverbal expressions are common in many real-world

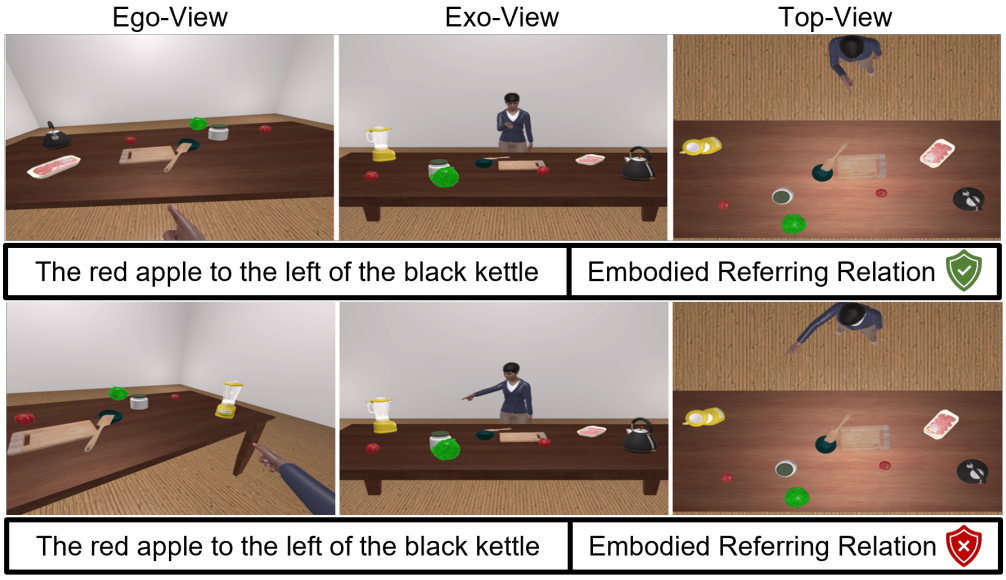


Figure 4.1: Embodied referring expressions generated using, CAESAR, with verbal and nonverbal modalities from multiple views. Top: verbal utterance and nonverbal gestures both referring to the same object (i.e., Apple). Bottom: verbal utterance refers to the Apple; however, the nonverbal gestures refer to the Blender.

settings. For example, humans often mistakenly describe one object while pointing to another object. People are adept at identifying these scenarios and involve themselves in a conversation to complete the communication. Similarly, an intelligent AI-agent should identify inconsistent interactions from multimodal referring expressions. However, existing datasets only contain congruent and complete verbal and nonverbal interaction signals. Therefore, to train a robust model, we need a dataset with contrastive data samples, enabling the agent to request additional information from human partners in cases of incongruent signals.

To address the shortcomings of the existing datasets, we have developed a novel embodied simulator, CAESAR, to generate large-scale datasets of referring expressions. To the best of our knowledge, CAESAR is the first simulator to generate multimodal referring expressions with verbal utterances and nonverbal gestures in a virtual environment. CAESAR has three novel aspects which differentiate it from other synthetic data generation systems (e.g., CLEVR [23] and Kubric [89]). First, CAESAR simulates scenarios in which verbal utterances and nonverbal cues (pointing gesture and gaze) refer to objects in an embodied setting (Fig. 4.1). We have collected real human pointing gesture data using an OptiTrack motion capture system [90] and emulated the same behaviors in CAESAR by incorporating a new stochastic deictic gesture generation approach. Second, CAESAR renders multiple views from different perspectives, such as ego-, exo-, and top-view, that can aid in training models to learn different perspectives for comprehending multimodal referring expressions. Third, taking inspiration from previous work [21], we have designed a module in

CAESAR to generate contrastive samples, where the virtual human is pointing to an object while verbally describing a different object.

One of the primary goals of developing CAESAR is to democratize the data generation process so that researchers without simulator development experience can have complete control of generating a diverse dataset to train and evaluate a learning model. Similar to existing data generation systems, the development of our simulator requires extensive knowledge of motion planning and game engine. Thus, to make it accessible to everyone, we have developed a tool that enables researchers to generate diverse samples without any simulator development experience. Using this tool, we have developed two large-scale datasets, CAESAR-XL and CAESAR-L, for understanding multimodal referring expression in an embodied virtual environment. A comparison of our developed datasets and other existing datasets for referring expression understanding is listed in Table 4.1. We have developed the simulator and generated data under an approved IRB (protocol number: 4627, Title: Understanding Multimodal Human Instruction in Embodied Environment).

Although several state-of-the-art visual-language models have been proposed for different tasks, such as spatial relation recognition [20], [21], [51], referring expression comprehension and visual question answering [26], [27], [46], these approaches use nonverbal embodied interactions from only an exocentric perspective. Thus, we have adopted state-of-the-art models and benchmarked on our datasets for grounding embodied spatial relations using multiple views and multimodal data. Our experimental results suggest that these models' performance varies with perspective, and nonverbal cues can improve it. Moreover, the results also indicate that we need to develop models that extract salient nonverbal cues and effectively fuse verbal utterances for robust performance.

The key contributions of this work are listed below:

- We have developed a novel embodied simulator, CAESAR, to generate referring expressions with verbal uttrances and nonverbal gestures captured from multiple perspectives.
- We have generated two large-scale and one small datasets of multimodal referring expressions in an embodied setting using CAESAR.
- We have benchmarked various models on our dataset. The results suggest these models cannot effectively learn perspective-taking, which opens new research directions to develop robust models for embodied referring expression comprehension.
- CAESAR allows researchers to tune the simulator's parameters without any development experiences to generate customized samples for training and diagnosing their models.

4.1 CAESAR: An Embodied Simulator

In this section, we present CAESAR, an embodied simulator capable of automatically generating multimodal referring expressions with verbal utterances and nonverbal cues (pointing gesture and gaze) to refer to an object. Generated embodied referring expressions are depicted in Fig. 4.1.

Table 4.1: Comparison of the datasets of referring expression comprehension. V, NV, E, C, and A denote verbal, nonverbal, embodied, contrastive samples, and ambiguous samples, respectively. *Average number of words.

Datasets	V	NV	E	Views			C	A	No. of Images	No. of Samples	Object Categories	Avg. Words*
				Exo	Ego	Top						
PointAt [147]	✗	✓	✓	✓	✗	✗	✗	✗	220	220	28	-
ReferAt [64]	✓	✓	✓	✓	✗	✗	✗	✗	242	242	28	-
IPO [148]	✗	✓	✓	✓	✗	✗	✗	✗	278	278	10	-
IMHF [149]	✗	✓	✓	✓	✗	✗	✗	✗	1716	1716	28	-
RefIt [69]	✓	✗	✗	✓	✗	✗	✗	✗	19,894	130,525	238	3.61
RefCOCO [150]	✓	✗	✗	✓	✗	✗	✗	✗	19,994	142,209	80	3.61
RefCOCO+ [150]	✓	✗	✗	✓	✗	✗	✗	✗	19,992	141,564	80	3.53
RefCOCOg [151]	✓	✗	✗	✓	✗	✗	✗	✗	26,711	104,560	80	8.43
Flickr30k [152]	✓	✗	✗	✓	✗	✗	✗	✗	31,783	158,280	44,518	-
GuessWhat? [153]	✓	✗	✗	✓	✗	✗	✗	✗	66,537	155,280	-	-
Cops-Ref [154]	✓	✗	✗	✓	✗	✗	✗	✗	75,299	148,712	508	14.40
CLEVR-Ref+ [22]	✓	✗	✗	✓	✗	✗	✗	✗	99,992	998,743	3	22.40
YouRefIt [24]	✓	✓	✓	✓	✗	✗	✗	✗	497,348	4,195	395	3.73
CAESAR-L	✓	✓	✓	✓	✓	✓	✓	✓	11,617,626	124,412	61	5.56
CAESAR-XL	✓	✓	✓	✓	✓	✓	✓	✓	841,620	1,367,305	80	5.32

For CAESAR we have created an environment where an embodied agent (an avatar) refers to various objects distributed on a table top through nonverbal gestures and verbal utterances by exploiting spatial relation with other objects in the scene. CAESAR generates the environment by dynamically loading various avatars, objects, walls, and tables.

4.1.1 Observer-Aware Object Generator

To ensure plenty of variation across data samples while limiting clutter, CAESAR randomly spawns between four to ten objects from our pre-populated object library. Among these spawned objects, CAESAR randomly chooses one object as the referred object, which will later be described through nonverbal cues and verbal utterances. We apply some constraints to an object to be declared properly generated. First, objects can only occur in a scene at most three times. Second, objects must be partially visible from both the ego and exo views. To promote object diversity, CAESAR varies spawned objects in rotation, color, size, and position. CAESAR does not vary some object colors, such as oranges, to ensure proper object appearance.

Figure 4.2: Comparison of real (top) and synthetic motion generated from CAESAR (bottom). We used real human motion using an OptiTrack motion capture system to synthesize gestures in our simulator. The results suggest that our synthetic generated motion are very similar to the real motion.

4.1.2 Embodied Referring Expression Generator

CAESAR generates both verbal and nonverbal referring expressions for each embodied interaction. To vary nonverbal expressions, different cues are used interchangeably by the human avatar to refer to objects. Nonverbal data consists of procedurally generated pointing gestures and a gaze that refers to an object. To achieve this, we dynamically calculate musculo-skeletal motion for eight different avatars [155], [156]. Additionally, verbal expressions are generated randomly from a set of templates.

Pointing Gesture Synthesis: One of the primary goals of the CAESAR simulator is to generate realistic pointing gestures that match the amount of variability in real human gestures. To accomplish this, we have researched prior works for procedural pointing gesture synthesis and developed a novel synthetic pointing gesture synthesis algorithm. Past algorithms generally fall under three categories: data-driven algorithms that use motion-capture data to fit constraints [157], physics-driven algorithms that build motion using a musculo-skeletal simulation [158], and hybrid algorithms combine aspects of the prior two. Our method of synthesizing motion [155] is a hybrid algorithm that synthesizes a motion path for pointing gestures based on motion-capture data (using similar timing and arc-like motion). We also determine joint rotations based on inverse kinematics to fit this motion path, and subsequently applies a physical simulation to account for gravity, momentum, and self-collision. We have used real human motion data using an OptiTrack motion capture system to synthesize human pointing gestures in CAESAR (Fig. 4.2).

Our gesture generation algorithm has five phases: rest, preparation, stroke, hold, and retraction. The rest phase consists of a static idle animation; the subsequent phases are layered onto this to

Table 4.2: Verbal referring expression generation templates. Here, <Obj>: Referred object name, <Obj-1>: Reference object name, <Obj-n Prop.>: Color or Size of object n, <SR>: Spatial relation. Note that spatial relations/locations are relative to either the observer (exo view) or embodied agent (ego view).

No.	Template of Verbal Referring Expression	Example Instance
1	<Obj>	The apple
2	<Obj-Prop.><Obj>	The red apple
3	<Spatial Location><Obj Prop.><Obj>	The center small apple
4	<Obj><SR><Obj-1 Prop.><Obj-1>	The apple to the left of the big cutting board
5	<Obj Prop.><Obj><SR><Obj-1 Prop.><Obj-1>	The small apple next to the brown cutting board

simulate the micro-movements that occur while muscles are under tension. For the preparation, stroke, and hold phases of the gesture, the motion path is determined by constructing a Catmull-Rom curve [159] through a set of three points: one at the rest-coordinates of the pointing hand, one at the peak-coordinates, where the hand is most extended in a pointing gesture, and a third point at the midpoint of the previous points, with a varying displacement as to randomly alter the shape of the motion. This curve is then converted into a Bezier Curve [160]. To allow human hands to travel along paths we used 3D Bezier Curves [161]. We also added a Cubic Easing function along the path and a basic physical particle simulation to the path-following object as a basis for our Two-Joint Inverse Kinematics (IK) target when creating the arm animation. To implement the IK and physical simulation, we used the Unity Animation Rigging [162] and Dynamic Bone [163] packages. The retraction phase of the gesture is implemented by easing off the IK constraint's strength, allowing for gravity to swing the limb back into its rest position, directly under the shoulder joint.

Gaze Synthesis: Our avatars' head and body orientation is calculated through a set of IK targets, also using the Unity Animation Rigging package [162]. They are layered on top of each other: first, the body is applied, and then the head is applied. The IK target weights are eased in according to a timing parameter shared between the pointing and gaze systems. This constraint ensures that when both gaze and pointing gestures are generated, the avatar will look towards the target and change their body's orientation before pointing.

Verbal Referring Expression Generation: Taking inspiration from previous works [22], [68], [164], we have developed five compositional templates to generate verbal referring expression, presented in Table 4.2. In these templates, the target object is referred to by verbal and nonverbal cues, a reference object is used to add context to the target object's location, and the object's properties, such as color, size, and spatial location (left, right, corner), are varied. For example, using Template-5, we can generate the verbal message "the red apple to the right of the black kettle", depicted in Fig. 4.1. We also varied the spatial relation/location of the target object by referring to it from either the observer's or the actor's perspectives, resulting in twelve verbal expressions formulated from the five templates.

4.1.3 Rendering Nonverbal Referring Expressions from Multiple Views

CAESAR generates nonverbal referring expressions in three scenarios - a person gazing at an object, a person pointing to an object, and a person gazing and pointing at an object. There is another setting where no human avatar is rendered and the scene only contains objects, named *no human* scenario. Nonverbal cues in three of the scenarios are captured from three camera views: ego, exo and top. We have also generated skeletal poses of a simulated human avatar using the Vectrosity package [165].

4.1.4 Contrastive Sample Generator

CAESAR generates contrastive embodied referring expressions where the given embodied referring expressions are insufficient to successfully ground an object. As described in Section 4.1.3, there are four different scenarios our simulator generates; CAESAR generates contrastive embodied referring expressions for all four of these different scenarios. In the situations of only gaze, only pointing gesture, and both gaze and pointing gesture, the human avatar points, gazes or both at an object that it is not verbally describing. This is made apparent in Fig. 4.1, where the humanoid verbally and nonverbally describing two different objects results in contrastive expressions. For the scenarios with *no human* avatar, CAESAR generates a verbal expression that describes an object not in the scene. These contrastive data samples can help to train models to ground embodied spatial relations. While generating these contrastive scenarios, we apply several constraints to ensure non-ambiguity in whether a scenario is contrastive or not. For example, a chosen object's proximity to different copies of that object is checked to ensure that a referred object can be distinguished from other referred objects. Additionally, the contrastive object selected for nonverbal expression is checked for being adequately spaced from the chosen object and of a different category. These constraints ensure the person's gaze or gesture is sufficiently different to make the sample contrastive.

4.1.5 Data Annotation

CAESAR generates detailed annotations of each data sample, including bounding box coordinates for all the generated objects from all three views, object attributes (color, size, absolute location), and their relative locations from the actor (ego view) and the observer (exo view). We found that using Unity's object mesh renderer for bounding box calculations provided large overestimates, so we calculate the position (x and y coordinates) of each vertex of an object relative to each camera, which leads to accurate bounding box annotations. Additionally, as the ego view camera constantly changes position and rotation (the other cameras remain static), we dynamically calculate bounding boxes during videos for the ego camera to effectively track where each object is relative to the moving camera. Moreover, CAESAR annotates each verbal referring expression according to object attributes and spatial relations. It also records environmental parameters, such as lighting conditions (number of lights, position, intensity) and background color.

4.1.6 Configurable Data Generation Interface

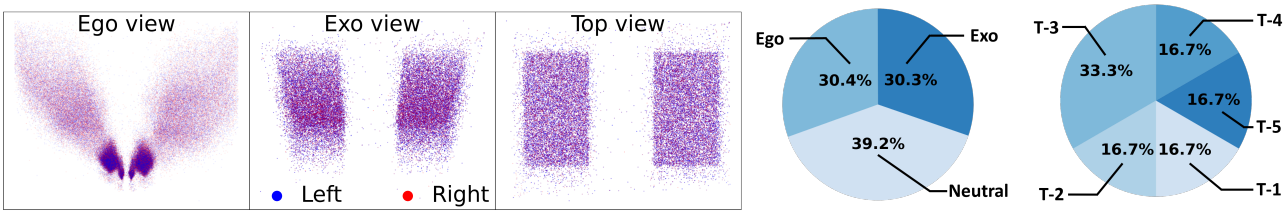
One of the challenges of many simulators that generate datasets is a lack of configurability or a requirement of extensive development experiences using certain libraries. We have developed a tool to configure CAESAR without programming or game engine experience through simply clicking buttons in the Unity inspector window to toggle features. This tool uses serialized fields inside our main manager, directly allowing users to configure different features that will be used internally. These configurable features include the ability to specify whether video should be recorded, activate different cameras (i.e., the skeletal camera), and designate the number of scenes to generate in parallel.

4.2 Dataset Analyses

Using CAESAR, we have generated two datasets. The first dataset, CAESAR-L, consists of $124,412$ data samples created from $11,617,626$ images at a resolution of 480×320 pixels. These data samples are divided into train, validation, and test data splits with $74,760$, $24,779$, and $24,873$ data samples, respectively. The second dataset, CAESAR-XL, consists of $1,367,305$ data samples, which were created from $841,620$ images by varying verbal expressions in the five different settings described in Section 4.1.3. These data samples are divided into train, validation, and test data splits with $1,123,886$, $122,157$, and $121,262$ data samples, respectively. These images were rendered with a resolution of 720×480 pixels using an object pool of size 80. The lower sample to image ratio in CAESAR-L dataset when compared to CAESAR-XL can be explained by the CAESAR-L dataset containing images, videos (rendered at 15 fps), and a skeletal pose. Table 4.1 shows an in-depth comparison between CAESAR-L, CAESAR-XL, and other similar datasets from the literature.

Similar to previous works [21], [51], one of the primary goals of CAESAR is to reduce the spatial location bias in generated data. For example, if the terms *"on the left"* and *"on the right"* always refer to objects located on the left or right side of the scene from the actor's perspective, models will exploit this bias to ground spatial relations. If actors give utterances such as *"on the left"* and *"on the right"* but from the view of the observer, instead of the view of the actor, these models will not be able to complete the perspective taking necessary to successfully ground these utterances. To address the issue, we randomly select verbal expressions from either the ego (actor) or exo (observer) perspective. We visualize the referred object location in each view (ego, exo, and top) for left and right spatial locations (Fig. 4.3(a)). These visualizations suggest that referred objects of these two locations are spread across both sides of the three views, meaning the object locations are not identifiable through solely verbal cues. This analysis indicates that our datasets are not biased in generating spatial locations, and thus, can force models to utilize nonverbal cues to succeed in embodied spatial relation grounding by recognizing which perspective given utterances come from.

As shown in Fig. 4.3(b), our datasets contain verbal utterances from multiple perspectives (ego,



(a) Referred object locations in different views. (b) Distribution of perspectives. (c) Distribution of templates.

Figure 4.3: Analysis of CAESAR-XL dataset. (a) CAESAR-XL has little to no bias as spatial-visual cues of object locations are less separable for a given *left* and *right* spatial location in verbal utterances. Note that the color being purple is a result of overlapping left and right points. (b) CAESAR-XL contains referring expressions from all the perspectives: ego (actor), exo (observer), and neutral (expressions that do not depend on perspective-taking). (c) CAESAR-XL generates verbal utterances using the templates described in Table 4.2. ($T-n$ denotes the $n-th$ template).

exo, neutral). Neutral utterances do not depend perspective to ground relations. For example, the term "the red apple" does not contain any spatial relation/location terms and thus does not require perspective to ground. The relatively equal distribution of multiple perspectives in our datasets promote the ability for models to learn perspective-taking in embodied settings. Fig. 4.3(c) shows the distribution of each verbal expression template presented in Table 4.2, where template three (the template involving spatial location, an object property, and then an object) was about twice as common as all other templates. This was done to ensure spatial relations and spatial locations were used at the same frequency (as templates four and five both use spatial relations). Thus, our dataset is not biased towards verbal expressions.

4.3 Embodied Relation Grounding Models

Existing models use verbal utterance and an exocentric view to recognize spatial relations. However, our dataset contains both verbal and nonverbal modalities captured from three views. Thus, we have adopted visual-language models to develop three representations learning models for the embodied spatial relation grounding task: a CLIP Model [92], a Dual-Encoder (ViT [166] + BERT [167]) model and a Late Fusion (ResNet [98] + BERT [167]) model (Fig. 6.8).

CLIP-based Model: CLIP model excels at aligning visual and language modalities [92]. Thus, we use the CLIP model to detect whether nonverbal cues and verbal utterances of an embodied expression refer to the same object. However, CLIP can take an image-text pair and produce verbal and visual representations. For this reason, we pair the verbal expression, T, to each of the views of the nonverbal expression (Ego (V_{ego}), Exo (V_{exo}), and Top (V_{top})) and pass each modality pair to CLIP models: $E_i^v, E_i^t = CLIP(V_i, T), i \in (ego, exo, top)$. Here, $E_i^v \in \mathbb{R}^{B \times S}$ and $E_i^t \in \mathbb{R}^{B \times S}$ are the visual and verbal embeddings from CLIP models, respectively. (B is the batch

size and S is the embedding dimension)

Visual-Language Transformer Models: We have extended two visual-language transformer models, ViLT [91] and VisualBERT [26], for grounding embodied relations. As these models were designed to produce representations from a single visual scene and a verbal utterance, we extend these models to extract visual-language representations from multiple visual scenes. First, we extract visual representations from multiple visual scenes using ResNet-50. Finally, we pass these visual tokens and tokenized verbal utterance to these visual-language transformer models. These models process these visual and verbal tokens using a single transformer model to extract combined visual-language representation. This representation is then used for grounding embodied relations.

Dual-Encoder Model: Like CLIP models, we pass each pair of visual and verbal modalities to Dual-Encoder models to extract verbal and non-verbal representations. We use ViT and BERT in Dual-Encoder models to encode visual and verbal modalities, respectively. Both of these encoders (ViT and BERT) use a Transformer to extract representations.

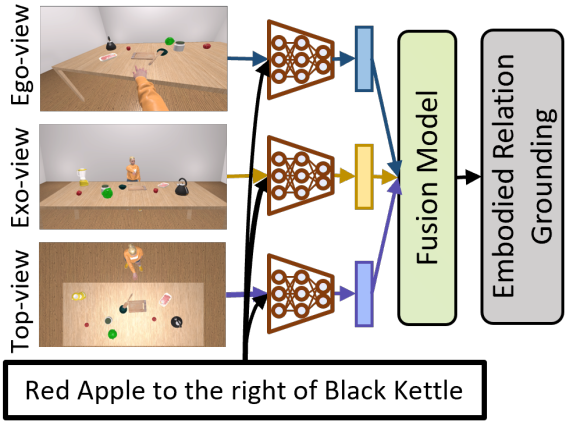


Figure 4.4: Embodied relation grounding model. Data from each pair of verbal and visual modalities is passed through a shared visual language models to extract representations, which are then fused for embodied spatial relation grounding.

Late Fusion Model: In the Late Fusion model, all the visual and verbal modalities are encoded independently using ResNet-50 [98] and BERT [167] models. We projected the extracted verbal and visual modalities representations to a fixed-sized embedding. Although Dual-Encoder and Late-Fusion models have similar architectures, as both of these models use separate encoders for visual and verbal modalities, there are two main differences. First, Dual-Encoder models use Transformers to design both visual and verbal encoders. This contrasts from how Late-Fusion models use different architectures for these two types of modalities, such as ResNet for the visual encoder and BERT for the verbal encoder. Second, in the Late-Fusion model, we first extract visual representations for all three views and fuse the visual and verbal representations using a Transformer-style multi-head attention approach [94]. This differs from the Dual-Encoder model, where we pair the verbal utterances with each view and pass each visual-verbal pair through the model to extract pairwise representations, which are concatenated to produce multimodal representations.

Multimodal Fusion: We fused the extracted verbal and visual representations from the above-mentioned models to produced multimodal representations, which are used to detect embodied

spatial relation. We used four fusion approaches: SUM, CONCAT, Self-Attention, and Cross-Attention. The first two fusion approaches summed and concatenated the verbal and visual representations. The self-Attention approach is similar to the Transformer-style self-attention [94] which attends each of the verbal and visual representations and sums the attended representations. We have also employed a Cross-Attention approach, which is similar to the co-attention approach from ViLBERT [27]. Cross-Attention is essentially a query-key-value style attention approach, where verbal embeddings are used as query and visual embeddings are used as key and values. Finally, the fused embedding is passed through a multilayer perceptron to detect embodied spatial relations.

Model Training Environment Setup: We projected the visual and verbal embeddings from CLIP, Dual-Encoder, Late Fusion models to 512, 768, and 768 sized embeddings, respectively. We fused all the embeddings from multiple views and pass them through a multilayer perceptron to classify whether verbal and nonverbal expressions refer to the same object. We trained the model using Cross-entropy loss on the CAESAR-XL dataset for four epochs. To train our model we used PyTorch-Lightning [100] and HuggingFace [168] to implement models and used the Adam optimizer with weight decay regularization [95]. An initial learning rate set to $3e^{-4}$ to train the models. We trained all the models in a distributed GPU cluster environment, where each node contains 4-8 GPUs.

4.4 Experimental Results and Discussion

We evaluated several models on our dataset CAESAR-XL by varying modalities, perspectives, and fusion methods. To evaluate the impact of modalities (verbal and nonverbal), we used all the views (ego, exo, and top) and employed the CONCAT fusion method. Moreover, we used the CONCAT fusion method to evaluate the impact of various perspectives using verbal, gaze, and pointing gesture data. Additionally, we used the verbal utterances from ego, exo, and neutral perspectives to evaluate whether the baseline models can effectively learn perspective to ground embodied referring relation. Finally, to evaluate the impact of the fusion methods, we used all the views with verbal, gaze, and pointing gesture data. The results are presented in Table 4.3.

Impact of modalities: The results in Table 4.3(a) suggest that verbal models without any nonverbal signals (e.g., BERT [167]) can not perform better than random guessing at the relation grounding task. The reasoning behind this performance is that we generated contrastive nonverbal data samples for the same verbal utterance. Additionally, incorporating nonverbal modalities (gaze and pointing gesture) improve embodied spatial relation grounding accuracy. For example, incorporating only gaze cues improved the model's performance compared to models that use verbal and visual modalities only. This performance improvement indicates the necessity of nonverbal modalities to ground embodied spatial relations. However, the performance degrades when models use both gaze and gesture cues, when compared to models using only gaze or gesture. As the evaluated baseline models encode the visual modalities to extract combined representation for pointing gestures and gaze feature representations, these models may not disentangle pointing gesture and gaze

Table 4.3: Embodied spatial relation grounding accuracy of baseline models. The results suggest that nonverbal cues increase embodied spatial relation grounding accuracy. However, the model's performance depends on how nonverbal interactions are captured and how representations from multiple views and modalities are fused. (V: Verbal, NH: Visual without Human, G: Gaze, P: Pointing Gesture, SA: Self-Attention, CA: Cross-Attention, LF: Late Fusion, DE: Dual-Encoder).

(a) Impact of Modalities

Model	V	V+NH	V+G	V+P	V+G+P
BERT	50.00	-	-	-	-
LF	-	78.44	81.51	81.18	76.00
DE	-	62.55	72.26	63.09	74.87
CLIP	-	64.78	83.01	77.31	75.21
ViLT	-	64.86	82.54	84.43	79.90
VisualBERT	-	68.07	80.07	77.94	75.61

(b) Impact of Multi-Perspectives

Model	Ego	Exo	Top	All
LF	60.72	76.51	88.37	76.00
DE	59.75	68.84	71.04	74.87
CLIP	59.12	78.97	78.86	75.21
ViLT	85.43	62.47	52.12	79.90
VisualBERT	57.75	70.16	66.32	75.61

(c) Impact of Fusion Methods

Model	SUM	CONCAT	SA	CA
LF	69.20	76.00	69.04	51.81
DE	77.02	74.87	72.50	74.89
CLIP	82.85	75.21	80.63	75.51

representation to comprehend referring cues accurately. Thus, we must carefully design the model architecture and training procedure to extract cues from nonverbal modalities and effectively fuse these representations to verbal modality to accurately recognize embodied spatial relation.

Impact of multiple perspectives: The results in Table 4.3(b) suggest that the models' performance varies with the perspectives. For example, the top view improves the performance of Late Fusion and Dual Encoder models compared to models using the exo view. These findings underscore that the exo view is not always the optimal perspective for ensuring robust performance. Additionally, although the performance of the single view-based models fluctuates, incorporating multiple perspectives helps achieve consistent performance across the models. In our experiments, we found that multiview models cannot outperform single view-based models, unlike findings from previous works [7], [8], [10], [11], [15], [28], [42], [81]. The reasoning behind this performance degradation is that nonverbal cues can be interpreted differently from multiple views. For example, a person pointing and verbally describing an object as the "left apple" can be visually interpreted by an observer as the "right apple". Thus, extracting synchronized cues across modalities is essential to validating an embodied spatial relation. Moreover, the evaluated baseline models do not explicitly learn to ground perspective to comprehend referring expressions. As the referring expressions in our datasets are generated from multiple perspectives, our datasets can be used to diagnose whether a model can effectively learn perspective-taking to comprehend embodied

referring expressions.

Impact of multimodal fusion: The results in Table 4.3(c) suggest that simple SUM and CONCAT fusion approaches performed better than more complex attention-based fusion methods. For example, CONCAT fusion model outperforms attention-based fusion models in almost all the evaluated settings. This performance degradation is likely because nonverbal cues are interpreted independently from different perspectives. Moreover, as attention-based fusion approaches try to align multiview representations, the conflicting nonverbal cues from different perspectives lead to sub-optimal representations in these baseline models. We can develop models that can jointly consider multiple perspectives in extracting complementary representations from multiview data to address this issue.

Results from human-subject study: We sampled 300 data samples of the exo view from the testing split of CAESAR-XL to conduct a human-subject study on Amazon Mechanical Turk under IRB Protocol: 4627. In this study, we showed the data samples to participants and asked them to indicate whether a virtual avatar was pointing, gazing, and verbally describing the same object. Each sample was shown to three participants, and we took the majority voting to determine the label of a sample. In this study, 397 participants took part where each participant's task approval ratting was at least 95%, and they were compensated for their time. The results suggest that the participants correctly validated the relations in 80.66% of the times.

4.5 Broader Impact

We have developed an easy-to-use simulator for researchers to generate datasets for different purposes. We believe that datasets generated using our simulator can also be used to train and evaluate models for various tasks in embodied settings, such as embodied question answering, object grounding, and conversational human-AI interactions. Moreover, researchers can use Unity plugins to generate other modalities (e.g., a depth map, point clouds, and object segmentation) and annotations (e.g., 3D object spatial locations/rotations) for developing novel multimodal learning models. We expect researchers to be able to generate datasets according to their needs - which CAESAR's configurable parameters allow for. Additionally, CAESAR-generated datasets can be used to pre-train models for embodied instruction comprehension, which can be transferred to robots for comprehending instructions in real-world human-robot interactions. Finally, the findings from our experimental results open some exciting research directions to develop robust models for embodied referring expression comprehension.

4.6 Limitations

Although we developed a 3D embodied environment in our simulator, we have rendered 2D image data in this work. In our future work, we will extend our simulator to render 3D data, such as point clouds. Using this 3D data we can develop models for multimodal instruction understanding in 3D

embodied environments. Moreover, our experimental results from baseline methods suggest room for improvements in embodied referring expression understanding by using multimodal and multi-view data. In the future, we plan to develop a model to extract complementary representations from multiple views and extract salient representations for nonverbal interactions to recognize embodied spatial relations accurately. Although we have developed and evaluated several visual-language transformer models, in future works, it will be an interesting avenue to investigate whether other visual-language models can effectively comprehend the embodied referring expressions. Specifically, as visual-language models take visual and verbal data as input together, it will be worth investigating whether these models can disentangle the nonverbal cues from the visual scene data and fuse the verbal data to produce salient multimodal representations.

Chapter 5
PERSPECTIVE-AWARE MULTITASK MODEL FOR REFERRING EXPRESSION GROUNDING

Humans naturally use multimodal cues, such as verbal utterances and non-verbal signals (gazes and pointing gestures), to refer to objects and events, known as referring expressions [55]–[57], [141]–[144], [146]. In prior work, understanding referring expressions has been generally modeled as grounding relations and objects in visual scenes using verbal utterances, which is known as referring expression comprehension (REF) [45], [52], [53], [150]. These models are often trained in non-embodied settings, where the visual scenes contain objects but disregard human nonverbal signals. Consequently, these models cannot generalize well in comprehending real-world human interactions.

Several recent works have attempted to address the task of comprehending referring expressions by incorporating nonverbal gestures with verbal utterances in embodied settings (known as embodied referring expression comprehension (E-REF)) [24], [64]. However, some crucial issues remain unaddressed in these recent works. Particularly, most embodied referring expression datasets only capture human interactions from an observer perspective with exo-centric views. People innately use an understanding of perspective, which can be observed in how humans interchangeably use perspectives from the speaker and the observer when referring to objects during interactions. For example, a person can refer to an object as "the red lamp to the left of the black hat" from the speaker's perspective or "the red lamp to the right of the black hat" from the observer's perspective (Fig. 5.1). Thus, understanding perspectives can help a model to ground relations and objects. However, the existing datasets do not contain data from other perspectives (e.g., speaker, observer, neutral) and visual views (e.g., exo, ego, top) to train such a model.

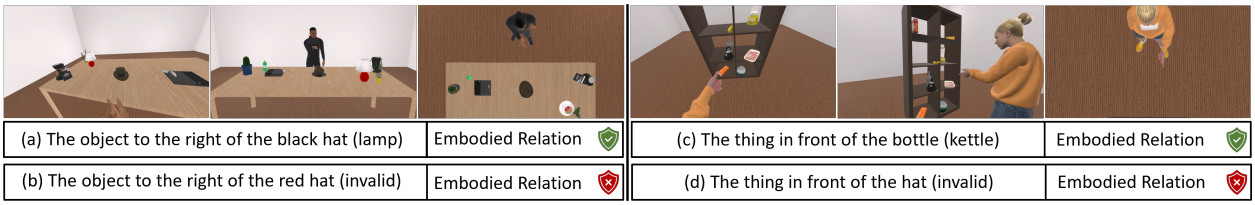


Figure 5.1: Comprehending embodied referring expressions requires an understanding of the perspective, i.e., whether an object is verbally described from the speaker's or observer's perspective. In these scenarios, nonverbal signals (gaze and pointing gesture) can complement verbal utterance to ground an object (a & c). However, sometimes people verbally describe an object and point to or gaze at another object (b & d). Thus, it is also crucial to ground relation for comprehending referring expressions.

Recent works studied REF and E-REF by designing two separate tasks: a relation grounding task [21], [65]–[67] and an object grounding task [20], [24], [68], [69]. In a non-embodied setting, the relation grounding task is defined as determining whether a verbal utterance appropriately describes the spatial relationships between objects in a visual scene. In an embodied setting, this relation grounding task is defined as determining whether a verbal utterance and nonverbal signals (gazes and pointing gestures) refer to the same object. The object grounding task aims to identify a referred object using a verbal utterance and nonverbal gestures. These tasks have many use-cases in real-world interactions. For example, if a person verbally describes an object but nonverbally points to another object, an AI-driven agent can identify these incoherent multimodal cues, using the relation grounding task, and request clarification. In another case, if a person points to an object and asks, "what is the object to the right of the black hat?", then the AI agent can use the object grounding task to identify the referred object (presented in Fig. 5.1). Thus, training models on these two related tasks (relations and objects grounding) and the previously mentioned perspective grounding task can enable achieve seamless human-AI interactions (HAI).

To address these challenges, we have developed a novel perspective-aware multitask model, PATRON, for the relation and object grounding task using multimodal cues. In PATRON, we have designed two cooperative tasks, one for the perspective grounding (the auxiliary task) and another for the relation and object grounding (the target task). In the auxiliary task module, PATRON learns disentangled representations, the auxiliary task-specific and task-guidance representations, to learn perspective grounding. In the target task module, PATRON uses our proposed guided fusion approach that utilizes task-guidance representations from the auxiliary task as prior information to extract guided multimodal representations. PATRON uses a self-attention-based fusion approach to extract supplementary target task-specific representations. Finally, PATRON fuses task-guided and target task-specific disentangled representations to learn relation and object grounding.

Additionally, to overcome the shortcomings of the existing datasets, we have developed a dataset, called CAESAR-PRO, to train and evaluate models for comprehending embodied referring expressions. In CAESAR-PRO, each embodied referring expression is captured from three visual views (ego, exo, and top), and the verbal utterances are generated from three perspectives: speaker, observer, and neutral. We have evaluated the performance of PATRON and state-of-the-art visual-language models by applying on the CAESAR-PRO dataset for perspective and relation-object grounding tasks. Our extensive experimental analysis suggests that perspective learning can improve the performance of visual-language models, including PATRON, for the relation-object grounding task. Moreover, our proposed perspective-aware guided fusion approach helps PATRON to outperform all the evaluated models by achieving the highest accuracy of 74.13% and 81.15% in relation-object and perspective grounding tasks, respectively. Moreover, our ablation study indicates that disentangling multitask representations can help extract salient multimodal features and significantly improve the performance of the relation-object grounding task. Our proposed perspective-aware E-REF model, the dataset, and the insights from our studies open new research directions in HAI.

5.1 PATRON: Perspective-aware Multitask Model

In PATRON, we have designed two tasks: an auxiliary task (perspective grounding) and a target task (relations and objects grounding). We combine relation and object grounding task in a single task (relation-object grounding), where the models identify the referred object if the verbal and nonverbal cues refer to the same object; otherwise, it will report a failed condition. In PATRON, the auxiliary task learns disentangled representations, auxiliary task-specific and task-guidance, where task-guidance representations are used to guide the target task to extract complementary representations. PATRON also learns disentangled representations for target tasks, task-guided and target task-specific, where task-guided representations are learned using task-guidance representations from the auxiliary task. In the following subsections, we present different modules of PATRON.

5.1.1 Unimodal Feature Encoders

PATRON uses modality-specific encoders to encode data from visual and verbal modalities. Visual modalities capture nonverbal gestures in three image views (X_{ego}, X_{exo}, and X_{top}). Verbal utterances (X_{verbal}) refer to an object from a perspective (ego, exo, and neutral). As different modalities have different feature characteristics, PATRON uses separate encoders to encode visual and verbal modalities. This architecture design enables PATRON to utilize state-of-the-art models (F_m) to extract salient unimodal representations (E_m). In our implementation of PATRON, we use ResNet and DistilBERT to extract unimodal representations:

$$E_m = F_m(X_m) \quad , \quad m \in (ego, exo, top, verbal) \tag{5.1}$$

Here, $E_m \in \mathbb{R}^{(B \times D^m)}$, B is the batch size, and D^m is the representation dimension of modality m.

5.1.2 Auxiliary Task Module

In PATRON, the auxiliary task module extracts task-specific and task-guidance disentangled representations from unimodal representations $E_u = (E_{ego}, E_{exo}, E_{top}, E_{verbal})$ (u indicates for unimodal). These disentangled representations are used together to learn perspective grounding, whereas task-guidance representations are also used to guide the target task module to extract perspective-aware complementary representations for relations and objects grounding.

5.1.2 Auxiliary Task-Specific Representation Learning:

In PATRON, we have designed a guided fusion approach to fuse unimodal representations. In the auxiliary task module, PATRON uses verbal representation (E_{verbal}) as queries to fuse visual

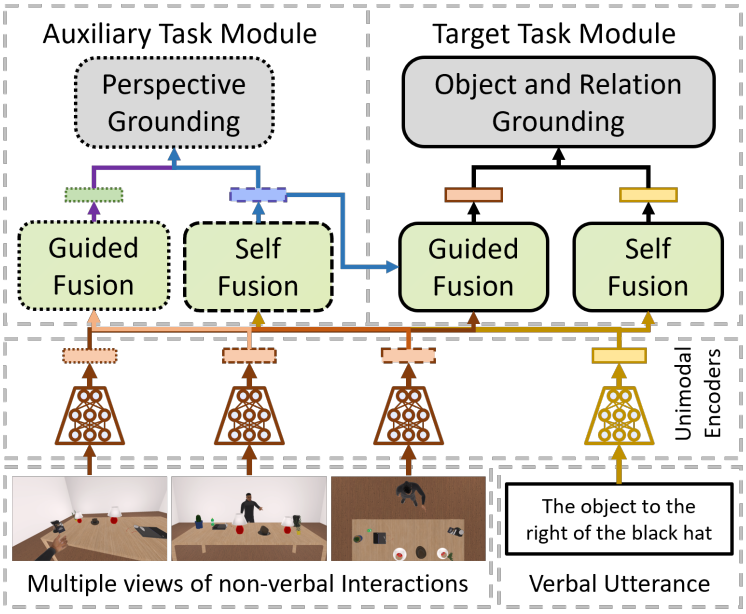


Figure 5.2: PATRON: Perspective-aware Multitask Learning Model. PATRON learns disentangled representations (i.e., auxiliary task-specific and task-guidance representations) for the auxiliary task (perspective grounding) and disentangled representations (i.e., task-guided and target task-specific) for the target task (relation and object grounding). Here, the proposed guided fusion approach extracts the task-guided representations using the task-guidance representations as prior information from the auxiliary task.

modalities ($E_{visual} = (E_{ego}, E_{exo}, E_{top})$) and produce task-specific representations. At first, PATRON projects (E_{verbal}) to produce queries (Q) and projects E_{visual} to produce key (K) and value (V) representations:

$$Q = E_{verbal}W^Q; K = E_{visual}W^K; V = E_{visual}W^V \tag{5.2}$$

Here, W^Q, W^K, and W^V are learnable parameters. Finally, queries are used to extract multimodal representations from keys and values in the following way:

$$E^{'} = \sigma\left(\frac{QK^T}{\sqrt{D^u}}\right)V \tag{5.3}$$

$$E^{aux}_{task\_specific} = W^o E^{'} \tag{5.4}$$

D^u is the unimodal representation dimension and W^o is a learnable parameter. As we also use guided fusion approach in the target task module, we can summarize this as,

$$E^{aux}_{task\_specific} = Guided\_Fusion(Query, E_u) \tag{5.5}$$

5.1.2 Task-Guidance Representation Learning:

PATRON uses self-attention approaches to fuse unimodal representations and extract task-guidance representations ($E^{aux}_{task\_guidance}$), which is disentangled from $E^{aux}_{task\_specific}$:

$$E^{aux}_{task\_guidance} = Self\_Attn(E_u) = \sum_{m \in M} \alpha_m E_m \tag{5.6}$$

$$\alpha_m = \frac{exp(\beta_m)}{\sum\limits_{m \in M} exp(\beta_m)} \quad , \quad m \in M \tag{5.7}$$

$$\beta_m = (W^{aux})^T E_m \quad , \quad m \in M \tag{5.8}$$

Here, M is the modality list (ego, exo, top, verbal), W^{aux} is a learnable parameter, and α_m is the attention score which is calculated using a 1D-CNN with a filter size of 1.

5.1.2 Perspective Grounding Task:

PATRON fuses disentangled representations ($[E^{aux}_{task\_specific}; E^{aux}_{task\_guidance}]$) using a self-attention approach ($Self\_Attn$: Eq. 5.6) to learn perspective. PATRON uses $E^{aux}_{task\_guidance}$ to learn perspective for ensuring that it contains perspective-aware information, which PATRON uses in the target task module:

$$E^{aux}_{fused} = Self\_Attn([E^{aux}_{task\_specific}; E^{aux}_{task\_guidance}]) \tag{5.9}$$

$$y_P = F_{Perspective}(E^{aux}_{fused}) \tag{5.10}$$

Here, $F_{perspective}$ is a multi-layer perceptron to learn perspective grounding.

5.1.3 Target Task Module

In PATRON, the auxiliary task module (perspective grounding) guides the target task module to extract salient multimodal representations for grounding relations and objects. PATRON uses Guided and Self Fusion modules to extract representations for target task learning.

5.1.3 Task-Guided Representation Learning:

PATRON uses our Guided Fusion approach (Section: Task-Specific Representation Learning and Eq.5.5), to fuse unimodal representations (E_u). In the target task module, the guided fusion approach aims to extract perspective-aware multimodal representations that can be used for grounding relations and objects. PATRON utilizes the guidance representations ($E^{aux}_{task\_guidance}$) from the auxiliary task module as prior information to extract salient multimodal representations:

$$E^{target}_{guided} = Guided\_Fusion(E^{aux}_{task\_guidance}, E_u) \tag{5.11}$$

5.1.3 Target Task-Specific Representation Learning:

Although a guided fusion approach helps PATRON to extract perspective-aware representations (E^{target}_{guided}), verbal and visual modalities can provide additional information to E^{target}_{guided}. PATRON uses $Self\_Attn$ (described in Section Task-Guidance Representation Learning and Eq. 5.6) to extract supplementary representations for relation-object grounding: $E^{target}_{task\_specific} = Self\_Attn(E_u)$.

5.1.3 Relation-Object Grounding Task:

PATRON grounds relations and objects together. PATRON identifies the target object that is referred to by multimodal cues - verbal utterances and nonverbal gestures (gazes and pointing gestures). If the verbal utterance and nonverbal gestures refer to two different objects, then the model should identify these inconsistencies (invalid embodied referring relations) and should not ground any objects. To accomplish this, PATRON fuses guided representations (E^{target}_{guided}) and target task-specific representations ($E^{target}_{task\_specific}$) through a self-attention approach ($Self\_Attn$: Eq. 5.6):

$$E^{target}_{fused} = Self\_Attn([E^{target}_{task\_guided}; E^{target}_{task\_specific}]) \tag{5.12}$$

$$y_{OR} = F_{OR}(E^{target}_{fused}) \tag{5.13}$$

Here, F_{OR} is a multi-layer perceptron to learn grounding relations and objects.

5.1.4 Multitask Learning

We use a multitask learning loss to train PATRON for jointly learning auxiliary (perspective grounding) and target tasks (relations and objects grounding). We use cross-entropy to calculate the loss for auxiliary (L_P) and target (L_{RO}) tasks:

$$L_P(y_P, \hat{y}_P) = \frac{1}{B}\sum_{i=1}^{B} y_{(P,i)} \log \hat{y}_{(P,i)} \tag{5.14}$$

$$L_{RO}(y_{RO}, \hat{y}_{RO}) = \frac{1}{B}\sum_{i=1}^{B} y_{(RO,i)} \log \hat{y}_{(RO,i)} \tag{5.15}$$

$$L_{multitask} = \gamma_p L_P + \gamma_{RO} L_{RO} \tag{5.16}$$

Here, γ_p and γ_{OR} are task loss weights. L_P helps to learn perspective-aware representations for grounding the perspective. This loss is also used to learn disentangled representation ($E^{aux}_{task\_guidance}$) for guiding the target task to learn perspective-aware multimodal representations.

5.2 CAESAR Dataset

We have used an embodied simulator, CAESAR [25] to develop a dataset of embodied referral expression. CAESAR allows to automatically generate datasets and synthesizing human gaze and gestures from multiple perspectives (ego, exo, and top). Moreover, CAESAR can generate contrastive situations where the person verbally and nonverbally referring two different objects.

5.2.1 New Environment Creation in CAESAR

We have developed an additional embodied environment in CAESAR, called a shelf environment, where various objects are located on a shelf (Fig. 5.1-right), whereas the original CAESAR simulator contains only a table-top environment (Fig. 5.1-left). These two environments allow us to generate diverse data samples with more spatial relations, such as above and below, enabling the model to understand spatial relations in three dimensions. In contrast, models trained only on the table-top environment can only understand spatial relations on the 2D plane of the table. Moreover, due to the locations of cameras in the shelf environment, the observer's point of view differs from the table-top environment (Fig. 5.1). The camera angle and perspective variation are significant as this new environment offers diversity from the contrasting perspectives of the table-top environment, where the observer is always placed in front of the speaker. Additionally, we have generated the depth map visual modality and segmentation mask of objects which can be used in other E-RFE tasks, such as scene segmentation.

5.2.2 Dataset Generation

To accomplish a realistic and sufficiently variable synthesis of human gaze and pointing gestures, we have used the CAESAR simulator, which uses a gesture synthesis algorithm on real-world data collected using a motion capture system. CAESAR uses inverse kinematics applied to both the

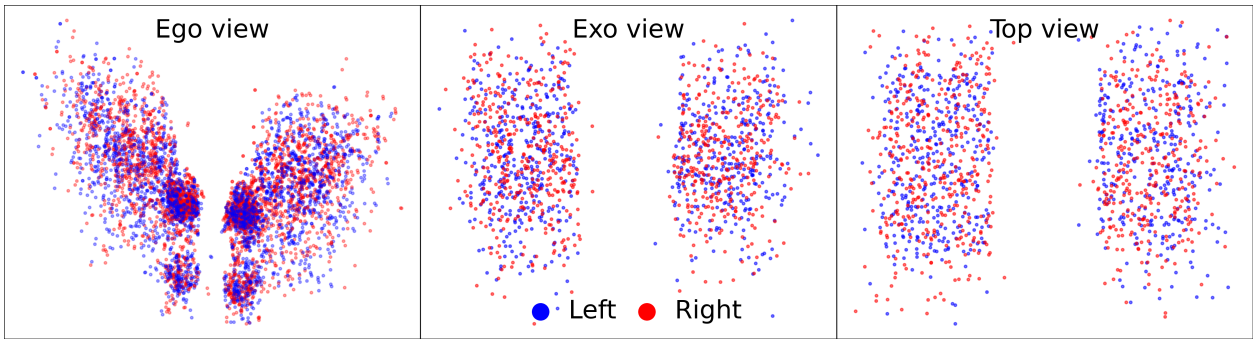


Figure 5.3: A visualization of referred object locations from different views in the table-top environment is presented here. These locations indicate that the CAESAR-PRO dataset has little to no bias toward object locations in visual scenes and is evenly distributed for a given *left* and *right* spatial relations in verbal utterances.

chest and head of the human to generate gestures. To construct verbal referring expressions, we used several templates from CAESAR. We have also included additional spatial relations, such as the above and below relation, which allows for generating more diverse data samples and training models to learn 3D spatial relations. The general structure of these templates: <referred object location><referred object properties><spatial relation>< reference object location><reference object properties>. We have varied this template structure and created eight unique sub-templates. Additionally, we have varied the object names, colors, sizes, locations, and spatial relations to generate diverse verbal expressions to identify one of up to ten objects in a scene from multiple perspectives.

5.2.3 Dataset Analyses

The CAESAR-PRO dataset consists of $128,100$ samples, with a train $(79,431)$, validation $(24,597)$, and test split $(24,072)$. The CAESAR-PRO dataset is composed of $229,036$ images, which are mixed with different verbal expressions from multiple perspectives. Images are rendered at a resolution of (480×320) pixels, and an object library of 61 objects is used. We randomly sampled 10 objects which are used as referred objects. Each data sample consists of RGB images, skeletal images, depth map images, and object segmentation mask images for three camera views and a verbal utterance. We also generated task labels for perspective, relation, and object grounding tasks.

We aim to generate a dataset that is not biased to object locations and spatial relations. For example, the term "on the left" always refers to objects on the left side of the scene from the observer's perspective would cause trained models to bias towards only using verbal cues, resulting in a model not being aware of the perspective-taking necessary to ground real-world referring

expressions with verbal and nonverbal signals. In Fig. 5.3, it is evident that the object locations in CAESAR-PRO (table-top environment) are not tied to the spatial relations in the verbal utterances. These diverse data ensure that models use nonverbal cues to ground objects rather than solely relying on verbal cues.

5.3 Experimental Setup

We have evaluated the performance of PATRON by comparing its performance against the following models for the perspective and relation-object grounding tasks: MuMu [11], VisualBERT [26], CLIP [92], Dual-Encoder (similar to Lit model [169]), and Late-Fusion (similar to HAMLET model [7]). These models take a visual image and a verbal utterance to produce visual-language representations for learning downstream tasks. We extend these models to process multiple visual views with a verbal utterance in our evaluation settings, as there are multiple visual views (ego, exo, and top) capturing nonverbal gestures. For VisualBERT, we extract visual representations of multiple views using ResNet-101 and pass these representations with a verbal utterance to learn visual-language representations. For the CLIP and Dual-Encoder models, we pair the verbal utterance to each visual view and pass each visual-verbal pair through the model to extract visual-language representations, which are later concatenated. For the Late-Fusion and MuMu models, we extract visual and verbal representations using ResNet-101 and DistillBERT [170], respectively. Late Fusion model fuses these representations using the Multi-Head Self-Attention approach [94], whereas MuMu uses a guided fusion approach.

We have evaluated these models' performances by applying on the CAESAR-PRO dataset. As some classes contain more data samples than others, we used macro-accuracy metrics to evaluate perspective, object, and relation grounding tasks. We trained each model for eight epochs with a learning rate set to $1e^{-5}$ in a distributed cluster environment with eight A100 GPUs in each cluster node. We train all the models using Pytorch-lightning [120] environment with a fixed seed to ensure reproducibility.

5.4 Experimental Results and Discussion

5.4.1 Comparison of Multitask Learning Approaches

We evaluated the performance of PATRON and other models by applying on the CAESAR-PRO dataset in single and multitask learning settings. In these experiments, a model takes multiple visual views (ego, exo, and top) and a verbal utterance from multiple perspectives (speaker, observer, and neutral) to learn two tasks: (i) perspective grounding task and (ii) relation-object grounding task. In the multitask model, we chose either perspective or relation-object grounding task as the auxiliary task (the first task in the model architecture) and another task as the target task (the second task in the model architecture). For example, in Task Order I, we chose perspective grounding as the auxiliary task and relation-object grounding as the target task. In Task Order II, we chose

(a) Single task models trained separately		
Models	Perspective (Pers.)	Relation-Object (RO)
Late Fusion	74.90	65.50
Dual Encoder	**77.87**	54.47
CLIP	71.23	65.26
VisualBERT	77.20	**66.53**

(b) Multitask models with different task order in model				
Models	Task Order I		Task Order II	
	Pers →	RO	RO →	Pers
Late Fusion	72.30	61.80	65.40	75.12
Dual Encoder	75.67	64.99	43.66	75.77
CLIP	74.52	68.14	56.82	73.02
VisualBERT	74.52	65.90	62.15	69.44
MuMu	73.65	67.48	63.22	75.27
PATRON	**79.85**	**74.13**	**67.63**	**81.15**

Table 5.1: Top-1 macro accuracy of various models of perspective and relation-object grounding tasks.

relation-object grounding as the auxiliary task and perspective grounding as the target task. We have also evaluated state-of-the-art visual-language models in single-task learning settings, where we trained perspective and relation-object grounding tasks using two separate models. We did not evaluate MuMu and PATRON in single-task learning settings, as these models are designed for multitask learning. We present the results of single task and multitask models in Table 5.1 (a) & (b), respectively.

Results: The results in Table 5.1 suggest that PATRON outperforms all the single and multitask models for grounding perspective and relation-object tasks by achieving 81.15% and 74.13% in macro-accuracy, respectively. Among the other visual-language multitask models, CLIP and Dual Encoder achieve the next highest accuracy for relation-object and perspective grounding tasks by achieving 68.14% and 75.77%, respectively. However, among the single task models, VisualBERT and Dual Encoder achieve the next highest accuracy for relation-object and perspective grounding tasks by achieving 66.53% and 77.87%, respectively.

Discussion: The results in Table 5.1 indicate that for both Task Orders (I & II), the performance of PATRON improves compared to the single and multitask models. Although MuMu uses a guided fusion approach and outperforms single task models for relation-object grounding, it fails to outperform PATRON. However, when considering the task order, some multitask models show improved results compared to their single task models. For example, when Task Order I was considered, the CLIP model showed better accuracy than its single task counterpart for both

Models	Non-Embodied		Embodied		
	V	V+NH	V+G	V+P	V+G+P
BERT	26.44	-	-	-	-
Late Fusion	-	**56.33**	54.91	55.30	61.80
Dual Encoder	-	51.53	53.51	56.93	64.99
CLIP	-	52.63	57.38	60.67	68.14
VisualBERT	-	54.45	58.87	57.05	65.90
PATRON	-	54.24	**65.24**	**66.65**	**74.13**

Table 5.2: Impact of nonverbal signals (gaze and pointing gesture) on the performance (Top-1 macro accuracy) of the multitask models in the relation and object grounding task. The results suggest that nonverbal signals improve the performance of the models. (V: Verbal, NH: Visual without Human, G: Gaze, P: Pointing Gesture).

grounding tasks. Similarly, for Task Order II, the CLIP model showed improved performance for the perspective grounding task; however, the performance degrades for the relation-object grounding task compared to the single task model. One can also observe performance degradation of several models in some multitask settings compared to single task settings. For example, the accuracy of the perspective grounding task degrades for the Dual Encoder and VisualBERT models, whereas the accuracy of the relation-object grounding task degrades for the Late Fusion and the VisualBERT models.

The reasoning behind the performance degradation of the multitask models compared to their single-task counterparts is that the baseline models try to learn a shared representation for all tasks in the multitask setting. As multiple tasks compete to maximize their task-specific representations, a shared representation can discard salient representations of individual tasks. On the other hand, PATRON extracts task-specific and task-guidance disentangled representations. In this process, PATRON uses the task-guidance representations to guide other tasks using our proposed guided fusion approach to extract salient multimodal representations. In the same way, PATRON also learns to extract disentangle representations for the target task and trains these tasks cooperatively, whereas most of the other models train these tasks independently. These findings indicate that a multitask model can improve the tasks' performance if the model can disentangle visual-language representations while training the model in a cooperative learning setting, where one task can guide the learning of other tasks.

5.4.2 Impact of Nonverbal Gestures

We aim to investigate how nonverbal cues impact the performance of the models in the relation-object grounding task. We have conducted this analysis in different settings by varying nonverbal gestures: two non-embodied settings (only verbal (no visual), verbal + visual (scenes without

Models	Training Perspectives			
	Speaker	Observer	Neutral	All
Late Fusion	60.42	53.71	60.53	**61.80**
Dual Encoder	59.43	45.23	57.95	**64.99**
CLIP	62.36	58.04	60.99	**68.14**
VisualBERT	55.71	43.46	49.68	**65.90**
PATRON	60.36	47.23	57.85	**74.13**

Table 5.3: Top-1 macro accuracy of the multitask learning models when trained on data samples from single and multiple verbal perspectives and tested on data samples from multiple visual and verbal perspectives.

human)), and three embodied settings (verbal + gaze, verbal + pointing gesture, and verbal + gaze + pointing gesture). We trained the models in a multitask learning setting (auxiliary task: prospective grounding, target task: relation-object grounding). We used visual scenes captured from multiple views (ego, exo, and top) and multiple verbal perspectives (speaker, observer, and neutral) to train the models. Table 5.2 shows the top-1 macro accuracy of the target task.

Results and Discussion: The results in Table 5.2 suggest that PATRON outperforms all the baseline models in all the evaluated settings for the target task (achieving the highest accuracy of 74.13%). The results also indicate that PATRON achieves the highest accuracy when both gaze and pointing gestures were used, compared to when only gaze or only pointing gestures were used in the embodied setting, and only verbal + visual (scenes without humans) were used in the non-embodied setting. Similarly, other baseline models' performances were also improved when nonverbal cues were used compared to the same model trained with a partial set of nonverbal cues or without any nonverbal cues. Additionally, when only verbal utterances were used, without visual scene (i.e., BERT model), the model achieved only 26.44% accuracy. As the dataset contains verbal expressions that can be interpreted differently from different perspectives, nonverbal gestures can help the models disambiguate and accurately perform the relation-object grounding task. These findings suggest that using nonverbal gestures can improve a model's performance in comprehending E-REF.

5.4.3 Importance of Multi-Perspectives

Here, we investigate how varying verbal perspectives (speaker, observer, and neutral) can impact the performance of the models. We trained PATRON and baseline models on the CAESAR-PRO dataset by varying the verbal perspectives while utilizing all the visual views (ego, exo, and top). During testing, we used all the verbal perspectives and visual views. These models are trained in a multiple-task learning setting (auxiliary task: prospective grounding, target task: relation-object grounding). We have reported the top-1 macro accuracy of the target task in Table 5.3.

Models	Auxiliary Task	Target Task	Guided Fusion	Accuracy	Std. Dev.	Significant Over §
M1	✗	✗	✗	61.38	0.97	None
M2: MuMu	✗	✗	✓	64.21	2.27	M1
M3	✗	✓	✓	64.25	1.07	M1
M4	✓	✗	✓	70.38	0.72	M1-3
PATRON	✓	✓	✓	**74.09**	**0.56**	**M1-4**

Table 5.4: The results (Top-1 macro accuracy) of the ablation study, where various components of the model are evaluated on the relation-object grounding task. The results of five runs with different initial parameters are presented. ✓and ✗ denote whether a task learns disentangled representations or not, respectively. § Significance analysis at level $\alpha = 0.05$ (Following Dror et al. [121]).

Results and Discussion: The results in Table 5.3 suggest that all the models demonstrated the highest performance in comprehending E-REF when the models were trained utilizing the data with all the perspectives. For example, training PATRON on multiple perspectives improves the performance of relation-object grounding tasks (achieved 74.13% accuracy) compared to training the same model only on a single perspective. Baseline models also gain similar performance improvement when training the models with data from multiple perspectives. These findings indicate that training models on data samples from multi-perspective can help the models to comprehend E-REF more accurately.

5.4.4 Ablation Study and Significance Analysis

We have conducted ablation studies to evaluate whether our proposed disentangle representation-based guided fusion approach can significantly improve the performance of the relation-object grounding task. We evaluated PATRON and the baseline models on our CAESAR-PRO dataset in the multitask setting (auxiliary task: prospective grounding, target task: relation-object grounding). These models disentangle representations for auxiliary task (task-specific and task-guidance) and target task (task-guided and task-specific). We have conducted a significance analysis ($\alpha = 0.05$) by evaluating these models five times with different parameters initialization (Following Dror et al. [121]). The results are presented in Table 5.4.

Results and Discussion: The results in Table 5.4 suggest that the models with guided fusion can improve the performance of relation-object grounding tasks compared to the model that does not use guided fusion. For example, MuMu (M2) can improve the performance of relation-object grounding by 2.83% compared to a model which does not use guide fusion (M1). Additionally, the models can significantly improve performance if they can disentangle the representation for auxiliary and target tasks (e.g., M3, M4, and PATRON) compared to the models that cannot (e.g., M1). For example, PATRON improves the performance of relation-object grounding tasks by 12.71% by disentangling multiple task representations and using these representations in the guided fusion

approach compared to M1. The reasoning behind this significant performance improvement is that learning disentangled representations allows these models to learn task-specific and task-guidance salient representations, which can be used to guide other tasks. On the other hand, models learning non-disentangle representations need to use the same representations for task learning and guiding other tasks. Consequently, the shared representations neglect task-specific salient representation for learning generalized representations for all tasks and degrade the task performance.

5.5 Broader Impact

The broader impact of this work extends beyond the immediate field of multimodal deep learning and into various applications that could benefit from an improved understanding of embodied referring expressions.

Firstly, the development of PATRON, a perspective-aware multitask learning model, could significantly enhance the performance of AI systems in tasks that involve understanding and interpreting human communication. This includes applications in robotics, where robots need to understand human instructions to perform tasks accurately, and in assistive technologies, where understanding the user's perspective is crucial for providing appropriate assistance.

Secondly, creating the synthetic dataset, CAESAR-PRO, provides a valuable resource for researchers in the field. This dataset could facilitate further advancements in developing models that can understand and interpret embodied referring expressions, thereby contributing to the overall progress in the field.

Finally, the insights gained from this study regarding the importance of perspective grounding can have implications for the design of future AI systems. By highlighting the need for models to learn perspective grounding, this work could guide the development of more effective and intuitive AI systems that can interact with humans in a more natural and understanding manner.

However, it's important to note that while this work has the potential for a significant positive impact, it also raises ethical considerations. As AI systems become more capable of understanding and interpreting human communication, issues related to privacy, consent, and the potential misuse of these technologies become increasingly relevant. Therefore, ethical considerations must guide future work in this area and include measures to mitigate potential risks.

5.6 Limitations

We have proposed a perspective-aware multitask learning model designed for relation and object grounding tasks in embodied settings. PATRON leverages verbal utterances and nonverbal cues and introduces a novel guided fusion approach, where perspective grounding guides the grounding tasks. Additionally, we present CAESAR-PRO, a synthetic dataset of embodied referring expressions with multimodal cues, to facilitate further research in this area. However, this work has some limitations in different real-world settings, which can be addressed to develop a robust model.

Firstly, the primary limitation of this study is the use of a synthetic dataset, CAESAR-PRO, for training and evaluating our model, PATRON. Although this dataset has been carefully curated to include a variety of embodied referring expressions with multimodal cues, it may not fully capture the complexity and diversity of human nonverbal and verbal expressions in real-world settings. Synthetic datasets, by their nature, are simulations and may lack the nuances, variability, and unpredictability inherent in human behavior.

Secondly, the model's performance in real-world settings remains untested. While PATRON outperforms other state-of-the-art visual-language models in our synthetic dataset, its effectiveness in real-world applications is yet to be determined. The transition from a controlled, synthetic environment to a dynamic, unpredictable real-world setting may pose challenges that were not encountered during the training phase.

Lastly, while our model has shown promise in understanding and grounding embodied referring expressions, it is currently limited to the specific tasks of relation and object grounding, such as question answering in the embodied settings.

In the future work, we aim to address these limitations. This involves testing and refining the model using real-world data, expanding the model's capabilities to include a broader range of tasks, and exploring ways to simulate better the complexity of human nonverbal and verbal expressions in synthetic datasets. Despite these limitations, we believe that our study provides a valuable foundation for future research in this area.

Chapter 6
EMBODIED QUESTION ANSWERING USING MULTIMODAL EXPRESSION

For an autonomous agent to seamlessly collaborate with people, it is vital for agents to comprehensively understand human instructions [24], [25], [71], [146]. To develop models to comprehend human instructions, several tasks have been designed, such as referring expression comprehension [45], [52], [53], [150], [171], spatial relations grounding [20], [21], [51]–[54], [68], [164], [172], [173], and visual question answering [22], [23], [174]–[181]. Among these tasks, visual question answering (VQA) has been widely studied, as it requires complex reasoning of several sub-tasks, such as answering verbal questions revolving around an object's presence and category using visual context [20], [176], [182].

Although many synthetic and real-world datasets have been developed for VQA, one of the crucial shortcomings of these datasets is that the questions are solely based on verbal utterances. This differs from how people naturally use multimodal expressions (verbal utterances and nonverbal gestures) while asking questions. Additionally, many studies indicate that nonverbal gestures often provide complementary information to seamlessly understand a verbal question [24], [25], [55]–[63], [71], [79], [80]. For instance, in a visual scene containing two balls with different colors, a pointing gesture can provide additional information to answer questions such as "what is the color of that ball?" Thus, the lack of nonverbal interactions in prior VQA datasets makes them less suitable for developing models to comprehend question-answering (QA) tasks in real-world interactions.

Following VQA, embodied question-answering (EQA) tasks have recently been studied in the literature [70], [75]–[78]. Based on the definition of embodied interactions, EQA can be designed in two ways. The first type of embodied interaction is defined from an agent's perspective, such as a virtual robot, where the agent navigates in an environment to answer verbal questions [70]. These works solely incorporate verbal questions. The second type of embodied interaction refers to multimodal expressions, where a human interacts with the environment using verbal utterances and nonverbal gestures [24], [25], [71]. Adopting the latter definition, we have designed embodied question-answering (EQA) tasks as comprehending questions by utilizing multimodal expressions (verbal utterances and nonverbal gestures) in embodied settings. For example, an EQA task can involve pointing to an object and asking "what is that object?" In this context, the EQA task requires reasoning over multimodal expressions to answer the question.

Another crucial shortcoming in most existing VQA and EQA datasets is that verbal utterances are from a single perspective (speaker or observer). This differs from real-world interactions where people use both perspectives interchangeably. Consider a question from the speaker's perspective, "What is the object to the *right of red mug*?" In this question, *right of red mug* can be considered as the *left of red mug* from the observer's perspective. The absence of data from multiple perspectives

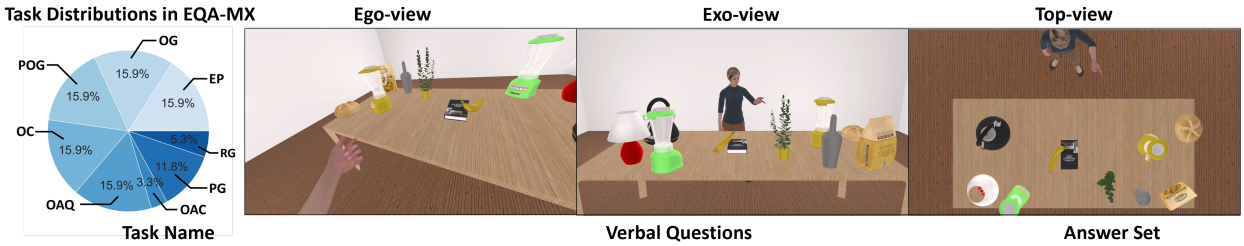
Datasets	V	NV	E	EQA	MT	MV	Views			C	A	No. of Images	No. of Samples	Object Categories	Avg. Words*
							Exo	Ego	Top						
PointAt [147]	✗	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗	220	220	28	-
ReferAt [64]	✓	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗	242	242	28	-
IPO [148]	✗	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗	278	278	10	-
IMHF [149]	✗	✓	✓	✗	✗	✗	✓	✗	✗	✗	✗	1716	1716	28	-
RefIt [69]	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	19,894	130,525	238	3.61
RefCOCO [150]	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	19,994	142,209	80	3.61
RefCOCO+ [150]	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	19,992	141,564	80	3.53
RefCOCOg [151]	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	26,711	104,560	80	8.43
Flickr30k [152]	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	31,783	158,280	44,518	-
GuessWhat? [153]	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	66,537	155,280	-	-
Cops-Ref [154]	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	75,299	148,712	508	14.40
CLEVR-Ref+ [22]	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	99,992	998,743	3	22.40
DAQUAR [183]	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	1449	124,68	37	11.5
FM-IQA [177]	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	157,392	316,193	-	7.38
Visual Madlibs [178]	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	107,38	360,001	-	6.9
Visual Genome [180]	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	108,000	1,445,332	37	5.7
DVQA [181]	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	300,000	3,487,194	-	-
VQA (COCO) [176]	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	204,721	614,163	80	6.2
VQA (Abs.) [176]	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	50,000	150,000	100	6.2
Visual 7W [179]	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	47,300	327,939	36,579	6.9
KB-VQA [184]	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	700	5826	23	6.8
FBQA [185]	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	2190	5826	32	9.5
VQA-MED [186]	✓	✗	✗	✗	✗	✗	✓	✗	✗	✗	✗	2866	6413	-	-
DocVQA [187]	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	12,767	50,000	-	-
YouRefIt [24]	✓	✓	✓	✗	✓	✗	✓	✗	✗	✗	✗	497,348	4,195	395	3.73
GRiD-3D [20]	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	8,000	445,000	28	-
EQA † [70]	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	5,000	5,000	50	-
MT-EQA † [70]	✓	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗	19,287	19,287	61	-
CAESAR-L [25]	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	✓	11,617,626	124,412	61	5.56
CAESAR-XL [25]	✓	✓	✓	✗	✗	✓	✓	✓	✓	✓	✓	841,620	1,367,305	80	5.32
EQA-MX ‡	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	750, 849	8, 243, 893	52	11.45

Table 6.1: Comparison of the QA datasets. Existing VQA and EQA datasets do not contain non-verbal human gestures (NV), multiple verbal perspectives (MV), contrastive (C) and ambiguous (A) data samples. ‡ Embodied (E) interactions refer to humans interacting with multimodal expressions. † Embodied interactions refer to an agent navigating in an environment. *Average number of words in questions. V: Verbal and MT: Multitasks.

in the existing datasets hinders the development of robust QA models.

Existing models for VQA and EQA tasks answer verbal questions from a single verbal and visual perspective [26], [27], [91], [188]. As multiple views can provide complementary information and interactions can be captured from different camera angles, aligning these visual representations before fusing them with verbal representations can help to learn generalized representations and comprehend interactions from different camera views robustly. Furthermore, existing models fuse continuous visual representations with discrete verbal representations. This inconsistency of embedding structures can lead to sub-optimal representations.

To address the shortcomings of existing VQA and EQA datasets, we have extended an embodied simulator to develop a large-scale novel dataset, EQA-MX, for training and diagnosing models



Existence Prediction [EP]	**Q**: Is there a plant in the scene? **A**: Yes	Yes/No
Object Grounding [OG]	**Q**: What is the name of that yellow object? **A**: Blender	All 52 objects
Perspective-Aware Object Grounding [POG]	**Q**: Considering the observer's perspective, what is the name of the yellow thing next to the bread loaf? **A**: Blender	All 52 objects
Object Counting [OC]	**Q**: How many objects are left of the yellow object? **A**: 7	{0-9}
Object Attribute Query [OAQ]	**Q**: What is the color of that thing? **A**: Yellow	All 12 colors
Object Attribute Compare [OAC]	**Q**: Is the color of that object the same color as the kettle? **A**: No	Yes/No
Perspective Grounding [PG]	**Q**: The blender to the right of the book. From which perspective is the object described? **A**: Speaker	Speaker/Observer/Neutral
Relation Grounding [RG]	**Q**: Considering the observer's perspective, the left yellow blender to the right of the bread, is the object referred to appropriately? **A**: No	Yes/No

Figure 6.1: EQA tasks for sample data from EQA-MX. Top-row: data distribution for each task in EQA-MX (left) and an embodied interaction with multiple visual perspectives (right). Bottom-row: name of the task (left), example questions and answers for the given task based on the visual scene above (middle), and the set of possible answers (right).

for comprehending EQA tasks (Table 6.1). Our simulator can be used to procedurally generate nonverbal interactions (gaze and pointing gestures) and verbal utterances in multiple embodied environments for different EQA tasks. We have generated and annotated data under an approved IRB (protocol number: 4627, Title: Understanding Multimodal Human Instruction in Embodied Environment). We have addressed the limitations of existing multimodal fusion approaches and developed a multimodal learning model for EQA tasks, VQ-Fusion, using vector quantization (VQ) [189], [190]. The VQ-based bottleneck plays a key role in disentangling the continuous visual representations into discrete embeddings and enables salient fusion with discrete verbal representations. We use a shared codebook in VQ to align multiview representations and learn the unified concept shared among multiple views. We highlight our *key contributions* below:

- We have developed a large-scale novel dataset (EQA-MX) that includes questions with multimodal expressions from multiple verbal perspectives.
- We have captured the nonverbal interactions from multiple visual perspectives to reduce the model's verbal and visual perspective bias.
- We have designed 8 new EQA tasks involving questions with multimodal expressions (verbal utterances and nonverbal gestures) that need to be answered using the visual context in an

embodied environment.

- To address the issue of fusing two different embedding structures (continuous visual and discrete verbal representations), we have developed a VQ-based multimodal learning model to learn salient representations from multiple visual and verbal perspectives.
- Our extensive experimental analyses indicate that our proposed model, VQ-Fusion, can help to improve the performance of EQA tasks up to 13%.

6.1 Embodied Question Answering Tasks

We have created 8 novel EQA tasks: Existence Prediction (EP), Object Grounding (OG), Perspective-Aware Object Grounding (POG), Object Counting (OC), Object Attribute Query (OAQ), Object Attribute Compare (OAC), Perspective Grounding (PG), and Relation Grounding (RG). Similar tasks have been developed in prior works [20], [150], [174], [176], [179], [185], however, those tasks involve only verbal questions. We are the first to design QA tasks in embodied settings where a human avatar asks questions using verbal utterances and nonverbal gestures in a virtual environment. Each of these tasks has multiple sub-templates for variation. In Fig. 6.1, we provide samples of these EQA tasks.

Existence Prediction (EP): Naturally, humans are able to determine what objects are present in a given scene. In scenarios where humans are interacting and an actor mistakenly references an object not in the scene, this allows observers to request more information. Created to mimic this situation, the existence prediction task involves determining whether the scene contains a particular object with some specific attributes, such as color.

Object Grounding (OG): Understanding which objects a human refers to using verbal and nonverbal cues is key to successful human-AI interaction. A model successfully able to ground objects has use cases such as assisting surgeons during a procedure by handing surgeons the correct tools. Thus, we design the object grounding task around this scenario, where models must identify the name of the object being referred to by verbal and nonverbal expressions.

Perspective-Aware Object Grounding (POG): Similar to the object grounding task, Perspective-Aware Object Grounding involves determining which object is being referred to, but this task includes the verbal perspective (either ego, exo, or neutral). Although real-world human-AI interactions will not always contain the perspective of a given relation, including the perspective allows us to determine whether or not understanding perspective can help in grounding objects.

Object Counting (OC): As understanding what object a human is referring to in a scene involves interpretation of the different number of objects inside that scene, understanding the number of objects in a scene can serve as an auxiliary task for the object grounding task. If models are able to create salient multimodal representations to attend to all the objects in a given scene, is is likely they will be able to ground particular objects better. Thus, in the object counting task the number of objects in a scene is asked based on different spatial relations.

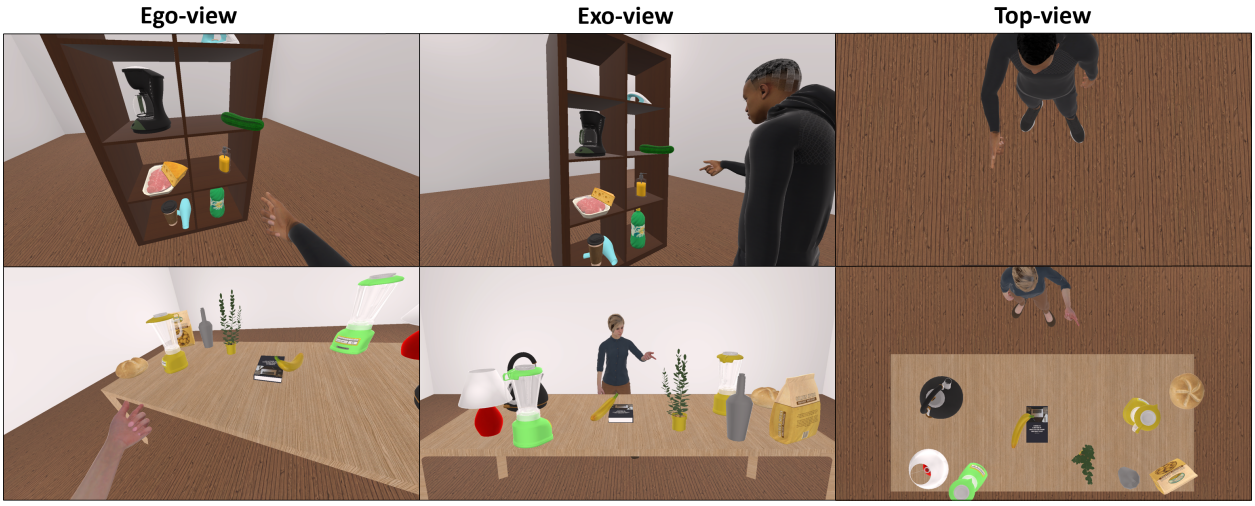


Figure 6.2: Sample data demonstrating the shelf environment vs. the table environment

Object Attribute Query (OAQ): It is often important in human-human interactions to identify particular attributes of objects. Additionally, this information can be used as auxiliary information for tasks such as the Object Grounding task, where the goal is to identify objects. We design the Object Attribute Query task around this particular situation, where the color of a given object is queried for.

Object Attribute Compare (OAC): Humans often exchange information throughout conversations through the use of comparison of different object attributes. This exchange of information can assist in understanding the different objects an actor is referring to. Thus, we design the object attribute compare task, where the attributes of two different objects in the scene are compared.

Perspective Grounding (PG): Understanding human verbal perspective is integral to successful human-AI communication, as humans interchangeably describe objects from their perspective as well as the perspective of others. We simulate this in the perspective grounding task using three different perspectives - neutral, egocentric (speaker), and exocentric (observer).

Relation Grounding (RG): As described in [25], the relation grounding task involves determining whether the supplied verbal and nonverbal signals align with respect to describing the same object. Understanding whether or not a human is accurately verbally and nonverbally referring to an object can enable the identification of human mistakes. We add complexity to this task through the variation of verbal perspective in the question.

6.1.1 EQA Task Templates

In this work we presented 8 EQA tasks. Each of these tasks has multiple sub-templates, which we present in more detail in Table 6.2. Each sub-template has multiple degrees of freedom from which to vary, ensuring generated embodied questions are diverse. For example, since most sub-

Task Name	Template	Task Example
Existence Prediction	Is there any/a/an <object name> in the scene?	Is there any cucumber in the scene?
	Is there any/a/an <object color> <object name> in the scene?	Is there any green cucumber in the scene?
Object Grounding	What is the name of that object/thing?	What is the name of that object?
	What is the name of the <object color> object/thing?	What is the name of that yellow thing?
	What is the name of that <object absolute location> <object color> object/thing?	What is the name of that right yellow object?
	What is the name of that <selected object absolute location> <selected object color> object/thing to the <spatial relation> of the <relational object absolute location> <relational object color> <relational object name>?	What is the name of that right yellow object to the right of the yellow cheese?
Perspective-Aware Object Grounding	Considering the <observer's/speaker's> perspective, what is the name of that object/thing?	Considering the observer's perspective, what is the name of that object?
	Considering the <observer's/speaker's> perspective, what is the name of the <object color> object/thing?	Considering the observer's perspective, what is the name of that yellow thing?
	Considering the <observer's/speaker's> perspective, what is the name of that <object absolute location> <object color> object/thing?	Considering the speaker's perspective, what is the name of that right yellow object?
Object Counting	How many objects are <spatial relation> of the object/thing?	How many objects are above the object?
	How many objects are <spatial relation> of the <object color> object/thing?	How many objects are left of the yellow thing?
Object Attribute Query	What it the color of that object/thing?	What it the color of that object/thing?
	What is the color of the <object name>?	What is the color of the hand soap dispenser?
Object Attribute Compare	Is the color of that object/thing the same color as the <relational object name>?	Is the color of that thing the same color as the cheese?
	Is the color of that <selected object name> the same color as the <relational object name>?	Is the color of that hand soap dispenser the same color as the soda bottle?
Perspective Grounding	<Referring expressions using the templates from CAESAR>. From which perspective is the object described?	The hand soap dispenser above the soda bottle. From which perspective is the object described?
Relation Grounding	<Referring expressions using the templates from CAESAR>, is the object referred to appropriately?	The hand soap dispenser above the cucumber, is the object referred to appropriately?
	Considering the observer's perspective, <Referring expressions using the templates from CAESAR>, is the object referred to appropriately?	Considering the observer's perspective, the hand soap dispenser below the cucumber, is the object referred to appropriately?
	Considering the speaker's perspective, <Referring expressions using the templates from CAESAR>, is the object referred to appropriately?	Considering the observer's perspective, the hand soap next to the coffee maker, is the object referred to appropriately?

Table 6.2: Templates for all 8 tasks in the EQA-MX dataset. The answers for these templates are based on the environment in the first row of Figure 6.2.

templates use the absolute location of an object, this absolute location can often times be described from either the observer or speaker perspective.

6.1.2 New Environments in EQA-MX

To increase dataset generalizability, we have added a shelf environment into the CAESAR simulator, and thus into the EQA-MX dataset. We visualize the three views (ego, exo, and top) for this and the table environment in Fig. 6.2. Because the exo and ego views in the table environment are on different sides of the table, the verbal perspectives differ. However, in the shelf environment, the exo and ego views are aligned meaning the verbal perspective is aligned. We created this environment in this way to ensure models have differing situations with regards to views and

Splits	EP	OG	POG	OC	OAQ	OAC	PG	RG
Train	1060k	1060k	1060k	1060k	1060k	218k	785k	349k
Valid	126k	126k	126k	126k	126k	27k	93k	41k
Test	126k	126k	126k	126k	126k	28k	93k	42k

Table 6.3: EQA-MX dataset splits for 8 EQA tasks.

perspective. Additionally, since the shelf has objects below/on top of one another, it adds diversity with respect to spatial relations/locations, ensuring models understanding these relations/locations in all 3 dimensions.

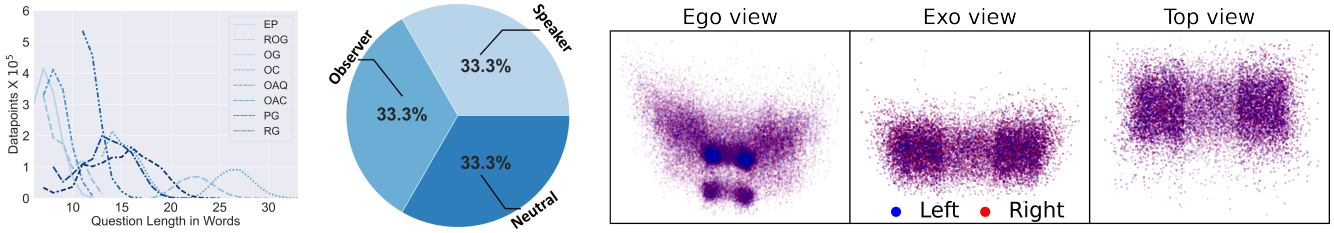
6.2 Dataset Generation with EQA Simulator

In this work, we have extended the CAESAR simulator [25] to generate data for different EQA tasks. CAESAR is used to randomly generate environments where an actor simulates nonverbal expressions through a pointing gesture and gaze in a scene (Fig. 6.3). Verbal expressions are created based on the visual scene. To increase the dataset's generalizability, we have used multiple environments. These environments differ in terms of camera views, object locations, and nonverbal/verbal expressions. In each visual scene, we generated four different situations, 1) a situation with no human and therefore no nonverbal expressions, 2) a situation with a human head gaze, 3) a situation with a human pointing gesture, and 4) a situation involving a human using a head gaze and a pointing gesture.

Generated nonverbal expressions consist of a pointing gesture and gaze. Pointing gestures are procedurally generated using inverse kinematics through the Unity engine. We create these pointing gestures based on random noise added onto real-world data of human pointing gestures captured using an Optitrack motion capture system [90]. Similarly, we have simulated human head gazes using inverse kinematics and an object location within the scene as a target. Verbal questions are generated based on different templates for each EQA task. The nonverbal and verbal expressions may describe the same object, or be contrastive, meaning the nonverbal and verbal expressions describe different objects. We use these contrastive instructions for the Relation Grounding task. Additionally, the absence of nonverbal gestures in situations with no humans generates ambiguous data samples.

6.3 Dataset Analysis

We have generated a novel large-scale dataset, EQA-MX, containing $8,243,893$ samples across the 8 tasks described in Sect. 6.1. The training, validation, and test set splits for each of these tasks is shown in Table 6.3. We removed some data samples to generate balanced dataset splits for the OAC, PG, and RG tasks.



(a) Lengths in words of all questions. (b) Verbal perspective (PG task). (c) Object locations with respect to spatial relations.

Figure 6.3: Dataset Analyses for the EQA-MX dataset. (a) demonstrates how the EQA-MX dataset contains questions with different lengths in words and thus amounts of contextual information for all the EQA tasks. (b) shows the ratios of data samples with different verbal perspectives for the perspective grounding (PG) task. (c) shows the object locations with respect to different spatial relations. As the object locations are not separable, the EQA-MX dataset is non-biased with respect to verbal and visual perspectives.

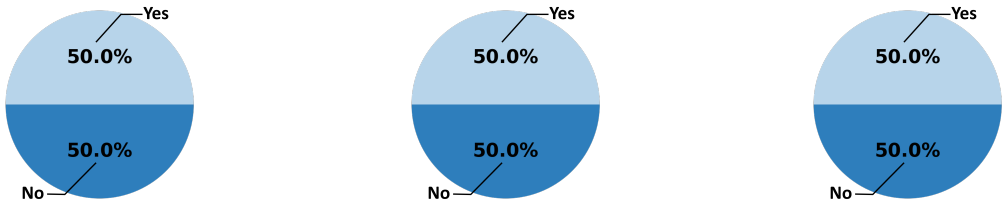
Our designed EQA tasks vary in terms of the goals (Fig. 6.1) and visual-verbal contextual information in the questions. This is made apparent by the variance in question lengths in words (Fig. 6.3(a)). Questions are as short as 6 words for the EP task and as long as 34 words for the OG task. Additionally, one of the main focuses of the EQA-MX dataset is to introduce data that varies in verbal and visual perspectives. Fig. 6.3(b) demonstrates the PG task's outcome of different verbal perspectives.

Similarly, Fig. 6.3(c) shows the location of objects based on spatial relations in questions from verbal perspectives. Fig. 6.3(c) also demonstrates how objects being referred to as on the left (blue) and right (red) are not linearly separable through the use of spatial relations, as different verbal perspectives use different relations to describe an object. For example, consider a speaker describing the red table lamp in Fig. 6.1. The speaker could state "the red lamp on the left". However, from the observer's perspective (exo view) the table lamp is on the right. Thus, given the verbal perspectives, spatial relations are non-separable in EQA-MX (Fig. 6.3(b)). This reduced verbal and visual perspective biases in EQA-MX dataset can help train robust models for comprehensively comprehending EQA tasks.

6.3.1 Task Output Distributions

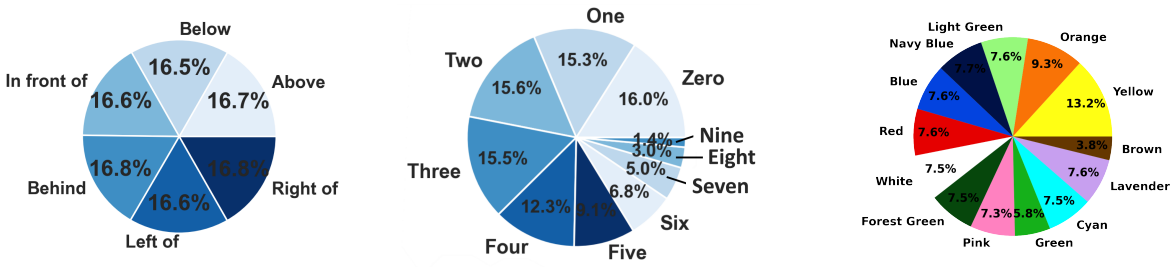
As shown in Figs. 6.4,6.6 we balance outputs of our task distributions where possible in order to ensure the EQA-MX dataset is not biased. For the OG and POG tasks, the output distribution of all 52 categories is balanced to ensure models do not bias a particular object.

Additionally, in Fig. 6.4, all binary tasks (EP, OAC, and PG) contain a $50/50$ split between *yes* and *no* answers. Because the CAESAR simulator randomly generates scenes populated with objects, the OC and OAC tasks do not have even task distributions. This can be explained by these



(a) Existence Prediction Task (b) Object Attribute Compare Task (b) Relation Grounding Task

Figure 6.4: Distributions of task outputs in the existence prediction (EP), object attribute compare (OAC), and relation grounding (RG) tasks. All these tasks have balanced binary outputs



(a) OC task spatial relations (b) Distribution of OC task output (c) Distribution of OAC task output

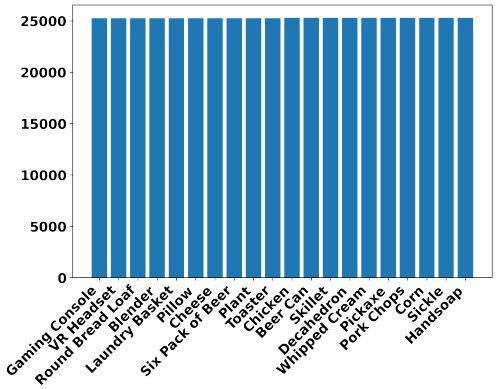
Figure 6.5: Distribution of task outputs in the object counting and object attribute compare tasks. Both distributions are not completely even due to different observed scene probabilities. For the object counting (OC) task, lower numbers have higher probabilities of occurring due to the number of objects in the scene ranging from 4 - 10, hence the imbalance in distributions. Similarly, in the object attribute compare task different object colors are queried for, and since the colors of objects is not completely balanced, the task distribution is imbalanced.

tasks involving observed characteristics in scenes where some characteristics are more common than others. For example, since the max number of objects that can be generated in a scene is 10, the probability of an object have 9 objects to the left of it is much lower than the probability of an object having 2 objects to the left of it. Similarly, certain colors are more common in objects inside of the CAESAR simulator. These distributions are made more apparent in Fig. 6.5 (we report macro accuracy for models trained on these tasks).

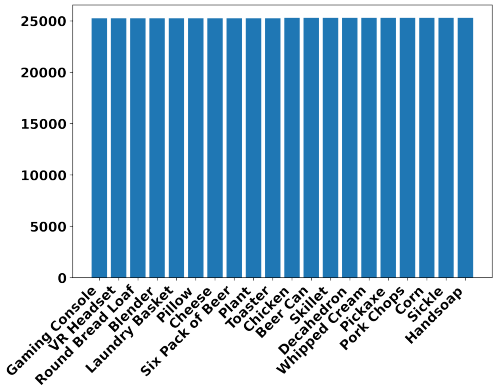
6.3.2 Object Locations Analyses

We visualize object locations inside the EQA-MX dataset to show how different spatial relations have/don't have bias (Fig. 6.7). Particularly, since one of our contributions is the creation of the

(a) Wordcloud based on all verbal expressions



(b) Distribution of outputs for the OG task



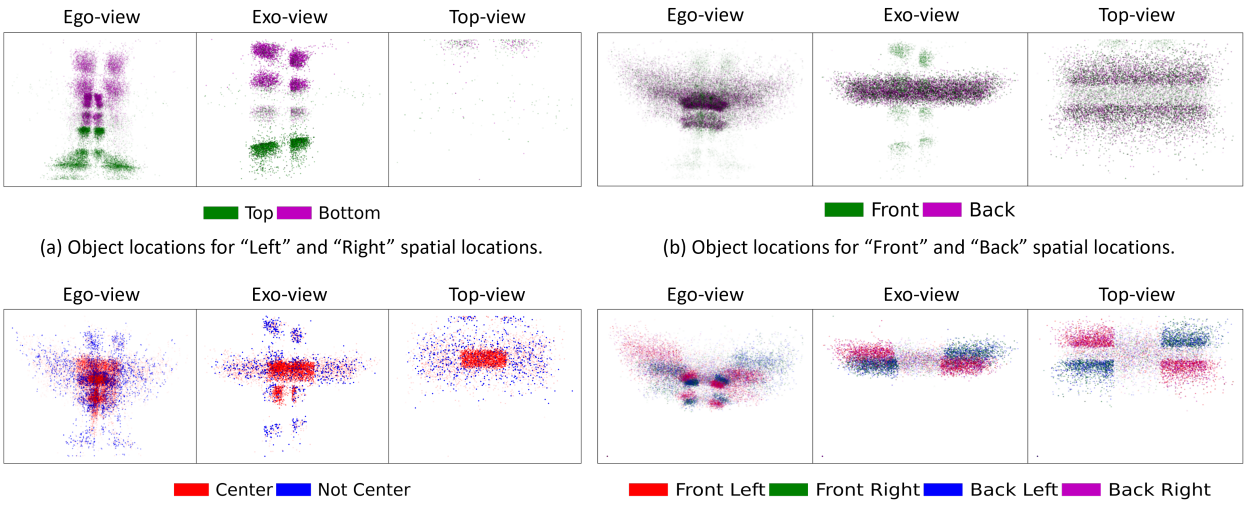
(c) Distribution of outputs for the POG task

Figure 6.6: A verbal expression Wordcloud for the EQA-MX dataset, as well as the output distribution for the object grounding (OG) and perspective-aware object grounding (POG) tasks. In the Wordcloud the size of words represents the frequencies that they occur in the verbal utterances. Therefore, the most frequent words describe general properties of objects or are general words inside questions - such as color, perspective, and spatial relations/locations. In the diagrams for object frequencies for the object grounding and perspective-aware object grounding tasks, the most referred objects all have the same frequencies (these tasks have the same object distributions).

shelf environment, we show how since its visual views are aligned certain visual cues have bias.

6.4 VQ-Fusion: VQ-based Multimodal Fusion

We develop a vector quantization-based multimodal fusion approach, VQ-Fusion, to learn visual-language representations. As EQA tasks in EQA-MX involve multiple visual views, VQ-Fusion extracts visual representations from multiple visual views (X_{ego}, X_{exo}, and X_{top}) and verbal ques-



(a) Object locations for "Left" and "Right" spatial locations.

(b) Object locations for "Front" and "Back" spatial locations.

(c) Object locations for "Center" spatial locations (Table-top).

(d) Object locations for "Corner" spatial location.

Figure 6.7: Object locations visualized for different spatial relations/locations across the EQA-MX dataset. The object locations are not easily separable based on spatial relations/locations that vary based on perspectives. (a & b) demonstrates how the shelf environment has more non-separable locations/relations due to the fact that verbal perspective in the shelf environment does not vary based on visual perspective. c is generally linearly separable, as expected, as the center of a given scene is objective. d demonstrates how opposing corners (i.e. front left and back right) are non-separable due to varying based on verbal perspectives).

tions (X_q) for different EQA tasks (Fig. 6.8 and Sect. 6.1). Following the existing adapter-based learning models [191]–[196], we design VQ-Fusion as an adapter model that can be used in existing models without significantly changing the existing model architecture.

Visual and Language Representation Learning: At first, VQ-Fusion extracts visual and language representations using a state-of-the-art visual encoder (e.g., ResNet [98] and ViT [166]) and language model (e.g., BERT [167]). VQ-Fusion uses shared models to extract the visual representations from multiple views independently:

$$E_m = F_m(X_m) \quad , \quad m \in (ego, exo, top, verbal) \tag{6.1}$$

Here, F_m is the visual or verbal encoders, $E_m \in \mathbb{R}^{D_m}$, and D_m is the representation dimension of modality m.

Discretization and Multimodal Fusion: Language models create discretized representations, whereas visual encoders produce continuous representations of visual scenes. Fusing these representations with different embedding structures can lead to sub-optimal multimodal representations

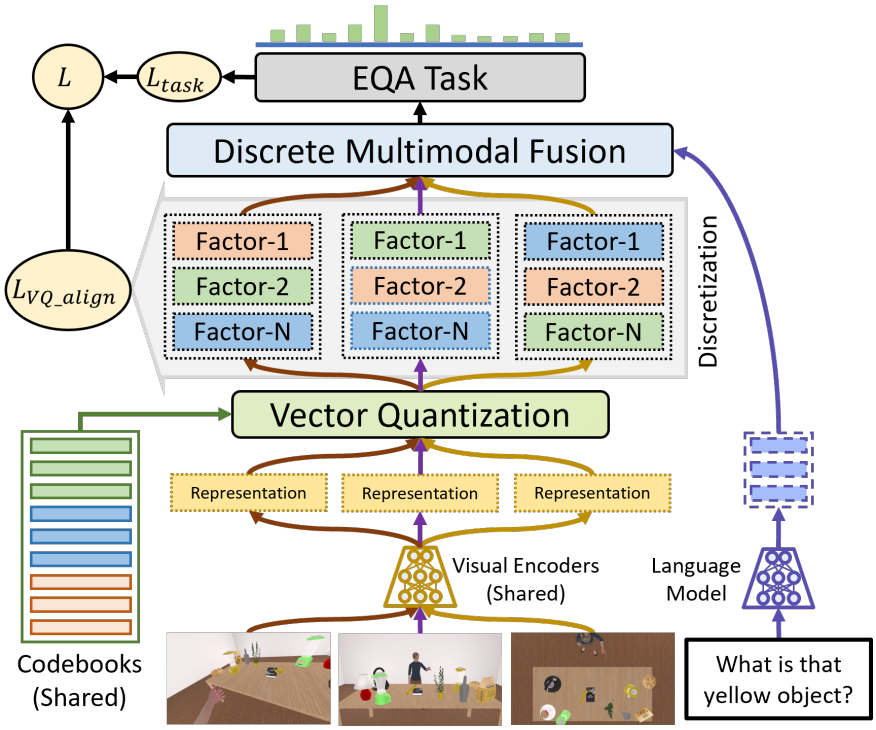


Figure 6.8: VQ-Fusion: Vector Quantization (VQ) based multimodal learning model architecture. VQ-Fusion extracts multiview visual representations using visual encoders, which are then discretized using shared codebooks. The shared codebooks' bottleneck allows the model to learn unified concepts across multiple views. Finally, discretized visual representations are fused with discrete verbal representations to produce multimodal representation.

[197]. For this reason, we discretize the visual representations before multimodal fusion.

In VQ-Fusion, we adopted the vector quantization (VQ) method from VQ-VAE [189] and Discrete-Value Neural Communication [190] works to discretize multiview visual representations, $E_m \in (E_{ego}, E_{exo}, E_{top})$. Previous works use VQ to discretize a representation using codebooks, whereas we use shared codebooks to discretize and align multiview representations to learn unified concepts across visual views for extracting salient multimodal representations. First, VQ-Fusion divides each E_m into G continuous segments $(s_{(m,1)}, s_{(m,2)}, \ldots, s_{(m,G)})$, where $E_m = \text{CONCAT}(s_{(m,1)}, s_{(m,2)}, \ldots, s_{(m,G)})$ and $s_{(m,i)} \in \mathbb{R}^{D_m/G}$. Second, VQ-Fusion independently maps continuous segment $s_{(m,i)}$ to discrete latent code $c_j \in \mathbb{R}^{L\times(D_m/G)}$ using shared codebooks C, where L is codebooks size (i.e., number of categorical codes in each codebook). We can find the optimal code for each continuous segment $s_{(m,i)}$ from the codebooks in the following way:

$$e_{(m,o_i)} = F^D(s_{(m,i)}), \quad o_i = \underset{j\in 1\ldots L}{\arg\max} ||s_{(m,i)} - c_j|| \tag{6.2}$$

Here, F^D is the discretization (D) method. Finally, we concatenate the discretized codes to produce discretized visual representation E_m^D in the following way:

$$E_m^D = \text{CONCAT}(F^D(s_{(m,1)}), \ldots, F^D(s_{(m,G)})) \tag{6.3}$$

Following the training procedure in [190] and [189], we calculate VQ loss to learn the codebooks:

$$\mathcal{L}_{VQ\_align} = \frac{\beta}{G} \sum_{i}^{G} ||s_i - sg(c_{o_i})||_2^2 \tag{6.4}$$

Here, sg is the stop-gradient operator that blocks gradients from flowing into c_{o_i}, and β is a hyperparameter that controls reluctance to change the code. We train the discretization module to learn codebooks using gradient descent with the other parts of VQ-Fusion. Additionally, as VQ-Fusion uses shared codebooks to discretize visual representation for multiple views, $\mathcal{L}_{VQ\_align}$ loss also guides the model to align multiview representations and learn unified concepts across views. This shared codebooks approach allows aligning multiview representation to answer the question with multimodal expressions effectively.

Finally, VQ-Fusion fuses these discretized visual and verbal representations using a self-attention approach to produce task representation E_{fused}:

$$E_{fused} = \sum_{m \in M} \alpha_m E_m \tag{6.5}$$

$$\alpha_m = \frac{exp(\gamma_m)}{\sum_{m \in M} exp(\gamma_m)}, m \in M \tag{6.6}$$

$$\gamma_m = (W)^T E_m, m \in M \tag{6.7}$$

Here, M is the modality list (ego, exo, top, verbal), W is a learnable parameter, and α_m is the attention score which is calculated using a 1D-CNN with a filter size of 1.

6.4.1 Task Learning

We use the fused representation, E_{fused}, to learn different EQA tasks T_k:

$$y_{T_k} = F_{T_k}(E_{fused}) \tag{6.8}$$

$$\mathcal{L}_{task,T_k}(y_{T_k}, \hat{y}_{T_k}) = \frac{1}{B} \sum_{i=1}^{B} y_{(T_k,i)} \log \hat{y}_{(T_k,i)} \tag{6.9}$$

Models	EP		OG		POG		OC	
	✗	✓	✗	✓	✗	✓	✗	✓
Dual Encoder	53.46	**55.78**	48.31	**49.96**	83.91	**84.28**	12.28	**12.38**
CLIP	53.17	**54.72**	54.06	**65.49**	70.92	**82.70**	09.65	**13.14**
VisualBERT	50.00	**54.51**	53.39	**54.50**	86.09	**87.09**	14.09	**14.35**
ViLT	90.24	**91.50**	59.74	**61.04**	86.10	**87.42**	11.14	**12.54**

Models	OAQ		OAC		PG		RG	
	✗	✓	✗	✓	✗	✓	✗	✓
Dual Encoder	63.71	**66.90**	57.92	**61.45**	66.72	**66.77**	75.78	**89.36**
CLIP	70.85	**74.32**	58.59	**70.59**	66.64	**66.99**	85.84	**89.93**
VisualBERT	51.43	**54.45**	58.56	**59.98**	66.37	**79.11**	89.13	**89.26**
ViLT	55.96	**59.47**	58.93	**60.16**	80.36	**81.23**	87.36	**88.68**

Table 6.4: Comparisons of VL models performance for EQA tasks. The results suggest that incorporating VQ-Fusion in VL models can improve the performance of EQA tasks. ✓: VL models with VQ-Fusion, and ✗: VL models without VQ-Fusion.

Here, F_{T_k} is the task learning module, which can be designed based on the EQA task properties. For example, we use a multi-layer perceptron for the object existence task (SectionSe 6.1). Moreover, $\mathcal{L}_{task,T_k}$ is the task learning loss of task T_k. Finally, we combine the task learning loss ($\mathcal{L}_{task,T_k}$) with the VQ loss ($\mathcal{L}_{VQ\_align}$) using task learning weights ($\mathcal{W}_{VQ}$ and $\mathcal{W}_{task}$) to train the VQ-Fusion model:

$$\mathcal{L} = \mathcal{W}_{VQ}\mathcal{L}_{VQ\_align} + \mathcal{W}_{task}\mathcal{L}_{task,T_k} \tag{6.10}$$

Variations of VQ-Fusion: VQ-Fusion allows to use state-of-the-art VL models (e.g., VisualBERT [26] & ViLT [91]) to extract these representations. As the architecture of these VL transformer models is limited to processing a single visual and verbal input, we need to pair the verbal question to each visual view and pass through these models to extract multiview visual and verbal representations. We use these representations in VQ-Fusion to discretize and fuse to produce multimodal representations.

6.5 Experimental Analysis

In this section, we have presented experimental analyses on our EQA-MX dataset to evaluate the impact of VQ-Fusion in VL models for EQA tasks. We have also conducted additional ablation studies and experimental analyses for human activity recognition task to evaluate the significance of VQ-Fusion for multimodal representation learning.

6.5.1 Baseline Models

Existing visual-language (VL) models for QA tasks are designed to answer a question using a single visual context. Since our proposed EQA tasks involve three visual views, we extend four VL models to learn multiview representations: Dual-Encoder (ViT+BERT) [166], [167], CLIP [92], VisualBERT [26], and ViLT [91]. For the Dual-Encoder (ViT+BERT) model, we independently extract visual representations for each view using a shared ViT model and verbal representations using a BERT model. We fuse these visual and verbal representations to produce task representations. For the CLIP models, we pair each visual view to a verbal question and pass this through the model to extract multiple visual and verbal representations and fuse them to produce task representations. For VisualBERT and ViLT, we use ResNet-101 [98] to extract visual representations that are passed through the model with verbal embeddings to produce task representations.

6.5.2 Training Setup

We developed all the models using the Pytorch (version: 1.12.1+cu113) [100] and Pytorch-Lightning (version: 1.7.1) [120] deep learning frameworks. We also used HuggingFace library (version: 4.21.1) for pre-trained models (BERT [1] [167], ViT [2] [166], VisualBERT [3] [26], Dual Encoder [4], ViLT [5][91], and CLIP [6] [92]). For the Dual-Encoder and CLIP models, we used an embedding size of 512, and for VisualBERT and ViLT, we used an embedding size of 768. We train models using the Adam optimizer with a weight decay regularization [95] and cosine annealing warm restarts at an initial learning rate: $3e^{-4}$, cycle length (T_0): 4, and cycle multiplier (T_{mult}): 2. We used batch size 128 and trained models for 8 epochs. We used the same fixed random seed (33) for all the experiments to ensure reproducibility. Lastly, all models are trained in distributed GPU clusters, where each node contains 8 A100 GPUs.

6.5.3 Comparison of Multimodal Learning Models

We evaluated state-of-the-art visual-language (VL) models with and without our VQ-Fusion to learn VL representations for 8 EQA tasks. We varied the number of codebooks to $\{2, 4, 8, 16\}$ in VQ for each task and reported the best performance. We trained and evaluated these models independently for each task as a single-task model on our EQA-MX dataset. We used data samples with varying nonverbal gestures: gaze and pointing gestures, only gaze, and only pointing gestures. All the visual views (ego, exo, and top) and verbal perspectives (speaker, observer, and neutral) are used to train models and evaluate whether the models can learn generalized representation from

[1] https://huggingface.co/docs/transformers/model_doc/bert
[2] https://huggingface.co/docs/transformers/model_doc/vit
[3] https://huggingface.co/docs/transformers/model_doc/visual_bert
[4] https://huggingface.co/docs/transformers/model_doc/vision-text-dual-encoder
[5] https://huggingface.co/docs/transformers/model_doc/vilt
[6] https://huggingface.co/docs/transformers/model_doc/clip

diverse data. We report macro-accuracy across all tasks to accurately gauge whether models can effectively understand EQA tasks and are not biased toward a particular class (Table 6.4).

Results: The results in Table 6.4 suggest that incorporating VQ-Fusion in VL models helps to successfully fuse extracted salient multiview representations with verbal representations, and thus improves model performance on EQA tasks. For example, the CLIP model without VQ-Fusion achieves 54.06% accuracy in the object grounding task (OG), whereas incorporating VQ-Fusion in the CLIP model increases the OG task's performance to 65.49%. Similarly, VQ-Fusion improved the CLIP model's performance on the object attribute query task (OAQ) by 12%, the VisualBERT model's performance on the perspective grounding task by 12.74%, the ViLT model's performance on the object attribute comparison (OAC) task by 3.5%, and the DualEncoder model's performance on the relation grounding task (RG) by 13.58%. These performance improvements validate the significance of VQ-Fusion in extracting salient multimodal representations from multiple visual and verbal perspectives for effectively learning EQA tasks.

Discussion: The primary reasoning behind the performance improvement by incorporating VQ-Fusion in VL models is that VQ-Fusion discretizes the multiview representations before fusing with the discrete verbal representations. VQ-Fusion uses codebooks to discretize the visual representations to be similar to the discrete verbal representation structure. On the other hand, existing VL models extract continuous wrappings of monolithic visual representations and fuse it to the discrete verbal representations. This structural mismatch of representations leads to suboptimal multimodal fusion, resulting in poor extraction of salient task representations and thus degraded task performance.

Moreover, as VQ-Fusion uses shared codebooks in the VQ information bottleneck to learn multimodal representations, this codebook sharing enables models to align the multiview representations and learn unified concepts. Learning unified concepts from multiple views is crucial, as multiple views capture the same interaction. Existing models are designed to learn visual and language representations from a single visual perspective. Thus, these models do not have any mechanisms to extract unified concepts from multiple visual views. VQ-Fusion enables these models to learn this unified concept using shared codebooks-based VQ.

Our experimental results also indicate that incorporating additional perspective-related information can help models to successfully ground objects. This is made apparent by the model performance on the perspective-aware object grounding (POG) task being consistently higher then the model performance on the object grounding (OG) task. This is particularly notable as the only difference between these tasks is the presence of the question's verbal perspective (Fig. 6.1). Thus, these results suggest models need to understand verbal perspective for successfully grounding objects in situations with multiple verbal perspectives.

Although all the VL models presented can achieve considerable performance for most of the EQA tasks, these models perform slightly better than random-guessing for the object counting (OC) task. As these models do not use object location-specific information, the models suffer at locating and counting objects given a spatial relation. One possible extension of these models to improve performance for the OC task is the incorporation of object locations in representation

G	PG	EQA Tasks							
		EP	OG	POG	OC	OAQ	OAC	PG	RG
✗	✗	51.03	26.65	52.79	09.94	24.01	51.22	48.95	56.75
✗	✓	53.87	60.66	71.08	11.51	64.69	60.63	66.31	**90.01**
✓	✗	53.51	63.49	70.90	**12.29**	69.43	**61.25**	66.67	87.23
✓	✓	**54.38**	**68.61**	**79.68**	11.86	**72.62**	60.74	**66.68**	89.59

Table 6.5: Impact of gaze (G) and pointing gestures (PG) in learning EQA tasks. The results suggest that incorporating gestures improves EQA task performance. G (✗) and PG (✗) indicate visual scenes that do not include humans.

learning. Our EQA-MX dataset contains rich annotations of object locations, which can easily be incorporated in developing new models.

6.5.4 Impact of Nonverbal Gestures

We evaluated the impact of nonverbal gestures on learning EQA tasks. We evaluated VQ-Fusion with CLIP models and 8 codebooks on the different splits of EQA-MX dataset: data samples with gaze and gestures, only gaze, only gestures, and without gaze and gestures (this data split contains visual scenes without human).

Results and Discussion: The results in Table 6.5 suggest that the model performs worse for EQA tasks if we train the model using data without nonverbal gestures. For example, the model trained using data without nonverbal gestures achieved only 26.65% accuracy for the object grounding (OG) task, whereas the model trained using data with gaze and pointing gestures achieved 68.61% accuracy for the OG task. This is a trend for all other tasks where the performance improved when gaze and/or pointing gestures were incorporated compared to when it only relied on the verbal message. The performance degradation indicates that the models must learn nonverbal gestures to answer questions with multimodal expressions for EQA tasks.

6.5.5 Impact of VQ Codebooks

We evaluated VQ-Fusion with the CLIP model for 8 EQA tasks by varying the number of codebooks in VQ: $\{2, 4, 8, 16\}$. We evaluated these models on our EQA-MX with varied nonverbal gestures (gaze and pointing gestures, only gaze, and only pointing gestures). We trained these models with multiple visual and verbal perspectives.

Results and Discussion: The results in Table 6.6 suggest that different codebooks help the model achieve the highest performance for different tasks. For example, VQ-Fusion with 8 codebooks can achieve the highest performance in existence prediction (EP), object grounding (OG), and object attribute compare (OAC) tasks, whereas VQ-Fusion with 2 codebooks can achieve the highest performance for perspective-aware object grounding (POG) and object counting (OC)

VQ CBs	EQA Tasks							
	EP	OG	POG	OC	OAQ	OAC	PG	RG
2	53.46	64.86	**82.70**	**13.14**	61.39	57.43	61.39	88.24
4	52.15	61.12	73.94	11.35	69.42	**70.59**	60.30	**89.93**
8	**54.72**	**65.49**	73.97	11.92	**70.85**	60.68	66.82	88.23
16	53.19	55.12	71.32	11.43	69.35	60.37	**66.99**	84.36

Table 6.6: Impact of the number of VQ codebooks (VQ CBs) in VQ-Fusion with the CLIP model in learning EQA tasks.

tasks. The number of codebooks depends on the task complexity of how many concepts need to be learned. As the OG task requires learning verbal perspective, the model requires more codebooks to learn perspective-related concepts. On the other hand, as perspective is already given in the POG task, VQ-Fusion requires fewer codebooks. Our results also show similar phenomena, where VQ-Fusion achieves 82.70% accuracy for the POG task with only 2 codebooks, whereas it achieves 65.49% accuracy for the OG task with 8 codebooks.

However, increasing codebooks more than optimal leads to decreasing task performance. For example, the object attributes compare (OAC) task accuracy degrades if we increase the number of codebooks by more than 4. As the OAC task involves whether two objects have the same attribute, the model can learn these simple concepts using fewer codebooks. Increasing the number may lead to sparsity in codebooks, i.e., many codes are left unutilized, limiting the models to extract salient representations. On the other hand, using a few codebooks for complex tasks, such as OG and OAQ, leads to tight bottlenecks, which limits the models to learning salient concepts. Hence task performance degrades. These results indicate that an optimal number of codebooks based on the task characteristics is required to achieve the highest performance for each task.

6.5.6 Impact of Multiple Visual Perspectives and Modalities

In real-world settings, robots are typically equipped with multiple camera views. Several studies have emphasized the significance of multiview data in accurately comprehending human actions and instructions[8], [11]. To further validate the importance of multimodal data (nonverbal gestures captured through visual views and verbal utterances) in understanding embodied question answering (EQA) tasks, we conducted extensive ablation studies with varying visual views (ego, exo and top) and verbal utterances (verbal utterance templates described in Table 6.2).

In the first setting, we used only verbal utterances for all eight EQA tasks (Table 6.7: Top). We used BERT [167] for learning the EQA tasks. The results suggest models using only a verbal modality can not effectively learn these EQA tasks. Conversely, if we utilized both verbal and nonverbal data, then the performance of these EQA tasks improved (Table 6.7). This degraded performance using only verbal data emphasizes the importance of utilizing both verbal and non-

Only Verbal	EP	OG	POG	OC	OAQ	OAC	PG	RG
	40.64	8.90	45.46	7.45	7.69	29.49	45.23	44.82

Train	Test	EP	OG	POG	OC	OAQ	OAC	PG	RG
Ego	Ego	53.86	59.92	70.98	10.60	68.56	61.86	64.41	87.54
Ego	Exo	52.61	17.28	62.45	8.96	15.06	56.62	63.39	82.33
Exo	Exo	53.67	39.46	69.96	11.24	56.76	60.20	66.39	88.58
Exo	Ego	52.84	21.39	69.70	10.78	25.03	58.68	64.49	88.20
All	All	54.72	65.49	82.70	13.14	74.32	70.59	66.99	89.93
All	Ego	54.32	60.63	82.31	12.22	69.84	60.89	66.71	89.03
All	Exo	54.17	59.14	78.02	12.55	61.71	62.25	66.53	89.26

Table 6.7: We trained CLIP models with VQ-Fusion using different combinations of modalities on the 8 tasks described in Figure 2 in the paper. Top Table: only verbal questions. Bottom Table: different visual modalities and verbal questions. The results suggest that multimodal models outperform those using only verbal data (Top Table). Additionally, training models with multiview data leads to robust performance, while using a subset of views results in performance degradation if the views change during testing (Bottom Table). Existence Prediction (EP), Object Grounding (OG), Perspective-Aware Object Grounding (POG), Object Counting (OC), Object Attribute Query (OAQ), Object Attribute Compare (OAC), Perspective Grounding (PG), Relation Grounding (RG).

verbal data modalities for appropriately learning EQA tasks. Additionally, it also indicates that our proposed EQA-MX dataset is less biased towards verbal data for comprehending EQA tasks.

In the second setting, we used verbal utterances and nonverbal gestures to learn EQA tasks. We varied the visual perspectives during training and testing through the use of different camera views (ego, exo, and top) to capture the nonverbal interactions. We used CLIP model to learn EQA tasks involving verbal utterances and visual views. The results suggest that models trained using multiple visual perspectives perform better than models trained using a single visual perspective (Table,6.7: Bottom). The reasoning behind this performance improvement is that models using multiple visual views can learn generalized multiview representations, which can improve the performance at inference time when visual views are varied.

6.5.7 Comparison of Single and Multitask Models

We evaluated the impact of learning multiple tasks in a visual-language model. We conducted this experimental analysis in two settings. In both settings, we used verbal utterances and multiple visual modalities to learn EQA tasks. In the first setting, we trained CLIP models for each EQA task separately. In the second setting, we trained CLIP models for a subset of EQA tasks. Finally, we used the extracted representation in each EQA task head, where these task heads are designed using an MLP.

ST	EP	OG	POG	OC	OAQ	OAC	PG	RG
	54.72	65.49	82.70	13.14	74.32	70.59	66.99	89.93

MT	EP	OG		EP	POG		EP	PG	
	53.25	40.76		52.68	73.90		52.62	49.86	

MT	EP	OAQ	OG		EP	PG	OAQ		PG	EQ	OAQ
	54.24	68.70	55.56		53.17	66.92	66.61		66.80	53.26	69.01

Table 6.8: We train CLIP models with VQ-Fusion in single task (ST) and multitask (MT) settings. We reported accuracy of these tasks. Tasks trained in an MT setting are grouped together. The results suggest that the performance of these models with multiple tasks degrades compared to models learning these tasks separately. Existence Prediction (EP), Object Grounding (OG), Perspective-Aware Object Grounding (POG), Object Counting (OC), Object Attribute Query (OAQ), Object Attribute Compare (OAC), Perspective Grounding (PG), Relation Grounding (RG).

The results in Table 6.8 suggest that the performance of models learning multiple tasks degrades compared to the models learning these tasks separately. As these tasks have different characteristics, learning these tasks together can compete in the representation learning space and degrades these tasks' performance. For example, training the CLIP model for the Existence Prediction (EP) and Object Grounding (OG) tasks together degrades the Object Grounding task performance to 40.76% compared to an accuracy of 65.49% for a separately trained CLIP model for OG task. Previous studies have observed similar performance degradation when learning multiple competing tasks. The primary reason behind the performance degradation is that the competing tasks have conflicting gradients among different tasks that introduce negative knowledge transfer and thus degrade these tasks' performance. Thus, an exciting future research direction would be to design novel multitask model architectures and training approaches where training on multiple tasks using multiple modalities improves the performance of every task in a shared model.

6.5.8 Generalizability of VQ-Fusion

To evaluate the generalizability of VQ-Fusion for another task involving multimodal representation learning, we incorporate VQ-Fusion in an existing multimodal learning model (HAMLET [7]) for human activity recognition tasks with multimodal sensor data (RGB videos, acceleration, gyroscope, and orientation). We have evaluated this modal on the MMAct dataset [8]. The MMAct dataset comprises 37 common daily life activities, each performed by 20 individuals and repeated five times. The dataset includes seven modalities, ranging from RGB data to acceleration and gyroscope measurements. Our experiments focused on utilizing two available viewpoints of RGB videos, as well as acceleration, gyroscope, and orientation data. Notably, the MMAct dataset also includes visually occluded data samples, providing an opportunity to evaluate the effectiveness of

Table 6.9: Cross-session performance comparison (F1-Score) of multimodal learning methods on MMAct dataset

Method	F1-Score (%)
SVM+HOG [110]	46.52
TSN (RGB) [111]	69.20
TSN (Optical-Flow) [111]	72.57
MMAD [8]	74.58
TSN (Fusion) [111]	77.09
MMAD (Fusion) [8]	78.82
Keyless [37]	81.11
HAMLET [7]	83.89
MuMu [11]	87.50
VQ-Fusion(HAMLET)	**87.69**

multimodal learning approaches in extracting complementary features for activity recognition.

In our experimental analyses, we adhered to the original session-based evaluation settings and reported the F1-score. The results indicated that the HAMLET model, which utilizes our proposed VQ-Fusion approach, outperformed all existing state-of-the-art multimodal human activity recognition (HAR) approaches in session-based evaluation settings on the MMAct dataset (Table 6.9). Specifically, the inclusion of VQ-Fusion enabled HAMLET to improve its F1-score by 4.2%, resulting in the highest reported F1-score of 87.69% (Table 6.9). These findings suggest that VQ-Fusion can effectively aid existing models in extracting salient multimodal representations, thereby enhancing the performance of downstream tasks in the field of HAR.

6.6 Broader Impact

Our dataset contains rich annotations of visual scenes, such as object locations, spatial relations, and multiple visual and verbal perspectives. These can be used to design new tasks to robustly comprehend embodied interactions. Moreover, our EQA-MX dataset can be used for diverse tasks in embodied settings, such as scene segmentation and conversational human-AI interactions with multimodal expressions. Additionally, our dataset can be used to develop and evaluate models that can be transferred to robots for comprehending embodied human instructions in real-world settings. Lastly, our experimental analysis provides valuable insights that can be used in designing robust VL models, such as using similar embedding structures for fusing continuous and discrete representations leading to performance improvements.

Limitations and Future Works As we developed separate models for different EQA tasks, there are several fascinating research avenues we can pursue in the future, such as developing robust multitask learning models, developing a mechanism to transfer multiple tasks from sim to real-world

settings, and developing a training mechanism to incorporate new tasks in the trained models. Due to the limited resource, we generated visual scene with image data. Thus, one possible extension can be generating video data and other modalities (depth, point cloud, and human 3D skeleton). We can use third-party Unity extensions to generate these modalities. As the visual-language models fail to perform better in the object-counting task, we can develop task-specific models to improve the performance.

In real-world settings, people's gestures and non-verbal interactions can vary greatly, which may not be fully captured in the simulated settings used to collect the EQA-MX dataset. This could limit the model's ability to interpret and respond to these cues in real-world applications accurately. Additionally, the complexity and texture of objects and scenes in the real world can also vary significantly, which may not be adequately represented in the synthetic dataset. This could affect the model's performance when applied to real-world scenarios. Although we have used a synthetic dataset with some limitations, we open several impactful research directions that can move the embodied human-AI interactions research field forward, such embodied-question answering with multimodal cues (verbal and nonverbal gestures) tasks and fusing continuous visual and discrete language representation can lead salient multimodal representations.

6.7 Limitations

In this work, we have designed novel embodied question-answering tasks, benchmarks, and novel visual-language models to comprehend these embodied question-answering tasks. This work has some limitations, which can be addressed to develop robust models for comprehending human-embodied interactions.

The primary limitation of this study is the use of a synthetic dataset, EQA-MX, for training and evaluating our model, VQ-Fusion. While this dataset has been carefully curated to include a variety of embodied QA data samples involving multimodal expressions from multiple visual and verbal perspectives, it may not fully capture the complexity and diversity of human interactions in real-world settings. Synthetic datasets, by their nature, are simulations and may lack the nuances, variability, and unpredictability inherent in human behavior.

The model's performance in real-world settings is tested on our Real-MADRID dataset which is presented in the next section. While VQ-Fusion shows promising results in our synthetic dataset, its effectiveness in real-world applications is yet to be determined. The transition from a controlled, synthetic environment to a dynamic, unpredictable real-world setting may pose challenges that were not encountered during the training phase.

Chapter 7
OBJECT GROUNDING USING MULTIMODAL EMBODIED INTERACTION CUES

Comprehending embodied interactions enables the AI assistants to ensure seamless with humans. Several tasks have been designed to develop models for comprehending embodied interactions, such as scene grounding thorough target object bounding box prediction, embodied path planning, referring expression grouding, and embodied question answering. Grounding objects through embodied interaction presents a challenge for extracting salient visual-language representations. This task necessitates the interpretation of verbal descriptions and nonverbal cues, such as gaze and pointing gestures and using this information to predict the bounding box of the object being referred to in a visual scene.

Current visual-language models predominantly utilize a cross-attention mechanism to merge visual and language representations. This method aligns the visual and language representations, enabling the model to identify key features from both modalities. However, this enforced alignment introduces a significant challenge. While the cross-attention mechanism streamlines the model architecture, it inherently aligns visual and language modalities to the task, potentially leading to less than optimal performance in tasks that demand a more salient balance between the modalities. For example, in tasks that require rich visual information, such as bounding box detection, the enforced alignment could result in the loss of crucial visual information. This loss leads to less than optimal multimodal fused representations and substandard task performance. Conversely, reinforcing the language representation can boost the model's performance in tasks where the language modality outweighs the visual information.

Thus, the challenge is to develop a model capable of identifying the key modalities and reinforcing the corresponding representation in the downstream task to enhance performance. This necessitates a model that can dynamically adjust the balance between visual and language modalities based on the specific requirements of the task at hand.

We have introduced a novel reinforced residual representation-based multimodal learning model designed to comprehend referring expressions in embodied interactions. This model is designed to extract and integrate multimodal representations from visual and language modalities, thereby enabling a comprehensive understanding of human verbal and nonverbal interactions. We aim to address the limitations of existing models, which often struggle to extract aligned and complementary representations for downstream task learning. We have developed a guided residual representation learning approach, which aids the model in extracting complementary representation to the aligned visual-language representations.

To evaluate the effectiveness of our model, we trained our proposed models and baseline on our CAESAR-PRO dataset. We conduct an extensive experimental analysis. Our experimental results, presented in Table 7.1, demonstrate the impact of reinforced representations in object bounding

box prediction. The inclusion of visual reinforced representation improves performance, from 46.60% to 51.60% for IOU-25, emphasizing the importance of visual cues in object grounding. Furthermore, when both visual and language representations are used as reinforced representations, the performance of the object grounding task is further enhanced. Our proposed model, ReReP, which incorporates both visual and language representation, outperforms the baseline model by a substantial 8.91% for IOU-25, underscoring the multimodal nature of the task and the need for a balanced approach to visual and language cues.

Moreover, our proposed model, ReReP, with reinforced representations, consistently surpasses the baseline across all IOU thresholds, validating its effectiveness and the importance of reinforced representations in object grounding tasks. The results also suggest that multimodal reinforced representations can complement the aligned representations extracted using the self-attention approach, improving task performance. Thus, our experimental analysis showcases the potential of reinforced representations in enhancing the performance of object grounding tasks and highlights the need for models that can dynamically adjust the balance between visual and language modalities based on task requirements. These findings provide a solid foundation for future research in this field of visual-language representation learning.

7.1 Problem Formulation

The task we are addressing involves grounding objects referred to by embodied interaction (verbal utterances and nonverbal gestures) in a visual scene. Embodied interaction combines verbal utterances and nonverbal cues such as gaze and pointing gestures. This complex task requires a model to effectively integrate and interpret both language inputs and visual scene data with nonverbal cues and scene descriptions.

Given a visual input V and a language input L, the model's task is to predict the bounding box of the object referred to by the embodied interaction. The visual input V can be an image or a video frame, and the language input L can be a spoken or written description of the object. In our work, we have used visual as an image from a video of an interaction, and language input is the text transcription of the verbal utterance in audio format. The task of bounding box prediction stands as a pivotal challenge within the realm of object detection, a cornerstone issue in the field of computer vision. This task, however, is further amplified in complexity due to the necessity of integrating language input. The model must recognize the object in the visual scene data and understand the language input and correctly associate it with the object in the visual input.

The bounding box is defined by coordinates (x_1, y_1, x_2, y_2), where (x_1, y_1) are the coordinates of the lower left corner of the box and (x_2, y_2) are the coordinates of the upper right corner of the box. The model's output is a prediction of these coordinates.

The challenge lies in designing a model that can effectively fuse the visual and language inputs to make accurate bounding box predictions. This requires the model to understand the language input in the context of the visual scene and vice versa. The model must also be robust to variations in the visual and language inputs, such as different camera view perspectives (egocentric and

exocentric), lighting conditions, object orientations, and language descriptions.

7.2 Reinforced Residual Representations for Robust Visual-Language Representation Learning

Existing visual-language models typically use a cross-attention mechanism to fuse visual and language representations. This approach aligns the visual and language representations, allowing the model to extract salient features from both modalities. However, this forced alignment also presents a significant challenge. The cross-attention mechanism inherently aligns visual and language modalities to the task. While using cross-attention simplifies the model architecture, this assumption can lead to suboptimal performance in tasks requiring a more nuanced balance between the modalities. For instance, in tasks that require more visual information, the forced alignment can result in the loss of valuable visual information, leading to suboptimal multimodal fused representations. The loss of visual representation can lead to suboptimal task performance where the visual information is more prominent than language modality in the downstream tasks, such as bounding box detection requires visual information and language as auxiliary information. In these cases, reinforcing visual information can help the model improve the downstream tasks' performance. Conversely, if language modality is more prominent than visual information in the downstream task, then reinforcing the language representation can help the model improve the task performance. Thus, we need a model which can identify the salient modalities and reinforce the corresponding representation in the downstream task to improve the performance.

Our proposed model, Reinforced Residual Representations ($ReReP$) is motivated by the need to overcome the above-mentioned challenges. The ReReP model is designed to allow for a more flexible fusion of visual and language representations, thereby preserving more information from both modalities and improving the performance on tasks that require a more nuanced balance between visual and language information. In the following sections, we will detail the architecture and mechanisms of the ReReP model, demonstrating how it addresses the aforementioned challenges and improves upon existing visual-language models.

7.2.1 Visual-Language Representation

The first step in our proposed model involves the extraction of visual and language representations from the given inputs. This is accomplished by passing the visual and language data through a visual-language model. The visual-language model is designed to process and understand both visual and language data. This model is trained to extract meaningful representations from visual and language data, which can then be used for various tasks. We can use a pretrained visual-language models, such as CLIP [92], DualEncoder [198], ViLT [91], VisualBERT [26], LXMERT [46], and VilBERT [27].

Given a visual scene V_{input} and a language input L_{input}, the visual-language model processes these inputs and outputs visual-language representations, denoted as V and L, respectively. We

can formulate this process in the following way:

$$V, L = \text{VL-Model}(V_{input}, L_{input}) \tag{7.1}$$

The visual representation V is a high-dimensional vector representation that captures the important visual features of the input, such as the shapes, colors, and spatial relationships of the objects in the image or video frame. Similarly, the language representation L is a high-dimensional vector representation that captures the semantic and syntactic features of the language input.

These representations are the result of a complex transformation process that involves multiple layers of cross-attention modules. The visual-language model learns to extract these representations during training by optimizing its parameters to minimize the difference between visual-language representations or using pretrained tasks, such as masked visual or language tasks. The extracted visual and language representations serve as the basis for the subsequent steps in our proposed model, which involve further processing and fusion of these representations to perform the task of object bounding box prediction.

7.2.2 Self-Attention based Multimodal Fusion

The next step in our proposed model involves the application of self-attention to fuse the extracted visual and language representations. The self-attention allows for a more flexible and context-aware fusion of visual and language representations. The self-attention works by assigning different levels of attention to different parts of the visual and language representations based on their relevance to the task at hand. This is accomplished by treating the visual and language representations as both the query and the key-value pairs in the self-attention computation.

Given the visual representation V and the language representation L, the self-attention mechanism computes attended visual and language representations, denoted as V^a and L^a, respectively:

$$V^a, L^a = \text{Self-Attention}(query = \{V; L\}, key = \{V; L\}, value = \{V; L\}) \tag{7.2}$$

The attended visual and language representations V^a and L^a are enhanced versions of the original representations, with more emphasis on the task's most relevant parts. This is achieved by weighting the original representations with attention scores, which are computed based on the similarity between the query and the key-value pairs.

The self-attention mechanism allows our model to focus on the most relevant parts of the visual and language inputs, thereby improving the quality of the fused representation. This is particularly important for tasks that require a nuanced balance between visual and language information.

7.2.3 Reinforcing Representation Using Guided Attention

We introduce guided attention to enhance our visual-language representation's robustness further. This module reinforces the attended visual and language representations by focusing on the most relevant parts of the visual or language representations.

Guided attention is similar to self-attention but has a crucial difference in input representations. In the guided attention, we use the original visual representation V as the query and the attended visual and language representations V^a and L^a as the key and value:

$$V^g, L^g = \text{Guided-Attention}(query = \{V^a; L^a\}, key = \{V^a; L^a\}, value = \{V^a; L^a\}) \quad (7.3)$$

This design allows the guided attention mechanism to focus on the parts of the visual input that are most relevant to the language input, thereby reinforcing the visual-language representation. This guided attention works as an information bottleneck to extract and reinforce task-specific representations.

The outputs of the guided attention, denoted as V^g and L^g, are then fused with the outputs of the self-attention mechanism to form the final representation. This fusion is performed by summation:

$$V^f, L^f = V^a + V^g, L^a + L^g \quad (7.4)$$

The fused representation V^f and L^f combines the strengths of the self-attention and guided attention mechanisms, providing a robust and flexible representation that can be used for the downstream task of object bounding box prediction or grounding. Using self-attention and guided attention mechanisms, our model can extract and emphasize the most relevant features from visual and language inputs, thereby improving the performance of tasks requiring a nuanced balance between visual and language information.

7.2.4 Training Model

We train our model for object bounding box prediction task. The model is trained by minimizing a loss function, which measures the difference between the predicted and ground truth bounding boxes. We have used data from multiple visual perspectives to train a robust model. Specifically, we use both egocentric and exocentric visual data. Egocentric visual data is captured from the first-person perspective, providing a view of the scene as seen by the person performing the task. Exocentric visual data, on the other hand, is captured from a third-person perspective. By using data from different visual perspectives, we can train a more robust model that can handle a wider range of scenarios and the model is not biased towards a single perspective.

The loss function we employ is the Mean Squared Error (MSE) loss, which is a standard choice for regression tasks such as bounding box prediction. The MSE loss is defined as the average of the squared differences between the predicted bounding box B_{pred} and the ground truth bounding box B_{gt}. The equation for the MSE loss can be written as:

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n}(B_{pred_i} - B_{gt_i})^2 \quad (7.5)$$

In this equation, n is the total number of elements in the bounding box (usually 4: lower-left-corner: (x_1, y_1) and upper-right-corner: (x_2, y_2)), and i indexes these elements. The MSE loss is suitable for minimization, with a lower MSE indicating a closer match between the predicted and ground truth bounding boxes. During training, the model learns to extract and fuse visual and language representations in a manner that is effective for the task of object bounding box prediction or grounding.

7.3 Experimental Setup

We developed all the models using the Pytorch (version: 1.12.1+cu113) [100] and Pytorch-Lightning (version: 1.7.1) [120] deep learning frameworks. Additionally, we used HuggingFace library (version: 4.21.1) for pre-trained models (BERT [1] [167], ViT [2] [166], and Dual Encoder [3]. For the Dual-Encoder model, we used an embedding size of 512.

The models were trained using the Adam optimizer with weight decay regularization set to 0 [95] and cosine annealing warm restarts with an initial learning rate of $3e^{-4}$, cycle length (T_0) of $2, 4, 6$, and cycle multiplier (T_{mult}) of 2. We utilized a batch size of 32 and trained the models for 14 epochs. To ensure reproducibility, we used the same fixed random seed (33) for all experiments. Finally, all models were trained on distributed GPU clusters, with each node equipped with 4 A100 GPUs.

7.4 Experimental Analysis

We conducted experimental analysis on the CAESAR-PRO dataset [71]. Through this detailed experimental analysis, we aimed to evaluate the effectiveness of our proposed model and its variations compared to the baseline model. We trained several variations of our proposed model by varying the reinforced representation of visual and language representations. We explored three distinct variations in the reinforced representation:

1. **Visual-Only Reinforced Representation**: In this variant, we solely used the visual representation as the reinforced representation. This approach emphasizes the importance of visual cues in the task of object bounding box prediction.

2. **Language-Only Reinforced Representation**: In this variant, we exclusively used the language representation as the reinforced representation. This approach underscores the significance of language cues in the task.

3. **Visual and Language Reinforced Representation**: In this variant, we used both visual and language representations as the reinforced representation. This approach recognizes the

[1] https://huggingface.co/docs/transformers/model_doc/bert
[2] https://huggingface.co/docs/transformers/model_doc/vit
[3] https://huggingface.co/docs/transformers/model_doc/vision-text-dual-encoder

Models	Reinforced Representations		Guided Reinforcement	Metrics		
	Visual	Language		IOU-25	IOU-50	IOU-75
Baseline	✗	✗	✗	46.60	13.84	1.26
ReReP	✓	✗	✗	51.60	20.40	2.30
ReReP	✓	✓	✗	55.51	21.70	2.70

Table 7.1: Comparisons of VL models performance for object grounding task of bounding box detection. The results suggest that reinforcing visual and language representation in VL models can improve the performance of object ground task. We evaluated several variations of ReReP by varying the reinforced representations.

multimodal nature of the task and attempts to balance the contributions of both visual and language cues.

In some variations, we employed guided attention to extracting the reinforced representations. Guided attention helps the model focus on the most relevant parts of the input, thereby improving the quality of the reinforced representations.

As a baseline, we trained the Dual-Encoder model on the CAESAR-PRO dataset. We used a pretrained Dual-Encoder model from the HuggingFace library[4] to extract the visual and language representations. These representations were then summed to produce the task representation for the bounding box detection task.

We trained all the models following the similar setup presented in Section 7.3. Following the prior work [24], we have reported IOU accuracy with various thresholds values $(25\%, 50\%, 75\%)$. The experimental results are presented in Table 7.1.

Results and Discussion: The experimental results, as presented in Table 7.1, show the effectiveness of reinforced representations for object bounding box prediction task. The results indicate that including visual reinforced representation enhances the task performance, improving from 46.60% to 51.60% for IOU-25. This enhancement underscores the importance of visual cues in object grounding and suggests that reemphasizing visual representation can lead to better performance.

Interestingly, the results also show that using both visual and language representations as reinforced representations can further boost the performance of the object grounding task. For instance, with both visual and language representation, our proposed model, ReReP, improves the baseline model performance by a substantial 8.91% for IOU-25. This finding highlights the multimodal nature of the task and suggests that a balanced consideration of both visual and language cues can lead to more accurate object grounding. Moreover, our proposed model ReReP with reinforced representations consistently outperforms the baseline across all IOU thresholds. This result validates our proposed model's effectiveness and emphasizes the importance of reinforced

[4] https://huggingface.co/docs/transformers/model_doc/vision-text-dual-encoder

representations in extracting salient features for object grounding tasks.

The experimental results also suggest that multimodal reinforced representations complement the aligned representations extracted using the self-attention approach. This finding indicates that while the self-attention mechanism effectively aligns visual and language modalities, the addition of reinforced representations complements the aligned representations, leading to improved task performance.

In conclusion, our experimental analysis demonstrates the potential of reinforced representations in improving the performance of object grounding tasks. It also highlights the need for models that can dynamically adjust the balance between visual and language modalities based on the specific requirements of the task at hand. These insights pave the way for future research in this exciting field of comprehending embodied interactions.

7.5 Findings

Our experimental analysis has led to several key findings that contribute to the field of visual-language representation learning, particularly in the context of comprehending embodied interactions:

1. **Significance of Reinforced Representations**: Our results highlight the importance of reinforced representations in the task of object bounding box prediction. This finding underscores the essential role of visual cues in object grounding and indicates that emphasizing visual representation can enhance performance.

2. **Advantages of Multimodal Reinforced Representations**: Utilizing both visual and language representations as reinforced representations further enhance the performance of the object grounding task. This finding emphasizes the multimodal nature of the task and suggests that a balanced approach to visual and language cues can lead to more precise object grounding.

3. **Efficacy of Proposed Model**: Our proposed model, ReReP, with reinforced representations, consistently surpasses the baseline across all IOU thresholds. This result validates our proposed model's efficacy and underscores the importance of reinforced representations in extracting salient features for object grounding tasks.

4. **Complementarity of Multimodal Reinforced Representations and Self-Attention**: The experimental results indicate that multimodal reinforced representations can complement the aligned representations extracted using the self-attention approach. This finding suggests that while the self-attention mechanism effectively aligns visual and language modalities, the addition of reinforced representations can provide a more nuanced balance between the modalities, leading to improved task performance.

These findings lay a solid groundwork for future research in this field and underscore the potential of reinforced representations in enhancing the performance of object-grounding tasks. They also highlight the need for models that can dynamically adjust the balance between visual and language modalities based on task requirements.

7.6 Limitations

While our research has yielded promising results, it is important to acknowledge its limitations. Our experimental analysis was conducted on the CAESAR-PRO dataset containing synthetic data. The use of synthetic data presents a potential limitation as it may not fully capture the complexity and variability of real-world human interactions.

Human gaze, pointing gestures, and verbal utterances can vary from synthetic to real-world environments. For instance, the synthetic data may not accurately represent the nuances of the human gaze and pointing gestures, which are influenced by many factors, including cultural norms, personal habits, and situational context. Similarly, verbal utterances in synthetic data may lack the diversity and complexity of real-world language use, including variations in accent, dialect, and speech patterns.

Training a model on synthetic data may therefore limit its performance on real-world datasets. While our proposed model, ReReP, demonstrated impressive performance on the synthetic CAESAR-PRO dataset, its effectiveness in real-world scenarios remains to be fully evaluated. While promising, the model's ability to dynamically adjust the balance between visual and language modalities based on task requirements may be challenged by the increased complexity and variability of real-world data.

Furthermore, while our model outperformed the baseline across all IOU thresholds in our experiments, it's important to note that the performance was evaluated based on a specific task (object bounding box prediction) and a specific metric (IOU). The model's performance may vary when evaluated on different tasks or metrics.

Thus, while our findings provide valuable insights into the potential of reinforced representations in visual-language representation learning, further research is needed to validate these findings in real-world settings and across different tasks and evaluation metrics.

Chapter 8
COMPREHENDING EMBODIED REFERRING EXPRESSIONS IN REAL-WORLD SETTINGS

Comprehending embodied interactions is essential for developing robust models that can function effectively in real-world settings. However, the diversity and complexity of real-world interactions often exceed the capabilities of existing synthetic datasets and simulators, such as our CAESAR simulator and the accompanying synthetic datasets (CAESAR-XL [25], CAESAR-L [25], and CAESAR-PRO [71]). While these have been instrumental in developing and diagnosing learning models, they may not fully capture the nuances of real-world human interactions.

In the literature, a few datasets have been developed to capture real-world embodied interactions, such as YouRefIt [24] and MoGaze [146]. However, these datasets have three crucial limitations that limit these datasets to develop robust models for understanding embodied interactions comprehensively. Some of these limitations are similar to the existing synthetic datasets described in Chapter 05. First, these datasets contain verbal utterances either from the speaker's or observer's perspective. For example, the verbal utterance "left ball" from the speaker's perspective can be interpreted as "right ball" from the observer's perspective. This perspective bias can create bias in the datasets; thus, the model developed using these datasets can comprehensively understand embodied interactions.

Second, embodied interactions in the existing datasets are captured using the exo or ego view. This single-view dependency creates view bias in the datasets, and the model trained on these datasets can not comprehend embodied interaction in diverse environments. For example, a referred object can not be viewed from one view due to the obstacle. However, in these settings, another view can observe an object occluded from one view. In the same way, the interactions (nonverbal gestures) occluded by one view can be captured by others. Thus, capturing the embodied interaction and scene is crucial using multiple views (ego, exo, and top).

Third, existing datasets partially capture nonverbal gestures. These datasets either capture pointing gestures or gaze. However, in embodied interactions, both signals provide complementary information to comprehend an interaction robustly. Fourth, existing datasets are collected indoors, mostly lab and at home. This drawback limits to training models to comprehend interactions in diverse settings, such as indoor (store, office, and home) and outdoor (e.g., store). Additionally, these datasets are collected from a stationary camera from a fixed angle. As a result, these datasets are biased to particular system settings. Thus, models trained with existing datasets can not be utilized to develop perception systems of autonomous systems, specifically mobile systems, to comprehend human interaction in diverse real-world settings.

To address these issues, we curated a diverse dataset, REMO, to comprehend human interaction in real-world settings. We collected the dataset in diverse indoor and outdoor settings with

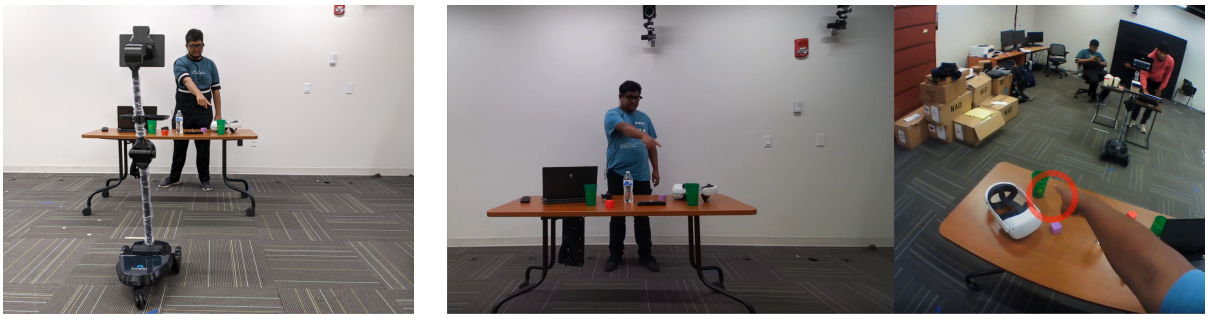


Figure 8.1: Real embodied data collection system and sample data. Left: Data collection system to collect embodied interaction in real-world settings. We have collected data using Azure Kinect DK mounted on the Ohmni robot and ego camera view and eye gaze using pupil smart glass. Right: A sample data collected using our data collection system is depicted.

varying environment attributes, such as lighting conditions, object arrangement, and environment appearance. We have used an Azure Kinect DK [199] device to capture the embodied interactions from different angles. This device was installed on the Ohmni Robotic system [200] to capture the ego view from the robot. Moreover, we have collected gaze and ego views from a human perspective using PupilCore Smart Glass (Invisible) [201]. Our data collection system has been depicted in Fig. 8.1. Finally, we have annotated these interactions using expert human annotators and curated verbal referring expressions. We have collected and annotated data under an approved IRB (protocol number: 4627, Title: Understanding Multimodal Human Instruction in Embodied Environment).

By providing a robust and diverse dataset that captures embodied interactions in real-world settings, we are enabling the development of more comprehensive and effective AI models. The dataset's diversity, in terms of environmental attributes and nonverbal signals, offers a rich resource for training and testing AI systems. This will allow researchers and developers to create models that can better understand and interact with the world, thereby enhancing the performance of AI systems in a wide range of applications, from autonomous vehicles to assistive technologies. Additionally, our proposed model's ability to extract and integrate multimodal representations offers a more comprehensive understanding of human interactions. This can help the AI systems seamlessly interact with humans, leading to more natural and effective communication. Furthermore, the insights gained from our experimental analysis could guide future research in the field, contributing to the development of more robust and effective models for understanding embodied interactions in real-world settings.

8.1 Data Collection

8.1.1 Data Collection System

We have developed a data collection system to synchronously collect data from different sensors. This system incorporates the Azure Kinect, which provides a multitude of sensory data, including visual, depth, infrared (IR), skeletal tracking using Azure SDK, and inertial measurement unit (IMU) data. In addition, we have integrated the Pupil Smart Glasses, which offer visual, IR, gaze tracking, and gesture recognition capabilities. To facilitate data collection in real-world scenarios, we have mounted these devices onto the Ohmi Lab's telepresence robot. This setup enhances the practicality of our system and encourages natural interactions with subjects, particularly when providing object-referential instructions to the robot. A visual representation of our integrated data collection system can be found in Figure 8.2.

The Azure Kinect DK sensor, a key component of our system, boasts the following specifications:

- RGB Camera (Heighest Resolution: 3840 x 2160 px @30 fps)
- Depth Camera (Method: Time-of-Flight, Highest Resolution: 640 x 576 px @30 fps)
- Motion Sensor (An LSM6DSMUS as an inertial measurement unit (IMU) with an accelerometer and a gyroscope with a sampling rate of 1.6 Hz.)
- Microphone (USB audio class 2.0, Channel: 7, Sensitivity: -22 dBFS (94 dB SPL, 1 kHz), Signal to noise ratio > 65 dB, Acoustic overload point: 116 dB)

The participant in our study was equipped with the Pupil Lab's Pupil Invisible Eye Tracker. This device is accompanied by an Android smartphone responsible for recording the participant's eye-tracking data. The data is subsequently transmitted to the Pupil Cloud via the Pupil Invisible Android application. This seamless hardware and software integration ensures efficient and reliable data collection and transmission. The Pupil Invisible Eye Tracker is a state-of-the-art device with a range of features designed to capture precise and accurate eye-tracking data. This device, in conjunction with the Azure Kinect and other sensors, forms a comprehensive system for embodied data collection, enabling us to gather a rich dataset for our research. The specifications of the Pupil eye tracker are as follows:

- Eye Cameras (200Hz @ 192x192 px IR illumination)
- Scene Camera (Detachable scene camera, 30Hz @1088 x 1080px, 82x82 FOV)

The participant stands before the Ohmin robot and provides instructions that may involve both verbal and non-verbal object-referencing gestures (Figure 8.2 (a)). The Azure Kinect DK's RGB

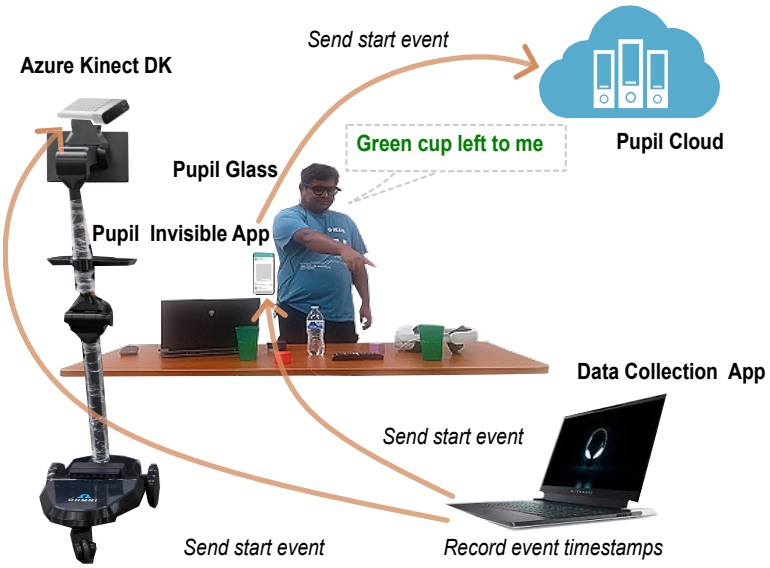


Figure 8.2: Real embodied data collection system

camera captures a continuous stream of RGB data that provides an external or exo-centric perspective of the participant. Moreover, the depth camera of Azure Kinect records depth and infrared data streams, further enriching the exo view of the participant. The system also allows to collect IR data using Azure Kinect's IR sensor. Additionally, we used the Azure Kinect Body Tracking SDK to capture the skeletal data streams. This SDK allows us to track 32 body joints, providing 3D coordinates and orientation information for each joint. The Kinect's microphone also records audio data from the user.

The Pupil Invisible Eye Tracker complements the data collected by the Azure Kinect by recording an RGB data stream that provides an ego-centric perspective of the participant. This combination of exo-centric and ego-centric views provides a comprehensive understanding of the participant's interactions. An Alienware m15 R4 laptop powers the entire system with an i7-10870H RTX processor. This high-performance machine serves as the backbone of our system, integrating all sensor components and ensuring their smooth operation.

We developed a Python-based application to facilitate coordination and synchronization among all system components. This application is central to the operation of our data collection system, ensuring that all components work together seamlessly to collect synchronized data from multiple sensors. We used this system to capture video from each camera, the time series data from the IMU and skeleton joints, and session metadata. We employed the pyKinectAzure library to interface with the Azure Kinect SDK sensor, while the Pupil Labs' Realtime API was used to communicate with the Pupil Eye camera.

One of the significant challenges we encountered was the synchronization of various data streams captured by different devices. To overcome this, we implemented a time-based synchro-

nization method. This method records the UNIX timestamps of different data capture events and data streams, enabling synchronization during post-processing. This synchronization is crucial to align the data streams captured from different devices. Our approach involved recording the timestamp at both the start and end of each interaction and the timestamp of the event when the participant pointed to an object. This was achieved using our Python-based system, operated by the individual recording the data collection sessions. We utilized different keystrokes on a standard keyboard to denote different events. The "Space" key was pressed at the start and end of an interaction, while the "G" key was pressed to indetify the canonical moment of an interaction. The canonical moment indicates when the participant point to an object using gaze or pointing gestures. Moreover, The "G" keystroke event time was used to identify the canonical frame, i.e., the frame where the participant actually pointed to an object. When the participant used cues other than pointing, such as gaze, the "G" key was pressed when the gaze event occurred. The "Space" keystroke event time was used to identify the start and end of an interaction, thereby facilitating the segmentation of interactions. The "Q" key was used to terminate a session. The corresponding UNIX timestamp for these keystroke events was recorded for both the Azure Kinect and Pupil Lab Eye tracker.

We stored the Azure Kinect recordings and the corresponding keystroke event time locally as MP4 and JSON files, respectively. For the Pupil eye tracker, the recordings of the participants' ego view and keystroke events were saved in the Pupil Cloud using the Pupil Lab Android app and Pupil API, respectively.

It's important to note that while our current system utilizes a time-based synchronization method to synchronize between two different devices (the Azure Kinect Sensor and Pupil Eye Tracker), it is designed to be extensible. For example, our system can be expanded to incorporate multiple Azure Kinect devices to capture multiple views of the participant during interaction rather than just the ego and exo views.

We use three different environments for collecting data: home indoor, home outdoor, and a laboratory environment. An indoor home environment includes living rooms, bedrooms, kitchens, etc.; an outdoor one includes balconies, parking lots, front yards, etc. The collaborative robotics lab (CRL) of the University of Virginia was chosen as the laboratory environment. While choosing objects, we prioritize those usually available in these environments. A complete list of the objects used in the dataset can be found in Appendix 1.

8.1.2 Data Collection Protocol and Procedure

The data collection process began with a comprehensive introduction to the subjects about the data collection system, the purpose of the dataset, and the protocol to be followed during data collection. Prior to participating in the data collection sessions, subjects were required to complete a demographic survey.

Each session involved subjects providing instructions that referenced objects in their surroundings, using both language and nonverbal gestures (gaze and pointing gestures). These instructions

were designed to facilitate natural interaction between the subject and a robot capable of interpreting verbal and nonverbal human instructions. We have two instructions for collecting data in constrained and unconstrained settings. In the constrained setting, subjects were briefed on the format of instructions and how they could employ various modalities (verbal and nonverbal) to make the interaction as natural as possible. We also suggested the participants use both verbal and nonverbal gestures to describe an object. In the unconstrained setting, we did not suggest whether to use verbal or nonverbal gestures to describe an object. We instruct the participant to describe an object to the robot.

The ultimate goal of this dataset is to enhance the ability of social robots to accurately interpret object referencing instructions. This involves identifying the object uniquely, which requires extracting the object's location and other attributes from the instruction. This task presents a challenge as humans often use diverse formats when providing verbal instructions, and these instructions may sometimes lack the necessary features for object identification. Incorporating nonverbal cues, such as pointing or referencing the object in relation to another object, can significantly improve the efficiency of interpreting object referencing instructions. Furthermore, object referencing instructions can be given from multiple perspectives, such as the subject's or the robot's perspective, which must be resolved for accurate object comprehension.

The participants were given the flexibility to choose any perspective (subject, robot, or neutral) when providing instructions. This approach allowed us to diversify our dataset by including object-referencing instructions with varied spatial referencing and perspectives. For instance, an object could be referenced in relation to another object, such as "The black box on top of the brown table." The object reference in the verbal instruction could be from the subject's perspective, e.g., "The couch to my right," or it could be from the robot's perspective, e.g., "The lamp to your left." To further diversify the dataset, some participants were not given any specific guidance on the interaction mode, allowing us to capture natural human instincts when providing instructions. This approach also helped eliminate biases that might be introduced by pre-guidance on the format of the instructions, allowing subjects to be flexible in their instruction delivery.

Each subject participated in multiple sessions, each lasting approximately one hour. During each session, the subject performed several interactions. Using our data collection system, we recorded the subject's ego view, exo view, IMU, skeleton, and audio data stream for each session. Upon completion of the sessions, subjects were asked to complete a post-task survey and sign a consent form to give permission to release the dataset. We compensated each participant with $\$15$ dollars for one hour of their time.

8.1.3 Data Post-processing

Each data collection session resulted in generating an Azure Kinect video file in MP4 format, generated by using our Python-based data collection application. We used the recorded timestamps to split each session's video data. This MP4 file encapsulates three data streams from the Azure Kinect's camera sensor: RGB, Depth, and Infrared. Accompanying JSON files contain the time

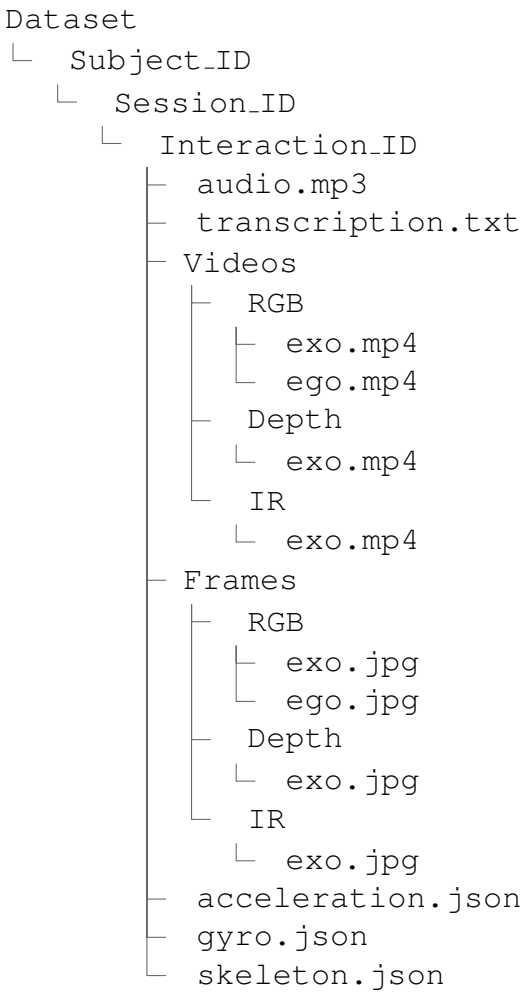


Figure 8.3: Dataset folder structure

series data for the IMU and skeleton joints, along with pertinent session metadata. We employed the FFmpeg library to separate the Kinect video streams into individual MP4 files and to extract the recording audio as an MP3 file. The IMU time series was divided into two distinct files for the accelerometer and gyroscope readings. Concurrently, the Pupil eye tracker generated an MP4 video file for each session, which was saved to the Pupil Cloud. We used the recorded timestamps from the Pupil Cloud to split each session video. We also compared the Azure Kinect and Pupil Cloud timestamps to determine the time lag and synchronize the video streams.

The primary challenge in post-processing the data was segmenting the interactions and synchronizing the data from the Azure Kinect and Pupil Lab. To segment each interaction from the Azure Kinect data streams, we identified the start and end times of that interaction. We also pin-

pointed the canonical frames, i.e., frames where the subject precisely points to an object. We used the FFmpeg library to split each interaction and canonical frame. Subsequently, we located the corresponding Pupil recording for the Azure Kinect recording in the Pupil Cloud using the Python Pupil Cloud API. We used the recording-start timestamp saved in the metadata file to find the matching Pupil recording. After downloading the Pupil video, we applied the same procedure as with the Azure Kinect recording to split the interactions and canonical frames at the timestamps recorded during data collection. Finally, we employed the OpenAI Whisper library to transcribe the Kinect audio data into corresponding text. It's important to note that we manually verified the synchronization and segmentation with the help of five human experts.

Our dataset comprises numerous data collection sessions and the results of data post-processing. Each session is organized in a specific folder structure which is depicted in FIgure 8.3. Here, 'transcription.txt' is the text transcription of 'audio.mp3'. In the 'Videos' subfolder, 'exo.mp4' and 'ego.mp4' refer to the videos from the Azure Kinect SDK camera and Pupil Eye Camera, respectively. Similarly, In the 'Frames' subfolder, 'exo.jpg' and 'ego.jpg' refer to the canonical frame from the Azure Kinect SDK camera and Pupil Eye Camera, respectively. We stored the IMU sensor data in the acceleration.json and gyron.json files. The skeleton data of the whole interaction is stored in skeleton.json file.

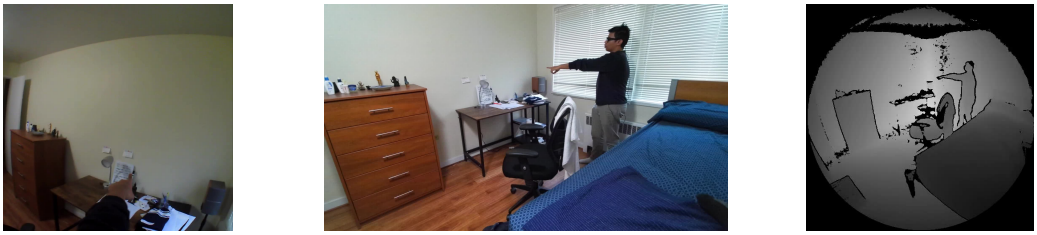
8.1.4 Participants

We recruited a total of 66 participants for our study from Charlottesville, Virginia, United States. The participant group was balanced in terms of gender, with 53.03% males (n = 35) and 46.97% females (n = 31). The majority of our participants were students from various academic levels and disciplines at the University of Virginia. To ensure the diverse verbal utterance data, the participants were not required to be native English speakers. The average age of the participants was 26.66 years, with a standard deviation of 3.36 years. Participants were asked to rate their level of experience with robots on a Likert scale ranging from "no experience" (1) to "expert-level experience" (5), resulting in a mean score of 2.25 and a standard deviation of 0.96. The majority of participants, 93.94% (n = 62), were right-handed, while 6.06% (n = 4) were left-handed. One participant did not consent to the publication of the data. Therefore, we have restricted the use of their data and excluded it from our dataset.

8.1.5 Dataset Statistics

We collected data in both constraiedn and unconstrained setting. Some sample data from our dataset has been shown in Figure. A detailed statistical breakdown of the dataset is presented in Table 8.1. The data collection phase involved 392 sessions split between home and lab environments. A total of 13,990 interactions were recorded, with 3,176 occurring at home and 10,814 in the lab. The total video time recorded was 17.62 hours, with 4.14 hours coming from at-home sessions and 13.48 hours from lab sessions. A total of 14,368 frames were captured. Each interaction

Verbal utterance: Cycle in front of me



Verbal utterance: The whiteboard to your left

(a) Sample data from constrained setting.

Verbal utterance: A container with three pingpong balls in it

Verbal utterance: On the top shelf is a three-drawer container with some papers on it

(b) Sample data from unconstrained setting.

Figure 8.4: Sample data from REMO dataset in both constrained and unconstrained settings.

lasted 4.53 seconds on average, and there were approximately 36.65 frames in each session. The average session length was 2.69 minutes. These statistics provide valuable insights into the scale and nature of the data collected and will serve as a solid foundation for subsequent analysis and interpretation.

Attribute	Value
Number of Sessions	392 (Home: 194, Lab: 198)
Number of Interactions	13990 (Home: 3176, Lab: 10814)
Total Number of Frames	14368
Total Video Time (hrs)	17.62 (Home: 4.14, Lab: 13.48)
Avg. Interaction Length (sec)	4.53
Avg. frames per session	36.65
Avg. Session Length (min)	2.697

Table 8.1: Dataset Statistics

8.1.6 Post Task Survey Analysis

We conducted a post-task survey to gain a deeper understanding of participants' preferences when instructing a robot. In this survey, participants were asked to indicate their preferred method of object referencing. The options provided were: using only verbal instructions, only gestures, or a combination of verbal and gestures.

The survey results revealed that a significant majority of participants, 96.97% (n = 63), preferred using both verbal and nonverbal gestures (gaze and pointing gesture). A small fraction of participants, 3.03% (n = 2), preferred using only verbal instructions. One participant did not permit to release the data, so we did not include that participant's response in this analysis. Additionally, none of the participants chose nonverbal gestures as their sole preferred method of communication.

These findings underscore the perception among humans that a combination of verbal and nonverbal forms of instruction is the most efficient way to convey object-referencing expressions. This aligns with our motivation to develop a dataset for object-referencing instructions that incorporates both language and visual cues. The results of this survey provide valuable insights into human communication preferences, which can inform the design and development of more intuitive and effective human-robot interaction systems.

8.2 Experimental Setup

We developed all the models using the Pytorch (version: 1.12.1+cu113) [100] and Pytorch-Lightning (version: 1.7.1) [120] deep learning frameworks. We also used HuggingFace library (version: 4.21.1) for pre-trained models (BERT [1] [167], ViT [2] [166], Dual Encoder [3], and CLIP [4] [92]). For the Dual-Encoder and CLIP models, we used an embedding size of 512. We train models using

[1] https://huggingface.co/docs/transformers/model_doc/bert
[2] https://huggingface.co/docs/transformers/model_doc/vit
[3] https://huggingface.co/docs/transformers/model_doc/vision-text-dual-encoder
[4] https://huggingface.co/docs/transformers/model_doc/clip

the Adam optimizer with a weight decay regularization set to 0 [95] and cosine annealing warm restarts at an initial learning rate: $3e^{-4}$, cycle length (T_0): $\{2, 4, 6\}$, and cycle multiplier (T_{mult}): 2. We used batch size 32 and trained models for 14 epochs. We used the same fixed random seed (33) for all the experiments to ensure reproducibility. Lastly, all models are trained in distributed GPU clusters, where each node contains 4 A100 GPUs.

8.3 Experimental Analysis

We performed our experimental analysis on our REMO dataset. This comprehensive analysis aimed to assess the performance of our proposed model and its variations compared to the baseline model. We trained multiple variations of our proposed model, each differing in the type of reinforced representation of visual and language modalities. We examined three distinct variations:

1. **Visual-Only Reinforced Representation**: This variant solely leverages the visual representation as the reinforced representation. This method emphasizes the crucial role of visual cues in the task of object bounding box prediction.

2. **Language-Only Reinforced Representation**: This variant solely utilizes the language representation as the reinforced representation. This method highlights the importance of language cues in the task.

3. **Visual and Language Reinforced Representation**: This variant employs both visual and language representations as the reinforced representation. This method acknowledges the multimodal nature of the task and strives to balance the contributions of both visual and language cues.

In some variations, we incorporated guided attention to extracting the reinforced representations. Guided attention enables the model to concentrate on the most pertinent parts of the input, thereby enhancing the quality of the reinforced representations.

For the baseline, we trained the Dual-Encoder model on the REMO dataset. We utilized a pretrained Dual-Encoder model from the HuggingFace library[5] to extract the visual and language representations. These representations were then summed to generate the task representation for the bounding box detection task.

All models were trained following the similar setup outlined in Section 7.3. We reported IOU accuracy with various threshold values $(25\%, 50\%, 75\%)$. The experimental results are presented in Table 8.2.

Results and Discussion: The experimental results, as presented in Table 8.2, provide valuable insights into the effectiveness of reinforced representations for the task of object bounding box prediction. The results indicate that including visual reinforced representation significantly enhances

[5] https://huggingface.co/docs/transformers/model_doc/vision-text-dual-encoder

Models	Reinforced Representations		Guided Reinforcement	Metrics		
	Visual	Language		IOU-25	IOU-50	IOU-75
Baseline	✗	✗	✗	70.05	41.03	10.40
ReReP	✓	✗	✗	70.30	42.50	11.70
ReReP	✓	✓	✗	70.30	43.60	12.30
ReReP	✓	✗	✓	72.01	45.79	12.80
ReReP	✓	✓	✓	73.49	51.63	21.21

Table 8.2: Comparisons of VL models performance for object grounding task of bounding box detection. The results suggest that reinforcing visual and language representation in VL models can improve the performance of object ground task. We evaluated several variations of ReReP by varying the reinforced representations.

task performance, with an improvement from 70.05% to 72.05% for IOU-25. This enhancement underscores the importance of visual cues in object grounding and suggests that reemphasizing visual representation can lead to better performance.

The results further indicate that incorporating both visual and language representations as reinforced representations can significantly enhance the performance of the object grounding task. For example, our proposed model, ReReP, which utilizes both visual and language representation, boosts the baseline model performance by a notable 3.45% for IOU-25. This outcome underscores the multimodal aspect of the task and implies that a balanced approach to both visual and language cues can result in more precise object grounding. Furthermore, equipped with reinforced representations, our proposed model ReReP consistently surpasses the baseline across all IOU thresholds. This experimental result validates our proposed model's effectiveness and highlights the significance of reinforced representations in extracting pertinent features for object grounding tasks.

While our proposed model, ReReP, shows modest improvements in task performance at lower IOU thresholds (25% and 50%), it demonstrates a significant performance boost at higher IOU thresholds. Specifically, ReReP achieves a 10.81% improvement in the bounding box prediction task at higher IOU thresholds. This is particularly noteworthy as achieving high performance at higher IOU thresholds is challenging. IOU, or Intersection over Union, is a metric used to evaluate the accuracy of an object detector on a specific dataset. Higher IOU thresholds are more challenging because they require the predicted bounding box to match the ground truth bounding box more closely. In other words, a model must correctly identify the objects with high precision to achieve a high score at these thresholds. The significant improvement achieved by ReReP at these thresholds suggests that our model is particularly effective at accurately identifying objects, even under stringent evaluation criteria.

However, our analysis also reveals the crucial role of guided reinforcement in our model's performance. Without the use of guided reinforcement, our model's performance drops by 8.91%

for IOU-75. This drop in performance underscores the importance of guided reinforcement in our proposed approach. Guided reinforcement helps the model focus on the most relevant parts of the input, thereby improving the quality of the reinforced representations. It essentially guides the learning process, helping the model to better leverage the reinforced representations for the task at hand. Our detailed analysis highlights the effectiveness of our proposed model, ReReP, particularly at higher IOU thresholds. It also emphasizes the importance of guided reinforcement in enhancing the model's performance. These findings provide valuable insights for developing effective visual-language representation learning models.

The experimental results also indicate that multimodal reinforced representations can effectively supplement the aligned representations derived using the self-attention approach. This observation suggests that while the self-attention mechanism is adept at aligning visual and language modalities, incorporating reinforced representations can offer a more refined balance between these modalities, leading to enhanced task performance.

Therefore, our experimental analysis showcases the potential of reinforced representations in augmenting the performance of object grounding tasks. It also underscores the necessity for models capable of dynamically adjusting the balance between visual and language modalities, contingent on the specific requirements of the task at hand.

8.4 Limitations

While our work has made significant strides in understanding embodied interactions in real-world settings, it is not without its limitations. Firstly, our study relies heavily on the REMO dataset, which, while diverse and robust, may not capture all the nuances of human interactions in every conceivable real-world setting. The dataset was collected in a variety of indoor and outdoor environments with varying attributes, such as lighting conditions, object arrangement, and environment appearance. However, there are other environmental factors or settings not covered in our dataset that could influence embodied interactions. For example, our dataset does not contain data from very dark environments where the depth RGB data may not capture the embodied interactions appropriately.

Secondly, our proposed model's ability to extract and integrate multimodal representations, while effective, may not perfectly capture the complexity of human interactions. Human interactions are incredibly nuanced and can vary greatly between individuals and contexts. Thirdly, while our model has shown promise in enhancing the performance of AI systems in a wide range of applications, from autonomous vehicles to assistive technologies, its real-world applicability and scalability have yet to be fully tested. Further research and testing are needed to determine how well our model can be integrated into different systems and how it performs in large-scale, real-world applications.

Lastly, our work is limited by the inherent challenges of working with embodied interactions, such as the difficulty of accurately capturing nonverbal gestures and the potential for perspective

bias in verbal utterances. Despite our best efforts to mitigate these issues, they remain inherent challenges in this field of research.

Despite these limitations, we believe our work represents a significant step forward in the field of artificial intelligence, particularly in understanding embodied interactions in real-world settings. We hope that our findings will inspire further research in this area, and we look forward to seeing how our work can be built upon and improved in the future.

8.5 Broader Impact

The broader impact of our work lies in its potential to significantly enhance the understanding of embodied interactions in real-world settings. By providing a robust and diverse dataset, REMO, that captures embodied interactions in various indoor and outdoor settings, we are enabling the development of more comprehensive and effective AI models. These models can be trained and tested using our dataset, which offers a rich resource of environmental attributes and nonverbal signals. This will allow researchers and developers to create models that can better understand and interact with the world, thereby enhancing the performance of AI systems in a wide range of applications.

For instance, in the field of autonomous vehicles, our work can contribute to the development of more sophisticated perception systems that can comprehend human interaction in diverse real-world settings. This can lead to safer and more efficient autonomous vehicles that can better navigate complex environments and interact with humans. Similarly, in the field of assistive technologies, our work can help develop more intuitive and responsive systems that can understand and respond to human interactions effectively. This can significantly improve the quality of life for individuals who rely on these technologies.

Furthermore, our proposed model's ability to extract and integrate multimodal representations offers a more comprehensive understanding of human interactions. This can help AI systems seamlessly interact with humans, leading to more natural and effective communication. This is particularly important in the development of AI assistants, which need to understand and respond to human interactions in a natural and intuitive manner.

Finally, the insights gained from our experimental analysis could guide future research in the field, contributing to the development of more robust and effective models for understanding embodied interactions in real-world settings. This can lead to significant advancements in the field of artificial intelligence, pushing the boundaries of what AI systems can understand and achieve.

Chapter 9
RELATED WORK

9.1 Multimodal Representation Learning

Multimodal Learning: Prior neuroscience studies suggest that multisensory animal cognition systems align unimodal features before their fusion [1], [116]. For example, a study on adult cats conducted by Meredith and Stein [1] found that multisensory stimuli (visual, auditory and somatosensory) converge in the deep laminae superior colliculus cell. In another study, Meredith et al. [116] found that multisensory interactions and convergence occur in deep laminae cells. Several studies also found that animals' multisensory systems inherently employ recurrent [115], [202], [203] and hierarchical [203], [204] information processing approaches to combine multimodal information. Similarly, in the analysis of the human visual system, Lamme et al. [114] found that feed-forward information processing is responsible for the unconscious vision and recurrent processing is responsible for attentive vision and visual awareness.

Several multimodal feature fusion approaches have been proposed in the multimodal learning literature, such as early, intermediate, and late fusion [35], [38]. Early fusion approaches combine raw sensor data and extract feature representation from this combined raw multimodal sensor data [7], [9]. Late fusion approaches extract unimodal feature representations and determine the task output by utilizing the unimodal representation separately. These approaches fuse the prediction from each modality independently [38]. However, intermediate fusion allows cross-modal interaction to fuse mid-level features for extracting complementary multimodal representation. Recent works in multimodal machine learning predominantly employ intermediate fusion over early and late fusion, as it allows to fuse mid-level features to obtain complementary multimodal representation [12], [28], [38], [42].

Intermediate fusion approaches create lateral connections among unimodal feature encoders to fuse mid-level features [12], [16], [40], [205]. For example, Feichtenhofer et al. [16] utilized directed lateral connections from audio to visual modality to fuse intermediate features. Moreover, Joze et al. [12] proposed squeeze and excitation operations to recalibrate and fuse the mid-level feature fusion. However, the interactions among heterogeneous modalities are chosen based on human intuition, which may introduce biases and limit the performance of feature encoders. Another approach is to employ attention mechanism to combine different unimodal features [7], [37]. For instance, Long et al. [37] proposed an attention approach to obtain unimodal features, which are concatenated for multimodal representation. While the aforementioned approaches have made significant advances, they all fuse features in a feed-forward manner, limiting the modality-specific encoders from aligning and refining their representations when observing features from other modalities.

Representations Alignment and Refinement: In recent years, representations alignment

and refinement approaches have been studied in the literature for various tasks, such as audio-video localization [198], [206], [207], answer prediction [208], text-video retrieval [209], and activity recognition [11]. For example, Wang et al. proposed a model for local feature alignment and temporally aggregated global feature alignment between text and video modalities [209]. The cross-modal similarities are used to train the model for the text-video retrieval task. Similarly, Wu et al. developed a Dual Attention Matching (DAM) module to model high-level event information and a global cross-check mechanism to extract local temporal information for audio-visual event localization tasks [198]. Apart from the feature alignment, several approaches have been proposed to use the fused multimodal representations to refine the task-specific representation. For example, in our prior work, we proposed a guided multimodal fusion approach that extracts activity-group specific representation to prioritize and fuse multimodal representations for activity recognition [11]. Similarly, Hu et al. proposed a multimodal transformer model to fuse the multimodal representations, which is then used for iterative answer decoding [208]. In the decoding step, the fused representations are used to refine the representation for answer decoding iteratively. Moreover, Ramaswamy and Das developed Audio-Visual Fusion Block (AVFB) to fuse audio-video features used by Segment-Wise Attention Block (SWAB) to match temporal feature segments for sound source localization [206]. Additionally, Wu and Yang developed a label refinement-based audio-visual joint representations learning model where modality-specific refined labels are used as a weakly supervised signal for cross-modal audio-visual contrastive learning [207]. One of the common properties of these learning architectures is that these architectures align or refine representations after extracting unimodal features.

Although several approaches have addressed the representation alignment and refinement related challenges for various tasks [11], [198], [206]–[209], there are some crucial issues we need to address to develop a robust multimodal learning model. First, these approaches align and/or refine multimodal representations after extracting unimodal representations and do not allow the unimodal learning models to observe the extracted representations from other modalities. Although some of these approaches match the local and global representations, these approaches do not allow the unimodal learning models to use the global representations to refine the local representations. Second, most of these learning architectures are constrained to two modalities, such as text and visual, or audio and visual modalities. Additionally, most of these works used a pair-wise feature representation matching approach to align multimodal representations. Thus, we need to develop and evaluate learning models which can align and refine representations from multiple modalities.

9.2 Perceiving Human Behavior

Multimodal Human-Activity Recognition: Activity recognition using a single data modality has been extensively studied with work using visual data [16], [134], [210], skeleton [31], [32] and physical sensor data [30]. However, relying on only one modality may lead to poor performance of activity recognition [35]. For example, in a low light environment, the physical sensor can provide a richer representation to recognize the activities than visual modality [7]. Thus, multimodal

sensors can provide complementary information to recognize activities accurately.

To overcome the limitations brought about by relying on a single modality, several works have taken a multimodal learning perspective to extract complementary features for activity classification [7]–[9], [12], [36], [37], [97], [211]–[213]. For example, Kong et al. [8] designed a knowledge distillation-based feature alignment approach to obtain complementary multimodal representation for HAR. Kazakos et al. [41] proposed a temporal window-based multimodal intermediate feature fusion approach for egocentric activity recognition. Moreover, Islam et al. [7] proposed a hierarchical attention-base fusion approach to extract and fuse multimodal representation for activity recognition. While prior approaches have achieved exemplary results, aligning and fusing relevant representations from heterogeneous modalities to obtain robust multimodal representation remains a challenging problem.

To address the aforementioned challenges in multimodal fusion, we have introduced a recurrent multimodal fusion approach by incorporating key insights from neuroscience studies. Our proposed method aims to overcome the limitations of state-of-the-art multimodal learning approaches that fuse feature in a feed-forward manner.

Unimodal Human-Activity Recognition: Human activity recognition has been extensively studied by analyzing and employing the unimodal sensor data, such as skeleton, wearable sensors, and visual (RGB or depth) modalities [214]. As generating hand-crafted features is found to be a difficult task, and these features are often highly domain-specific, many researchers are now utilizing the deep neural network-based approaches for human activity recognition.

Deep learning-based feature representation architectures, especially convolutional neural networks (CNNs) and long-short-term memory (LSTM), have been widely adopted to encode the spatio-temporal features from visual (i.e., RGB and depth) [16], [31], [93], [215]–[217] and non-visual (i.e., sEMG and IMUs) sensors data [9], [30], [218]. For example, Li et al. [93] developed a CNN-based learning method to capture the spatio-temporal co-occurrences of skeletal joints. To recognizing human activities from video data, Wang et al. proposed a 3D-CNN and LSTM-based hybrid model to detect compute salient features [219]. Recently, the graphical convolutional network has been adopted to find spatial-temporal patterns in unimodal data [32].

Although these deep-learning-based HAR methods have shown promising performances in many cases, these approaches rely significantly on modality-specific feature embeddings. If such an encoder fails to encode the feature properly because of noisy data (e.g., visual occlusion or missing or low-quality sensor data), then these activity recognition methods suffer to perform correctly.

Attention mechanism for Human-Activity Recognition: Attention mechanism has been adopted in various learning architectures to improve the feature representation as it allows the feature encoder to focus on specific parts of the representation while extracting the salient features [18], [19], [27], [37], [220]–[223]. Recently, several multi-head self-attention based methods have been proposed, which permit to disentangle the feature embedding into multiple features (multi-head) and to fuse the salient features to produce a robust feature embedding [94].

Many researchers have started adopting the attention mechanism in human activity recogni-

tion [36], [37]. For example, Xiang et al. proposed a multimodal video classification network, where they utilized an attention-based spatio-temporal feature encoder to infer modality-specific feature representation [37]. The authors explored the different types of multimodal feature fusion approaches (feature concatenation, LSTM fusion, attention fusion, and probabilistic fusion), and found that the concatenated features showed the best performance among the other fusion methods. To date, most of the HAR approaches have utilized attention-based methods for encoding the unimodal features. However, the attention mechanism has not been used for extracting and fusing salient features from multiple modalities.

9.3 Comprehending Human Interactions

Spatial Relation Grounding Datasets: Several synthetic datasets of various images of objects are generated for spatial relations grounding tasks using game engines, such as Unity [20]–[22], [172], [173]. For example, Goyal et al. [21] uses Blender to generate synthetic dataset, Rel3D, for spatial relationship recognition. However, in this dataset, the visual scene only contains two objects which simplifies the task. Lee et al. [20] addresses issues of Rel3D by generating multiple objects in the scene. Unlike these synthetic datasets, other datasets use real-world images [69], [150], [151]. For example, SpatialSense [51] uses real-world images for spatial relationship detection. In these datasets, verbal referring expressions are generated using either template-based methods [22], [68], [164] or human annotators [51], [68], [69], [150], [180]. For example, the ReferIt3D dataset [68] uses a compositional template ($< target >< spatial - relation >< reference >$) to generate verbal utterances. However, one of the limitations of these datasets is the absence of the nonverbal cues (pointing gestures and gaze).

Datasets Generator: Existing synthetic data generation tools [20]–[23] work adequately for generating referring expressions in non-embodied settings. For example, the GRiD-3D dataset [20] uses Blender [224] to generate referring expressions for relation grounding, object identification, and visual question answering. However, these tools were not designed to generate nonverbal gestures in embodied settings. Moreover, it is non-trivial to extend the existing simulators [20]–[23] and procedurally generate non-verbal gestures. In our previous work [155], we found that a template-based approach produces pointing gestures that deviate from realistic gesture.

Embodied Spatial Relation Grounding Datasets: A few datasets have been developed for embodied referring expression comprehension [24], [64], [147]–[149]. For example, the YouRefIt [24] dataset was developed in a real-world setting, which has several advantages over synthetic data. However, this dataset is limited in sample size and lack detailed annotations, which only contains 4,195 unique visual scenes. Moreover, the nonverbal interactions in the existing datasets have been captured only from the exocentric view, which may limit the model's ability to learn multiple perspectives. Although prior works have shown the importance of contrastive data samples to train models [21], the existing datasets of embodied referring expressions do not include contrastive data samples. Furthermore, the existing datasets do not include ambiguous expressions, which can be used to develop conversational embodied agents to ensure seamless human-AI interactions.

Additionally, the existing datasets do not explicitly consider generating samples with the occluded objects from a particular view, which can help diagnose the model's robustness to ground embodied spatial relations. A comparison of various referring expression datasets has been presented in Table 4.1.

Visual Question Answering Datasets: Many datasets have been developed to study visual question answering tasks [20], [22], [23], [72], [174]–[181], [183]–[187], [225]–[235]. These datasets primarily involve answering verbal questions using the visual scene as a context. For example, Antol et al. [176] developed a VQA dataset and introduced QA tasks involving an image and verbal questions about the image. This dataset contains both real-world images from the MS-COCO dataset [236], and the synthetic virtual scene contains clipart. Ren et al. [174] generated synthetic QA pairs using an algorithm that converts image description into QA form. The primary drawback of these datasets is that questions are formed using only verbal expression, whereas humans naturally use verbal and nonverbal gestures to questions in real-world settings. This limits these datasets from being used in embodied settings to comprehend questions with verbal and nonverbal gestures. Additionally, these datasets are curated from a single visual and verbal perspective, whereas humans use multiple perspectives interchangeably (e.g., speakers and observers). This single perspective dependency can create perspective biases in the dataset and limit the model to robustly comprehend questions and context.

Embodied Question Answering Datasets: In the literature, embodied question answering (EQA) tasks are designed as agents (e.g., virtual robots) navigate an environment to answer a verbal question [70], [75]–[78]. For example, Das et al. [70] developed a synthetic EQA dataset in a virtual House3D environment, where a robot navigates the environment and gathers visual information from an egocentric view to answer a verbal question. Yu et al. [75] extend this dataset and incorporate questions with multiple visual targets, such as finding multiple objects by navigating the environment. The primary limitation of these datasets is that questions include verbal expression, whereas in real-world embodied settings, humans use verbal and nonverbal gestures to question.

Although these works define *embodied interaction* as a virtual agent interacting with the environment, in another research thrust, embodied interaction refers to humans interacting with multimodal expressions. Recently, a few datasets have been developed with multimodal expressions [24], [64]. For example, Chen et al. developed a dataset, YouRefIt, with multimodal expressions for referring expression tasks. However, YouRefIt dataset contains data samples from a single visual and verbal perspective. Most importantly, as comprehending EQA tasks differs from referring expression tasks and requires complex reasoning of question and multimodal context, these datasets are less suitable for understanding QA-related tasks.

Visual-Language Model for VQA: Several visual-language representation learning models have been developed for VQA tasks [26], [27], [91], [92]. For example, Liunian et al. [26] developed VisualBERT to learn multimodal representation from a visual scene and a verbal expression to answer a question using the visual scene as a context. Kim et al. [91] designed Vision-and-Language Transformer model (ViLT) with monolithic processing of visual input to learn visual-

language representations without regional supervision of object detection. Although these models work adequately for VQA tasks, these models were designed to learn from a single visual and verbal perspective. In our work, EQA tasks involve answering a question with multimodal expressions from multiple visual and verbal perspectives. Thus, the existing models can not effectively align and learn salient representations from multiple visual and verbal perspectives to answer EQA tasks.

Chapter 10
CONCLUSION

My dissertation has two primary goals: perceiving human behavior and comprehending embodied interactions. We have developed multimodal learning models to extract salient and complementary multimodal representations from heterogeneous, missing, and noisy multimodal data. We developed Multi-GAT, a novel graphical attention-based multimodal feature learning approach for human activity recognition. Moreover, we have developed a cooperative multitask learning-based guided multimodal fusion approach, MuMu. As most of the existing multimodal models fuse representation in a feedforward way that limits the model to learn salient multimodal representations, we have developed MAVEN, a recurrent multimodal fusion approach to learning complementary multimodal representations. Our extensive experimental evaluations on three state-of-the-art multimodal human activity datasets suggest that our proposed models can help to improve human activity recognition performance.

To facilitate the research on comprehending embodied interactions, we have introduced a novel embodied simulator, CAESAR, to generate referring expressions with verbal utterances and nonverbal cues. Our simulator captures nonverbal interactions from multiple views and generates verbal expressions from multiple perspectives (actor and observer). Using CAESAR, we have developed two large-scale datasets of embodied referring expressions. Our experimental results suggest that nonverbal cues improve model performance and that existing models cannot effectively learn multiple perspective-taking to ground embodied spatial relations accurately. We believe that our simulator will help generate situated interactional datasets and training models for diverse tasks in embodied settings. Moreover, We developed a perspective-aware multitask learning model, PATRON, for comprehending referring expressions in embodied settings. We also curated a dataset of embodied referring expressions, CAESAR-PRO, to develop and evaluate learning models. Our extensive experimental results suggest that our perspective-aware guided fusion approach can extract salient multimodal representations for relation and object grounding.

We have extended our simulator to develop embodied question-answering datasets to facilitate developing and diagnosing multimodal models for comprehending question-answering interaction in embodied settings. To develop models for comprehending embodied interactions, we designed 8 novel EQA tasks requiring comprehension of questions with multimodal expressions (verbal and nonverbal gestures). To train and diagnose models for these EQA tasks, we developed a novel large-scale dataset, EQA-MX, which contains questions with multimodal expressions from multiple verbal and visual perspectives. Moreover, we developed a vector quantization-based multimodal representation learning model, VQ-Fusion, to learn salient multimodal representation from multiple visual and verbal perspectives. Our extensive experimental analyses suggest that VQ-Fusion can effectively fuse continuous multiview visual and discrete verbal representation, which

helps to improve the visual-language model's performance for all EQA tasks up to 13%.

While synthetic datasets have been instrumental in our previous works, they can not fully capture the nuances of real-world human-embodied interactions. To bridge this gap, we have curated a diverse and comprehensive dataset with multimodal embodied interactions in real-world settings. We have also introduced a novel model, Reinforced Guided Residual Representation (ReRep). The ReRep model is designed for a more flexible fusion of visual and language representations, preserving more information from both modalities and improving performance on tasks requiring a nuanced balance between visual and language information. This approach enhances the comprehension of embodied interactions, paving the way for more robust and effective AI systems capable of understanding and interacting with the world in a more natural and effective manner.

10.1 Summary of Contributions

Development of Multimodal Representation Learning Models: We have developed several multimodal representation learning models to robustly perceive human behavior. These models address the challenge of fusing heterogeneous multimodal data, as demonstrated in our IROS-2020 paper [7]. These models can extract complementary multimodal representations, as shown in our IEEE RAL-2021 [10] and AAAI-2022 [11] papers. Furthermore, these models recurrently fuse multimodal representations to produce robust task representation, as detailed in our IEEE Transaction Multimedia-2022 paper [81].

Cooperative Multitask-Based Multimodal Representation Learning: We have proposed a cooperative multitask-based multimodal representation learning approach to robustly perceive human actions. This approach, presented in our AAAI-2023 paper [11], allows the model to learn salient and complementary multimodal representations, thereby improving the performance of multiple tasks.

Development of a Simulator and Datasets for Embodied Interactions: We have developed a simulator (NeurIPS-2022 [25]) and datasets (NeurIPS-2022 [25] and AAAI-2023 [71]) to comprehend embodied interactions using multimodal cues, such as verbal utterances and nonverbal gestures. These tools provide a platform for training and evaluating models for comprehending human embodied interactions.

Benchmark Models for Embodied Interactions: We have proposed benchmark models (NeurIPS-2022 [25]) for embodied interactions. These models serve as a standard for evaluating the performance of other models in comprehending embodied interactions.

Perspective-Aware Multimodal Multitask Representation Learning Model: We have developed a perspective-aware multimodal multitask representation learning model (AAAI-2023 [71]). This model is designed to comprehend embodied interactions from multiple perspectives, thereby addressing the challenge of perspective bias in existing models.

Embodied Question-Answering Model and Datasets: We have developed an embodied question-answering model and datasets to ensure seamless interactions in embodied settings. This

model addresses the challenge of fusing continuous visual and discrete language representations by discretizing the visual representations using a vector-quantization approach.

Development of a Real-World Embodied Interaction Dataset: In contrast to previous datasets that contain synthetic data in simulated settings, we have developed a real-world embodied interaction dataset. This dataset captures embodied interactions in various indoor and outdoor settings, offering a rich resource for training and evaluating models for comprehending embodied interactions in real-world settings. The diversity of this dataset, in terms of environmental attributes and nonverbal signals, allows for a more comprehensive understanding of embodied interactions in real-world settings.

Reinforced Guided Residual-Based Multimodal Representation Model: Finally, we have proposed a novel reinforced guided residual-based multimodal representation model to comprehend embodied referring expressions. This model allows for a more flexible fusion of visual and language representations, thereby preserving more information from both modalities and improving the performance on tasks that require a more nuanced balance between visual and language information.

10.2 Lesson Learned

I have distilled the key insights and learnings I gathered throughout my Ph.D. journey. Each lesson presents a crucial aspect of the researcher's journey, potentially serving as a roadmap for others who embark on this path in the future.

10.2.1 Applying Transfer Learning from Simulated to Real-World Environments

Transfer learning is a critical technique in machine learning that allows for knowledge acquired from one domain, usually a richly resourced one, to be applied to another, potentially less-resourced domain. I leveraged this approach to train a model on synthetic data from the PATRON dataset [71] and subsequently applied it to real-world datasets, namely REMO and YouRefIt. However, the transition from simulated to real-world environments highlighted some significant challenges and resulted in performance degradation.

The core issue stemmed from the considerable difference in object textures and properties of the simulated environment compared to the real-world setting. The synthetic data lacked the complexity and the variability intrinsic to real-world data, such as nuanced human gestures or differing lighting conditions. This discrepancy meant the model was inadequately equipped to handle real-world data, leading to a dip in performance when applied to the REMO and YouRefIt datasets.

A critical distinction between the simulated and real-world settings was human behavior. For instance, in a real-world setting, people sometimes opt to physically interact with an object, such as touching it, instead of using indicative gestures, such as pointing. This contrast in behavior was

not well-represented in the synthetic dataset, and thus, the model was ill-prepared to interpret such scenarios.

The experience of applying transfer learning from simulated to real-world environments underscored the importance of considering the generalizability of a model during the design phase. It highlighted that synthetic data, while valuable for initial training, might not capture the full breadth of variability present in real-world scenarios. Therefore, for future work, there is a need to incorporate a more diverse range of scenarios, behaviors, and environmental factors in synthetic datasets to enhance their realism. Alternatively, refining our transfer learning techniques to better adapt models from synthetic to real-world data could be another promising direction. This experience illustrates the continuous process of learning and adaptation required in machine learning and will inform the approach to similar challenges in the future.

10.2.2 Developing Robust Models Using Multi-Domain Datasets

Constructing a robust model capable of performing across different environments is a cornerstone of advanced machine learning research. I discovered that harnessing datasets collected across diverse settings, such as laboratory environments, indoor and outdoor scenarios, aids in the creation of more robust models. A broader dataset can better represent the variability and complexity found in real-world situations, improving a model's capacity to generalize.

While incorporating more diverse data enhances a model's robustness, my experimental analyses also indicated that performance could degrade when training a model using multiple domain data. This degradation arises from the inherent differences between domains, such as varying scales, lighting conditions, object textures, and human behaviors. When combined into a single model, these disparate characteristics can introduce noise and confusion, hindering learning.

The principal lesson learned from this experience was the nuanced balance required in leveraging multi-domain datasets for model training. While diverse data undoubtedly improves model robustness by ensuring the representation of a wide range of scenarios, care must be taken when integrating data from different domains.

In the future, advanced strategies might need to be employed to effectively incorporate multi-domain data, such as domain adaptation techniques or specialized architectures that can handle the diverse characteristics inherent in such data. This could involve creating separate components or layers in the model for each domain or employing techniques like normalization or feature extraction to reduce the differences between domains.

This experience has broadened my understanding of the complex dynamics of training robust models. The challenge of successfully integrating multi-domain data is a rich area for future research and exploration, reminding us that the path to creating robust models is both a scientific and an artful pursuit.

10.2.3 The Complex Dynamics of Multitask Learning

Multitask learning can enhance the robustness and generalizability of a model by training it on multiple tasks simultaneously. A model with multiple tasks can gain a more holistic and diverse representation of the data, improving its performance on individual tasks.

During the experimental analysis, I found a potential competition for representation in the shared learning space when multiple tasks are trained concurrently. Essentially, different tasks may require different representations, and when trained together, they could interfere with each other, leading to performance degradation. This is particularly noticeable when the tasks are unrelated or have divergent objectives, as the model can struggle to find a shared representation that effectively addresses all tasks.

We can follow two potential solutions based on my experiences to counteract these challenges. Firstly, we should consider the compatibility of tasks before deciding to train them together. Tasks that share similar objectives or data representations are likely to benefit more from multitask learning. Secondly, devising an effective training approach that minimizes the potential negative impact of multitask learning is crucial. This could involve strategies such as alternating between different tasks during training or assigning different weights to the tasks based on their compatibility or importance.

10.2.4 Multimodal Learning and Mitigating Negative Knowledge Transfer

Using multiple modalities in representation learning offers a more comprehensive understanding of complex data, bolstering the robustness of learned representations. Different modalities, such as RGB images, depth information, and infrared (IR) data, can provide complementary perspectives that enrich a model's understanding of the task.

However, during my research, I encountered an intriguing paradox. While training on multiple modalities often bolstered representation learning, in certain instances, it introduced negative representation learning. This phenomenon essentially refers to instances where incorporating certain modalities into the learning process hampers, rather than enhances, the performance of downstream tasks. An example of this issue came when I attempted to train a model on RGB, depth, and IR images concurrently. Instead of enhancing the model's performance on tasks such as human activity recognition and embodied question answering this multimodal approach degraded task performance. Although, we have addressed this issue in our prior works [11], [81], more thorough studies will be required to learn generalized representations from a diverse set of modalities.

These experiences highlighted the importance of developing a more nuanced approach to multimodal learning. Specifically, designing a training approach that can identify the most salient modalities for a given task and focus on extracting robust representations from those modalities is paramount. This lesson underscores the importance of thoughtful modality selection and a strategic approach to training when working with multi-modal data. Future work might focus on devising strategies or mechanisms to identify and counteract negative knowledge transfer during training.

10.3 Future Work

This dissertation opens up several avenues for future research. Some potential future directions are outlined below:

10.3.1 Developing Multimodal Foundation Models For Robustly Perceiving Human Behavior and Interactions

Our multimodal learning models have shown promising results in perceiving human behavior and comprehending embodied interactions. Future work could focus on extending these models to understand human verbal and nonverbal interactions in both virtual and real-world settings. This could involve incorporating additional modalities or refining the fusion mechanisms to better capture the nuances of human interactions. Moreover, pretraining on large-scale multimodal data can help models learn generalizable representations that can be fine-tuned for robustly perceiving human behavior and interactions. Future work could focus on designing new pretraining tasks that can better capture the complexities of multimodal interactions.

Most of our current models rely on labeled data for training. However, labeled data is often expensive and time-consuming to collect. Future work could explore unsupervised or semi-supervised learning methods that can learn from unlabeled data. This could involve techniques such as self-supervised learning, contrastive learning, or generative modeling. Developing a multimodal foundation model using these techniques and unlabeled data will be an exciting future avenue of research for perceiving human behavior and interactions.

10.3.2 Continually Learn New Modalities and Domains

AI systems may need to learn and incorporate new modalities as they become available continually. Future work could explore extending our models to incorporate new modalities over time without disrupting the representations learned from existing modalities. This could involve modular architectures, dynamic routing, or multimodal meta-learning techniques for robust perception.

Our models are also trained and evaluated on specific datasets and domains. However, AI systems often need to adapt to new domains or tasks in the real world. Future work could explore domain adaptation or transfer learning methods that can adapt our models to new settings to effectively perceive dynamically changing human behavior and diverse interaction patterns.

10.3.3 Deploying in Real-World Settings

As virtual settings become increasingly prevalent, improving human interactions in these environments is a crucial challenge. Our models could be adapted to better understand and respond to human behaviors in virtual environments, thereby enhancing user experience and engagement. Moreover, Our multimodal learning models could be extended to understand instructions from people with disabilities, ensuring improved usability of assistance systems. This could involve

developing models that can interpret a wider range of modalities, including non-standard forms of communication used by individuals with specific disabilities.

Additionally, our proposed learning framework could be utilized to improve the user experience of AI assistants, such as Amazon Alexa, Microsoft Cortana, Google Home, and Apple Siri. Integrating our models into these systems could enable more effective multimodal interactions. Our models could be applied to various domains, such as online shopping and education. For example, multimodal human instruction and product content understanding could be used to enhance user interaction and product recommendations in online shopping platforms. Similarly, online educational platforms could use content understanding approaches to gauge student engagement and provide personalized learning recommendations.

Finally, we believe that our proposed multimodal human interaction simulator can help advance the Human-AI Interaction field. Future work could focus on using this simulator to develop and evaluate learning models for Human-AI Interaction systems, thereby contributing to advancing this important research field.

REFERENCES

[1] M. A. Meredith and B. E. Stein, "Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration," *Journal of Neurophysiology*, vol. 56, no. 3, 1986.

[2] M. T. Wallace and B. E. Stein, "Development of multisensory neurons and multisensory integration in cat superior colliculus," *Journal of Neuroscience*, vol. 17, no. 7, pp. 2429–2444, 1997.

[3] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *ICML*, 2011.

[4] C. Spence, "Multisensory perception," in *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*. American Cancer Society, 2018, pp. 1–56, ISBN: 9781119170174.

[5] "Multisensory perception," in *Multisensory Perception*, K. Sathian and V. Ramachandran, Eds., Academic Press, 2020, pp. xiii–xv, ISBN: 978-0-12-812492-5.

[6] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[7] M. M. Islam and T. Iqbal, "Hamlet: A hierarchical multimodal attention-based human activity recognition algorithm," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 10 285–10 292. DOI: 10.1109/IROS45743.2020.9340987.

[8] Q. Kong, Z. Wu, Z. Deng, M. Klinkigt, B. Tong, and T. Murakami, "MMAct: A large-scale dataset for cross modal human action understanding," in *IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 8658–8667.

[9] A. Kubota, T. Iqbal, J. A. Shah, and L. D. Riek, "Activity recognition in manufacturing: The roles of motion capture and semg+ inertial wearables in detecting fine vs. gross motion," in *ICRA*, 2019.

[10] M. M. Islam and T. Iqbal, "Multi-gat: A graphical attention-based hierarchical multimodal representation learning approach for human activity recognition," in *IEEE Robotics and Automation Letters (RA-L)*, 2021.

[11] M. M. Islam and T. Iqbal, "Mumu: Cooperative multitask learning-based guided multimodal fusion," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 1, pp. 1043–1051, Jun. 2022. DOI: 10.1609/aaai.v36i1.19988. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/19988.

[12] H. R. V. Joze, A. Shaban, M. L. Iuzzolino, and K. Koishida, "MMTM: Multimodal transfer module for cnn fusion," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[13] Y. Zhang, C. Cao, J. Cheng, and H. Lu, "Egogesture: A new dataset and benchmark for egocentric hand gesture recognition," *IEEE Transactions on Multimedia*, vol. 20, no. 5, pp. 1038–1050, 2018.

[14] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks," in *CVPR*, 2016, pp. 4207–4215.

[15] S. Samyoun, M. M. Islam, T. Iqbal, and J. Stankovic, "M3Sense: Affect-agnostic multitask representation learning using multimodal wearable sensors," in *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, 2022.

[16] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF Con- ference on Computer Vision and Pattern Recognition*, 2019.

[17] M. K. Hasan, W. Rahman, A. Bagher Zadeh, *et al.*, "UR-FUNNY: A multimodal language dataset for understanding humor," in *EMNLP-IJCNLP*, 2019, pp. 2046–2056.

[18] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE/CVF Con- ference on Computer Vision and Pattern Recognition*, 2017, pp. 375–383.

[19] K. Xu, J. Ba, R. Kiros, *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015, pp. 2048–2057.

[20] J. H. Lee, M. Kerzel, K. Ahrens, C. Weber, and S. Wermter, "What is right for me is not yet right for you: A dataset for grounding relative directions via multi-task learning," *arXiv preprint arXiv:2205.02671*, 2022.

[21] A. Goyal, K. Yang, D. Yang, and J. Deng, "Rel3d: A minimally contrastive benchmark for grounding spatial relations in 3d," *Advances in Neural Information Processing Systems*, vol. 33, pp. 10 514–10 525, 2020.

[22] R. Liu, C. Liu, Y. Bai, and A. L. Yuille, "Clevr-ref+: Diagnosing visual reasoning with referring expressions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4185–4194.

[23] J. Johnson, B. Hariharan, L. Van Der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick, "Clevr: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2901–2910.

[24] Y. Chen, Q. Li, D. Kong, *et al.*, "Yourefit: Embodied reference understanding with language and gesture," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1385–1395.

[25] M. M. Islam, R. M. Mirzaiee, A. Gladstone, H. N. Green, and T. Iqbal, "CAESAR: A multimodal simulator for generating embodied relationship grounding dataset," in *NeurIPS*, 2022.

[26] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," in *Advances in Neural Information Processing Systems*, 2019.

[27] J. Lu, D. Batra, D. Parikh, and S. Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *NeurIPS*, 2019.

[28] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, 2019.

[29] T. Iqbal and L. D. Riek, "Human-robot teaming: Approaches from joint action and dynamical systems," *Humanoid robotics: A reference*, pp. 2293–2312, 2019.

[30] A. E. Frank, A. Kubota, and L. D. Riek, "Wearable activity recognition for robust human-robot teaming in safety-critical environments via hybrid neural networks," in *IROS*, IEEE, 2019, pp. 449–454.

[31] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3d action recognition," in *Proceedings of the IEEE/CVF Con- ference on Computer Vision and Pattern Recognition*, 2017, pp. 3288–3297.

[32] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, 2018.

[33] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3d skeletal data: A review," *Computer Vision and Image Understanding*, vol. 158, pp. 85–105, 2017.

[34] T. Iqbal, M. Moosaei, and L. D. Riek, "Tempo adaptation and anticipation methods for human-robot teams," in *RSS, Planning HRI: Shared Autonomy Collab. Robot. Workshop*, 2016.

[35] T. Baltrušaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[36] G. Liu, J. Qian, F. Wen, X. Zhu, R. Ying, and P. Liu, "Action recognition based on 3d skeleton and rgb frame fusion," in *IROS*, Nov. 2019, pp. 258–264.

[37] X. Long, C. Gan, G. De Melo, *et al.*, "Multimodal keyless attention fusion for video classification," in *AAAI*, 2018.

[38] S. Münzner, P. Schmidt, A. Reiss, M. Hanselmann, R. Stiefelhagen, and R. Dürichen, "CNN-based sensor fusion techniques for multimodal human activity recognition," in *Proceedings of the 2017 ACM International Symposium on Wearable Computers*, Maui, Hawaii, 2017, pp. 158–165.

[39] M. K. Hasan, W. Rahman, A. Bagher Zadeh, *et al.*, "Ur-funny: A multimodal language dataset for understanding humor," *EMNLP-IJCNLP*, 2019.

[40] F. Xiao, Y. J. Lee, K. Grauman, J. Malik, and C. Feichtenhofer, "Audiovisual slowfast networks for video recognition," *arXiv preprint arXiv:2001.08740*, 2020.

[41] E. Kazakos, A. Nagrani, A. Zisserman, and D. Damen, "Epic-fusion: Audio-visual temporal binding for egocentric action recognition," in *ICCV*, 2019, pp. 5492–5501.

[42] D. Ramachandram and G. W. Taylor, "Deep multimodal learning: A survey on recent advances and trends," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 96–108, 2017.

[43] E. Macaluso, "Multisensory processing in sensory-specific cortical areas," *The neuroscientist*, vol. 12, no. 4, pp. 327–338, 2006.

[44] C. E. Schroeder and J. Foxe, "Multisensory contributions to low-level,'unisensory'processing," *Current opinion in neurobiology*, vol. 15, no. 4, pp. 454–458, 2005.

[45] A. Akula, V. Jampani, S. Changpinyo, and S.-C. Zhu, "Robust visual reasoning via language guided neural module networks," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[46] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 5100–5111. DOI: 10.18653/v1/D19-1514. [Online]. Available: https://aclanthology.org/D19-1514.

[47] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, PMLR, 2021, pp. 8748–8763.

[48] J. Roh, K. Desingh, A. Farhadi, and D. Fox, "Languagerefer: Spatial-language model for 3d visual grounding," in *Conference on Robot Learning*, PMLR, 2022, pp. 1046–1056.

[49] A. Akula, S. Gella, K. Wang, S.-c. Zhu, and S. Reddy, "Mind the context: The impact of contextualization in neural module networks for grounding visual referring expressions," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 6398–6416.

[50] L. Yu, Z. Lin, X. Shen, *et al.*, "Mattnet: Modular attention network for referring expression comprehension," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1307–1315.

[51] K. Yang, O. Russakovsky, and J. Deng, "Spatialsense: An adversarially crowdsourced benchmark for spatial relation recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2051–2060.

[52] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, "Mdetr-modulated detection for end-to-end multi-modal understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1780–1790.

[53] S. Yang, G. Li, and Y. Yu, "Cross-modal relationship inference for grounding referring expressions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4145–4154.

[54] L. Ye, M. Rochan, Z. Liu, and Y. Wang, "Cross-modal self-attention network for referring image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 502–10 511.

[55] D. McNeill, *How language began: Gesture and speech in human evolution*. Cambridge University Press, 2012.

[56] M. A. Arbib, K. Liebal, and S. Pika, "Primate vocalization, gesture, and the evolution of human language," *Current anthropology*, vol. 49, no. 6, pp. 1053–1076, 2008.

[57] U. Liszkowski, M. Carpenter, A. Henning, T. Striano, and M. Tomasello, "Twelve-month-olds point to share attention and interest," *Developmental science*, vol. 7, no. 3, pp. 297–307, 2004.

[58] C. Colonnesi, G. J. J. Stams, I. Koster, and M. J. Noom, "The relation between pointing and language development: A meta-analysis," *Developmental Review*, vol. 30, no. 4, pp. 352–366, 2010.

[59] V. Corkum and C. Moore, "The origins of joint visual attention in infants.," *Developmental psychology*, vol. 34, no. 1, p. 28, 1998.

[60] G. Butterworth, F. Franco, B. McKenzie, L. Graupner, and B. Todd, "Dynamic aspects of visual event perception and the production of pointing by human infants," *British Journal of Developmental Psychology*, vol. 20, no. 1, pp. 1–24, 2002.

[61] M. Scaife and J. S. Bruner, "The capacity for joint visual attention in the infant," *Nature*, vol. 253, no. 5489, pp. 265–266, 1975.

[62] J. M. Iverson and S. Goldin-Meadow, "Gesture paves the way for language development," *Psychological science*, vol. 16, no. 5, pp. 367–371, 2005.

[63] M. Halina, F. Rossano, and M. Tomasello, "The ontogenetic ritualization of bonobo gestures," *Animal cognition*, vol. 16, no. 4, pp. 653–666, 2013.

[64] B. Schauerte and G. A. Fink, "Focusing computational visual attention in multi-modal human-robot interaction," in *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, ser. ICMI-MLMI '10, Beijing, China: Association for Computing Machinery, 2010, ISBN: 9781450304146.

[65] B. Dai, Y. Zhang, and D. Lin, "Detecting visual relationships with deep relational networks," in *Proceedings of the IEEE conference on computer vision and Pattern recognition*, 2017, pp. 3076–3086.

[66] B. Zhuang, L. Liu, C. Shen, and I. Reid, "Towards context-aware interaction recognition for visual relationship detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 589–598.

[67] H. Zhang, Z. Kyaw, S.-F. Chang, and T.-S. Chua, "Visual translation embedding network for visual relation detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5532–5540.

[68] P. Achlioptas, A. Abdelreheem, F. Xia, M. Elhoseiny, and L. J. Guibas, "ReferIt3D: Neural listeners for fine-grained 3d object identification in real-world scenes," in *16th European Conference on Computer Vision (ECCV)*, 2020.

[69] S. Kazemzadeh, V. Ordonez, M. Matten, and T. Berg, "ReferItGame: Referring to objects in photographs of natural scenes," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 787–798. DOI: 10.3115/v1/D14-1086. [Online]. Available: https://aclanthology.org/D14-1086.

[70] A. Das, S. Datta, G. Gkioxari, S. Lee, D. Parikh, and D. Batra, "Embodied question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1–10.

[71] M. M. Islam, A. Gladstone, and T. Iqbal, "PATRON: Perspective-aware multitask model for referring expression grounding using embodied multimodal cues," in *AAAI [Under Review]*, 2023.

[72] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6700–6709.

[73] J. Pfeiffer, G. Geigle, A. Kamath, *et al.*, "Xgqa: Cross-lingual visual question answering," *arXiv preprint arXiv:2109.06082*, 2021.

[74] F. Liu, E. Bugliarello, E. M. Ponti, S. Reddy, N. Collier, and D. Elliott, "Visually grounded reasoning across languages and cultures," *arXiv preprint arXiv:2109.13238*, 2021.

[75] L. Yu, X. Chen, G. Gkioxari, M. Bansal, T. L. Berg, and D. Batra, "Multi-target embodied question answering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6309–6318.

[76] H. Luo, G. Lin, Z. Liu, F. Liu, Z. Tang, and Y. Yao, "Segeqa: Video segmentation based visual attention for embodied question answering," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9667–9676.

[77] D. Gordon, A. Kembhavi, M. Rastegari, J. Redmon, D. Fox, and A. Farhadi, "Iqa: Visual question answering in interactive environments," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4089–4098.

[78] S. Tan, W. Xiang, H. Liu, D. Guo, and F. Sun, "Multi-agent embodied question answering in interactive environments," in *European Conference on Computer Vision*, Springer, 2020, pp. 663–678.

[79] S. Kita, *Pointing: Where language, culture, and cognition meet*. Psychology Press, 2003.

[80] S. W. Cook, Z. Mitchell, and S. Goldin-Meadow, "Gesturing makes learning last," *Cognition*, vol. 106, no. 2, pp. 1047–1058, 2008.

[81] M. M. Islam, M. S. Yasar, and T. Iqbal, "MAVEN: A memory augmented recurrent approach for multimodal fusion," in *IEEE Transaction on Multimedia*, 2022.

[82] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[83] P. Xu, A. Madotto, C.-S. Wu, J. H. Park, and P. Fung, "Emo2Vec: Learning generalized emotion representation by multi-task training," in *Proceedings of the 9th Workshop on*

Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, ACL, 2018.

[84] L. Zhou, Z. Cui, C. Xu, *et al.*, "Pattern-structure diffusion for multi-task learning," in *CVPR*, Jun. 2020.

[85] A. Achille, M. Lam, R. Tewari, *et al.*, "Task2vec: Task embedding for meta-learning," in *ICCV*, Oct. 2019.

[86] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *CVPR*, Jun. 2018.

[87] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *CVPR*, 2019, pp. 1871–1880.

[88] X. Sun, R. Panda, R. Feris, and K. Saenko, "Adashare: Learning what to share for efficient deep multi-task learning," in *NeurIPS*, vol. 33, 2020, pp. 8728–8740.

[89] K. Greff, F. Belletti, L. Beyer, *et al.*, "Kubric: A scalable dataset generator," 2022.

[90] *Optitrack*, `https://optitrack.com/`, Accessed: 2022-06-03.

[91] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *International Conference on Machine Learning*, PMLR, 2021, pp. 5583–5594.

[92] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, PMLR, 2021, pp. 8748–8763.

[93] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," *IJCAI*, 2018.

[94] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, pp. 5999–6009, 2017, ISSN: 10495258. arXiv: `arXiv:1706.03762v5`.

[95] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2017.

[96] L. Xia, C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*, IEEE, 2012, pp. 20–27.

[97] C. Chen, R. Jafari, and N. Kehtarnavaz, "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," in *2015 IEEE ICIP*, 2015, pp. 168–172.

[98] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[99] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," in *ICLR*, 2016.

[100] A. Paszke, S. Gross, F. Massa, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.

[101] M. F. Bulbul, Y. Jiang, and J. Ma, "Dmms-based multiple features fusion for human action recognition," *IJMDEM*, vol. 6, no. 4, 2015.

[102] J. Imran and P. Kumar, "Human action recognition using rgb-d sensor and deep convolutional neural networks," in *ICACCI*, 2016.

[103] C. Li, Y. Hou, P. Wang, and W. Li, "Joint distance maps based action recognition with convolutional neural networks," *IEEE Signal Processing Letters*, vol. 24, no. 5, pp. 624–628, 2017.

[104] P. Wang, S. Wang, Z. Gao, Y. Hou, and W. Li, "Structured images for rgb-d action recognition," in *Proceedings of the IEEE/CVF Con- ference on Computer Vision and Pattern Recognition Workshop*, 2017, pp. 1005–1014.

[105] Y. Hou, Z. Li, P. Wang, and W. Li, "Skeleton optical spectra-based action recognition using convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, 2016.

[106] T. Liu, J. Kong, and M. Jiang, "RGB-D action recognition using multimodal correlative representation learning model," *IEEE Sensors Journal*, vol. 19, no. 5, pp. 1862–1872, Mar. 2019.

[107] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps," in *Proceedings of the IEEE/CVF Con- ference on Computer Vision and Pattern Recognition*, 2018.

[108] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural computation*, 1991.

[109] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *Advances in Neural Information Processing Systems*, 2015.

[110] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Berkeley mhad: A comprehensive multimodal human action database," in *WACV*, IEEE, 2013, pp. 53–60.

[111] L. Wang, Y. Xiong, Z. Wang, *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *European Conference on Computer Vision (ECCV)*, 2016.

[112] A. Goyal, A. Lamb, J. Hoffmann, *et al.*, *Recurrent independent mechanisms*, 2019. arXiv: 1909.10893 [cs.LG].

[113] F. Locatello, D. Weissenborn, T. Unterthiner, *et al.*, "Object-centric learning with slot attention," in *Advances in Neural Information Processing Systems*, 2020.

[114] H. Tang, M. Schrimpf, W. Lotter, *et al.*, "Recurrent computations for visual pattern completion," *Proceedings of the National Academy of Sciences*, vol. 115, no. 35, pp. 8835–8840, 2018.

[115] V. A. Lamme and P. R. Roelfsema, "The distinct modes of vision offered by feedforward and recurrent processing," *Trends in neurosciences*, vol. 23, no. 11, pp. 571–579, 2000.

[116] M. Meredith, M. Wallace, and B. Stein, "Visual, auditory and somatosensory convergence in output neurons of the cat superior colliculus: Multisensory properties of the tecto-reticulo-spinal projection," *Experimental Brain Research*, vol. 88, pp. 181–186, 1992.

[117] H. Bahuleyan, L. Mou, O. Vechtomova, and P. Poupart, "Variational attention for sequence-to-sequence models," *arXiv preprint arXiv:1712.08207*, 2017.

[118] Y. Deng, Y. Kim, J. Chiu, D. Guo, and A. Rush, "Latent alignment and variational attention," in *Advances in Neural Information Processing Systems*, 2018, pp. 9712–9724.

[119] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *stat*, vol. 1050, p. 1, 2014.

[120] W. Falcon, "Pytorch lightning," *GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning*, 2019.

[121] R. Dror, S. Shlomov, and R. Reichart, "Deep dominance-how to properly compare deep neural models," in *ACL*, 2019, pp. 2773–2785.

[122] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[123] M. Sabokrou, M. Pourreza, M. Fayyaz, *et al.*, "Avid: Adversarial visual irregularity detection," in *Asian Conference on Computer Vision*, Springer, 2018, pp. 488–505.

[124] T. Iqbal and L. D. Riek, "Human robot teaming: Approaches from joint action and dynamical systems," *Humanoid Robotics*, 2017.

[125] T. Iqbal and L. D. Riek, "Temporal anticipation and adaptation methods for fluent human-robot teaming," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 3736–3743.

[126] M. S. Yasar and T. Iqbal, "A scalable approach to predict multi-agent motion for human-robot collaboration," in *IEEE Robotics and Automation Letters (RA-L)*, 2021.

[127] M. S. Yasar and T. Iqbal, "Robots that can anticipate and learn in human-robot teams," in *HRI*, 2022.

[128] H. N. Green, M. M. Islam, S. Ali, and T. Iqbal, "Who's laughing nao? examining perceptions of failure in a humorous robot partner," in *HRI*, 2022, pp. 313–322.

[129] H. N. Green, M. M. Islam, S. Ali, and T. Iqbal, "Ispy a humorous robot: Evaluating the perceptions of humor types in a robot partner," in *AAAI Spring Symposium*, 2022.

[130] T. Iqbal, S. Li, C. Fourie, B. Hayes, and J. A. Shah, "Fast online segmentation of activities from partial trajectories," in *ICRA*, 2019.

[131] E. Pakdamanian, S. Sheng, S. Baee, S. Heo, S. Kraus, and L. Feng, "Deeptake: Prediction of driver takeover behavior using multimodal data," in *CHI*, 2020.

[132] M. S. Ryoo, B. Rothrock, C. Fleming, and H. J. Yang, "Privacy-preserving human activity recognition from extreme low resolution," in *AAAI*, 2017.

[133] H. Zhang and L. E. Parker, "4-dimensional local spatio-temporal features for human activity recognition," in *IROS*, 2011.

[134] L. Fan, W. Huang, C. Gan, S. Ermon, B. Gong, and J. Huang, "End-to-end learning of motion representation for video understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6016–6025.

[135] M. M. Arzani, M. Fathy, H. Aghajan, A. A. Azirani, K. Raahemifar, and E. Adeli, "Structured prediction with short/long-range dependencies for human activity recognition from depth skeleton data," in *IROS*, 2017.

[136] T. Iqbal, S. Rack, and L. D. Riek, "Movement coordination in human-robot teams: A dynamical systems approach," *IEEE Transactions on Robotics*, vol. 32, no. 4, pp. 909–919, 2016.

[137] I. Batzianoulis, S. El-Khoury, E. Pirondini, M. Coscia, S. Micera, and A. Billard, "Emg-based decoding of grasp gestures in reaching-to-grasping motions," *RAS*, 2017.

[138] A. Roitberg, N. Somani, A. Perzylo, M. Rickert, and A. Knoll, "Multimodal human activity recognition for industrial manufacturing processes in robotic workcells," in *ICMI*, 2015.

[139] J.-M. Perez-Rua, V. Vielzeuf, S. Pateux, M. Baccouche, and F. Jurie, "MFAS: Multimodal fusion architecture search," in *Proceedings of the IEEE/CVF Con- ference on Computer Vision and Pattern Recognition*, Jun. 2019.

[140] G. Awad, A. Butt, K. Curtis, *et al.*, "Trecvid 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search," in *Proceedings of TRECVID 2018*, 2018.

[141] U. Liszkowski, M. Carpenter, T. Striano, and M. Tomasello, "12-and 18-month-olds point to provide information for others," *Journal of cognition and development*, vol. 7, no. 2, pp. 173–187, 2006.

[142] M. Tomasello, *Origins of human communication*. MIT press, 2010.

[143] N. Tang, S. Stacy, M. Zhao, G. Marquez, and T. Gao, "Bootstrapping an imagined we for cooperation.," in *CogSci*, 2020.

[144] S. Stacy, Q. Zhao, M. Zhao, M. Kleiman-Weiner, and T. Gao, "Intuitive signaling through an" imagined we'".," in *CogSci*, 2020.

[145] D. Batra, A. X. Chang, S. Chernova, *et al.*, "Rearrangement: A challenge for embodied ai," *arXiv preprint arXiv:2011.01975*, 2020.

[146] P. Kratzer, S. Bihlmaier, N. B. Midlagajni, R. Prakash, M. Toussaint, and J. Mainprice, "Mogaze: A dataset of full-body motions that includes workspace geometry and eye-gaze," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 367–373, 2020.

[147] B. Schauerte, J. Richarz, and G. A. Fink, "Saliency-based identification and recognition of pointed-at objects," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 4638–4643. DOI: 10.1109/IROS.2010.5649430.

[148] D. Shukla, O. Erkent, and J. Piater, "Probabilistic detection of pointing directions for human-robot interaction," in *2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, 2015, pp. 1–8. DOI: 10.1109/DICTA.2015.7371296.

[149] D. Shukla, Ö. Erkent, and J. Piater, "A multi-view hand gesture rgb-d dataset for human-robot interaction scenarios," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2016, pp. 1084–1091. DOI: 10.1109/ROMAN.2016.7745243.

[150] L. Yu, P. Poirson, S. Yang, A. C. Berg, and T. L. Berg, "Modeling context in referring expressions," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., Cham: Springer International Publishing, 2016, pp. 69–85.

[151] J. Mao, J. Huang, A. Toshev, O. Camburu, A. Yuille, and K. Murphy, "Generation and comprehension of unambiguous object descriptions," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 11–20. DOI: 10.1109/CVPR.2016.9.

[152] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 2641–2649. DOI: 10.1109/ICCV.2015.303.

[153] H. De Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, and A. Courville, "Guess-what?! visual object discovery through multi-modal dialogue," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5503–5512.

[154] Z. Chen, P. Wang, L. Ma, K.-Y. K. Wong, and Q. Wu, "Cops-ref: A new dataset and task on compositional referring expression comprehension," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 086–10 095.

[155] R. Mirzaiee and T. Iqbal, "Stochastic synthesis of pointing gestures for virtual agents," *Under Review*,

[156] *Mixamo*, https://www.mixamo.com, Accessed: 2022-06-03.

[157] J. K. Hodgins, W. L. Wooten, D. C. Brogan, and J. F. O'Brien, "Animating human athletics," in *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '95, New York, NY, USA: Association for Computing Machinery, 1995, pp. 71–78, ISBN: 0897917014. DOI: 10.1145/218380.218414. [Online]. Available: https://doi.org/10.1145/218380.218414.

[158] M. Gleicher, "Retargetting motion to new characters," in *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '98, New York, NY, USA: Association for Computing Machinery, 1998, pp. 33–42, ISBN: 0897919998. DOI: 10.1145/280814.280820. [Online]. Available: https://doi.org/10.1145/280814.280820.

[159] C. Yuksel, S. Schaefer, and J. Keyser, "Parameterization and applications of catmull–rom curves," *Computer-Aided Design*, vol. 43, no. 7, pp. 747–755, 2011, The 2009 SIAM/ACM Joint Conference on Geometric and Physical Modeling, ISSN: 0010-4485. DOI: https://doi.org/10.1016/j.cad.2010.08.008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0010448510001533.

[160] S. Tayebi Arasteh and A. Kalisz, "Conversion between cubic bezier curves and catmull–rom splines," Jul. 2021.

[161] S. Y. Kula, *Bezier solution*, version 2.3.0, Accessed: 2022-06-03. [Online]. Available: https://github.com/yasirkula/UnityBezierSolution.

[162] *Animation rigging*, https://docs.unity3d.com/Packages/com.unity.animation.rigging@0.2, Accessed: 2022-06-03.

[163] W. Hong, *Dynamic bone for unity3d*, version 1.3.2. [Online]. Available: https://assetstore.unity.com/packages/tools/animation/dynamic-bone-16743%5C#description.

[164] F. Liu, G. E. T. Emerson, and N. Collier, "Visual spatial reasoning," *ArXiv*, vol. abs/2205.00363, 2022.

[165] *Unity vectrosity package*, https://assetstore.unity.com/packages/tools/particles-effects/vectrosity-82, Accessed: 2022-06-03.

[166] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.

[167] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[168] T. Wolf, L. Debut, V. Sanh, *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[169] X. Zhai, X. Wang, B. Mustafa, *et al.*, "Lit: Zero-shot transfer with locked-image text tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 123–18 133.

[170] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[171] Y.-C. Chen, L. Li, L. Yu, *et al.*, "Uniter: Universal image-text representation learning," in *European conference on computer vision*, Springer, 2020, pp. 104–120.

[172] J. Viethen and R. Dale, "The use of spatial relations in referring expression generation," in *Proceedings of the Fifth International Natural Language Generation Conference*, 2008, pp. 59–67.

[173] K. van Deemter, I. van der Sluis, and A. Gatt, "Building a semantically transparent corpus for the generation of referring expressions.," in *Proceedings of the Fourth International Natural Language Generation Conference*, 2006, pp. 130–132.

[174] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," *Advances in neural information processing systems*, vol. 28, 2015.

[175] D. Gurari, Q. Li, A. J. Stangl, *et al.*, "Vizwiz grand challenge: Answering visual questions from blind people," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3608–3617.

[176] S. Antol, A. Agrawal, J. Lu, *et al.*, "VQA: Visual Question Answering," in *International Conference on Computer Vision (ICCV)*, 2015.

[177] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, "Are you talking to a machine? dataset and methods for multilingual image question," *Advances in neural information processing systems*, vol. 28, 2015.

[178] L. Yu, E. Park, A. C. Berg, and T. L. Berg, "Visual madlibs: Fill in the blank image generation and question answering," *arXiv preprint arXiv:1506.00278*, 2015.

[179] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7w: Grounded question answering in images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4995–5004.

[180] R. Krishna, Y. Zhu, O. Groth, *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017.

[181] K. Kafle, B. Price, S. Cohen, and C. Kanan, "Dvqa: Understanding data visualizations via question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5648–5656.

[182] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the v in vqa matter: Elevating the role of image understanding in visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 6904–6913.

[183] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A deep learning approach to visual question answering," *International Journal of Computer Vision*, vol. 125, no. 1, pp. 110–135, 2017.

[184] P. Wang, Q. Wu, C. Shen, A. v. d. Hengel, and A. Dick, "Explicit knowledge-based reasoning for visual question answering," *arXiv preprint arXiv:1511.02570*, 2015.

[185] P. Wang, Q. Wu, C. Shen, A. Dick, and A. Van Den Hengel, "Fvqa: Fact-based visual question answering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 10, pp. 2413–2427, 2017.

[186] S. A. Hasan, Y. Ling, O. Farri, J. Liu, H. Müller, and M. Lungren, "Overview of imageclef 2018 medical domain visual question answering task," 10-14 September 2018, Tech. Rep., 2018.

[187] M. Mathew, D. Karatzas, and C. Jawahar, "Docvqa: A dataset for vqa on document images," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 2200–2209.

[188] Z. Huang, Z. Zeng, B. Liu, D. Fu, and J. Fu, "Pixel-bert: Aligning image pixels with text by deep multi-modal transformers," *arXiv preprint arXiv:2004.00849*, 2020.

[189] A. Van Den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[190] D. Liu, A. Lamb, K. Kawaguchi, *et al.*, "Discrete-valued neural communication in structured architectures enhances generalization," 2021.

[191] T. Beck, B. Bohlender, C. Viehmann, *et al.*, "Adapterhub playground: Simple and flexible few-shot learning with adapters," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2022, pp. 61–75.

[192] A. Ansell, E. M. Ponti, J. Pfeiffer, *et al.*, "Mad-g: Multilingual adapter generation for efficient cross-lingual transfer," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021, pp. 4762–4781.

[193] A. Rücklé, G. Geigle, M. Glockner, *et al.*, "Adapterdrop: On the efficiency of adapters in transformers," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 7930–7946.

[194] J. Pfeiffer, A. Rücklé, C. Poth, *et al.*, "Adapterhub: A framework for adapting transformers," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 46–54.

[195] J. Pfeiffer, I. Vulić, I. Gurevych, and S. Ruder, "Mad-x: An adapter-based framework for multi-task cross-lingual transfer," *arXiv preprint arXiv:2005.00052*, 2020.

[196] J. Pfeiffer, A. Kamath, A. Rücklé, K. Cho, and I. Gurevych, "Adapterfusion: Non-destructive task composition for transfer learning," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2021, pp. 487–503.

[197] P. P. Liang, A. Zadeh, and L.-P. Morency, "Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions," *arXiv preprint arXiv:2209.03430*, 2022.

[198] Y. Wu, L. Zhu, Y. Yan, and Y. Yang, "Dual attention matching for audio-visual event localization," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6292–6300.

[199] Microsoft, *Azure kinect dk*, Microsoft Corp. Available: https://azure.microsoft.com/en-us/services/kinect-dk/, 2019.

[200] OhmniLabs, *Ohmni robotic system*, Available: https://ohmnilabs.com/ohmni-robotic-system/, 2022.

[201] P. Labs, *Pupil invisible*, Available: https://pupil-labs.com/, 2021.

[202] N. Ramalingam, J. N. McManus, W. Li, and C. D. Gilbert, "Top-down modulation of lateral interactions in visual cortex," *Journal of Neuroscience*, vol. 33, no. 5, pp. 1773–1789, 2013.

[203] C. D. Gilbert and W. Li, "Top-down influences on visual processing," *Nature Reviews Neuroscience*, vol. 14, no. 5, pp. 350–363, 2013.

[204] E. B. Issa, C. F. Cadieu, and J. J. DiCarlo, "Neural dynamics at successive stages of the ventral visual stream are consistent with hierarchical error signals," *Elife*, vol. 7, e42870, 2018.

[205] D. Hu, C. Wang, F. Nie, and X. Li, "Dense multimodal fusion for hierarchically joint representation," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 3941–3945. DOI: 10.1109/ICASSP.2019.8683898.

[206] J. Ramaswamy and S. Das, "See the sound, hear the pixels," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 2970–2979.

[207] Y. Wu and Y. Yang, "Exploring heterogeneous clues for weakly-supervised audio-visual video parsing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1326–1335.

[208] R. Hu, A. Singh, T. Darrell, and M. Rohrbach, "Iterative answer prediction with pointer-augmented multimodal transformers for textvqa," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9992–10 002.

[209] X. Wang, L. Zhu, and Y. Yang, "T2vlad: Global-local sequence alignment for text-video retrieval," in *Proceedings of the IEEE/CVF Con- ference on Computer Vision and Pattern Recognition*, 2021, pp. 5079–5088.

[210] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *ECCV*, 2018, pp. 305–321.

[211] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in *Proceedings of the IEEE/CVF Con- ference on Computer Vision and Pattern Recognition*, 2019, pp. 244–253.

[212] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE/CVF Con- ference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.

[213] N. C. Garcia, P. Morerio, and V. Murino, "Modality distillation with multiple stream networks for action recognition," in *ECCV*, 2018.

[214] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations," in *Twenty-Third AAAI*, 2013.

[215] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6450–6459.

[216] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *ECCV*, 2018, pp. 803–818.

[217] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *ICCV*, 2015, pp. 4489–4497.

[218] M. S. Totty and E. Wade, "Muscle activation and inertial motion data for noninvasive classification of activities of daily living," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 5, pp. 1069–1076, 2017.

[219] X. Wang, L. Gao, J. Song, and H. Shen, "Beyond frame-level cnn: Saliency-aware 3-d cnn with lstm for video action recognition," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 510–514, 2016.

[220] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, 2015.

[221] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[222] V. Mnih, N. Heess, A. Graves, *et al.*, "Recurrent models of visual attention," in *NeurIPS*, 2014, pp. 2204–2212.

[223] P. Gao, H. You, Z. Zhang, X. Wang, and H. Li, "Multi-modality latent interaction network for visual question answering," in *ICCV*, 2019.

[224] *Blender*, `https://www.blender.org/`, Accessed: 2022-06-03.

[225] K. Kafle and C. Kanan, "An analysis of visual question answering algorithms," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1965–1973.

[226] A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, and A. Farhadi, "A diagram is worth a dozen images," in *European conference on computer vision*, Springer, 2016, pp. 235–251.

[227] S. E. Kahou, V. Michalski, A. Atkinson, Á. Kádár, A. Trischler, and Y. Bengio, "Figureqa: An annotated figure dataset for visual reasoning," *arXiv preprint arXiv:1710.07300*, 2017.

[228] L.-C. Huang, K. Kulkarni, A. Jha, S. Lohit, S. Jayasuriya, and P. Turaga, "Cs-vqa: Visual question answering with compressively sensed images," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, IEEE, 2018, pp. 1283–1287.

[229] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Neural module networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 39–48.

[230] S.-H. Chou, W.-L. Chao, W.-S. Lai, M. Sun, and M.-H. Yang, "Visual question answering on 360deg images," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020, pp. 1607–1616.

[231] A. Singh, V. Natarajan, M. Shah, *et al.*, "Towards vqa models that can read," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8317–8326.

[232] A. F. Biten, R. Tito, A. Mafla, *et al.*, "Scene text visual question answering," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4291–4301.

[233] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "Ok-vqa: A visual question answering benchmark requiring external knowledge," in *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 2019, pp. 3195–3204.

[234] A. Mishra, S. Shekhar, A. K. Singh, and A. Chakraborty, "Ocr-vqa: Visual question answering by reading text in images," in *2019 international conference on document analysis and recognition (ICDAR)*, IEEE, 2019, pp. 947–952.

[235] A. Kembhavi, M. Seo, D. Schwenk, J. Choi, A. Farhadi, and H. Hajishirzi, "Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension," in *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, 2017, pp. 4999–5007.

[236] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft coco: Common objects in context," in *European conference on computer vision*, Springer, 2014, pp. 740–755.

Multimodal and Multitask Representation Learning For Perceiving Embodied Interactions

---

A

Dissertation

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

---

in partial fulfillment

of the requirements for the degree

Doctor of Philosophy

by

Md Mofijul Islam

August 2023

APPROVAL SHEET

This

Dissertation

is submitted in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

Author: Md Mofijul Islam

This Dissertation has been read and approved by the examing committee:

Advisor: Tariq Iqbal

Advisor:

Committee Member: Laura Barnes

Committee Member: Aidong Zhang

Committee Member: Sara Riggs

Committee Member: Dan Bohus

Committee Member:

Committee Member:

Accepted for the School of Engineering and Applied Science:

Jennifer L. West, School of Engineering and Applied Science

August 2023

Appendix A

In real-world embodied data collection, we conduct surveys in two stages: demographic and post-task surveys. In the demographic survey, we intended to gather basic demographic information from participants. This information includes age, gender, the highest level of education, level of experience with robots, and dominant hand (Figure A.1). The survey was administered in person, prior to the participant performing the task.

We conducted a post-task survey to delve deeper into the preferences of participants when guiding a robot. This questionnaire requested that participants disclose their favored method of interaction (Figure A.2). They had the potential to choose from exclusively using spoken directives, solely employing gestures, or utilizing a blend of both verbal instructions and gestures.

This survey aimed to understand better how people would prefer to interact with a robot when they need to reference an object. The three options offered to the participants represented the broad categories of communication techniques (Figure A.2). The first one, 'using only verbal instructions', means that participants would give directions to the robot using language only, without any physical movement or hand signals.

The second option, 'only gestures', implies that the participants would refer to the object using physical gestures like pointing without using any words. This could be a more intuitive way for some people, especially in noisy environments or when a more visual representation is helpful.

The third option was 'a combination of verbal and gestures'. This means the participants would use words and physical gestures to instruct the robot. This could increase the clarity of the instructions and ensure a better understanding of the robot. This combination might be preferred when the tasks are complex or when precision is needed in the instructions.

The participants' preferences could greatly inform how we design robot interaction in the future. If one method is preferred over the others, it could potentially improve robotic systems' effectiveness and user satisfaction.

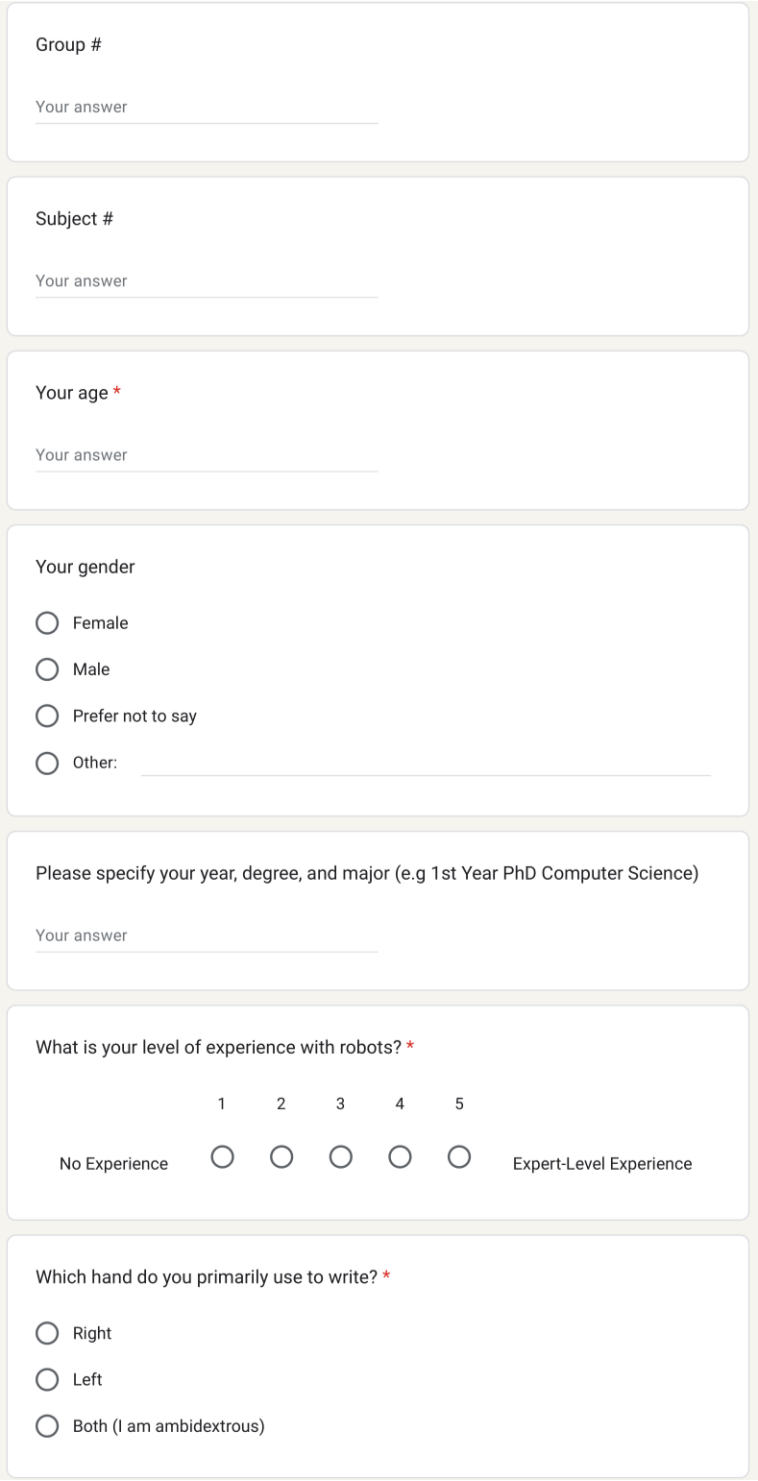


Figure A.1: Demographic Survey

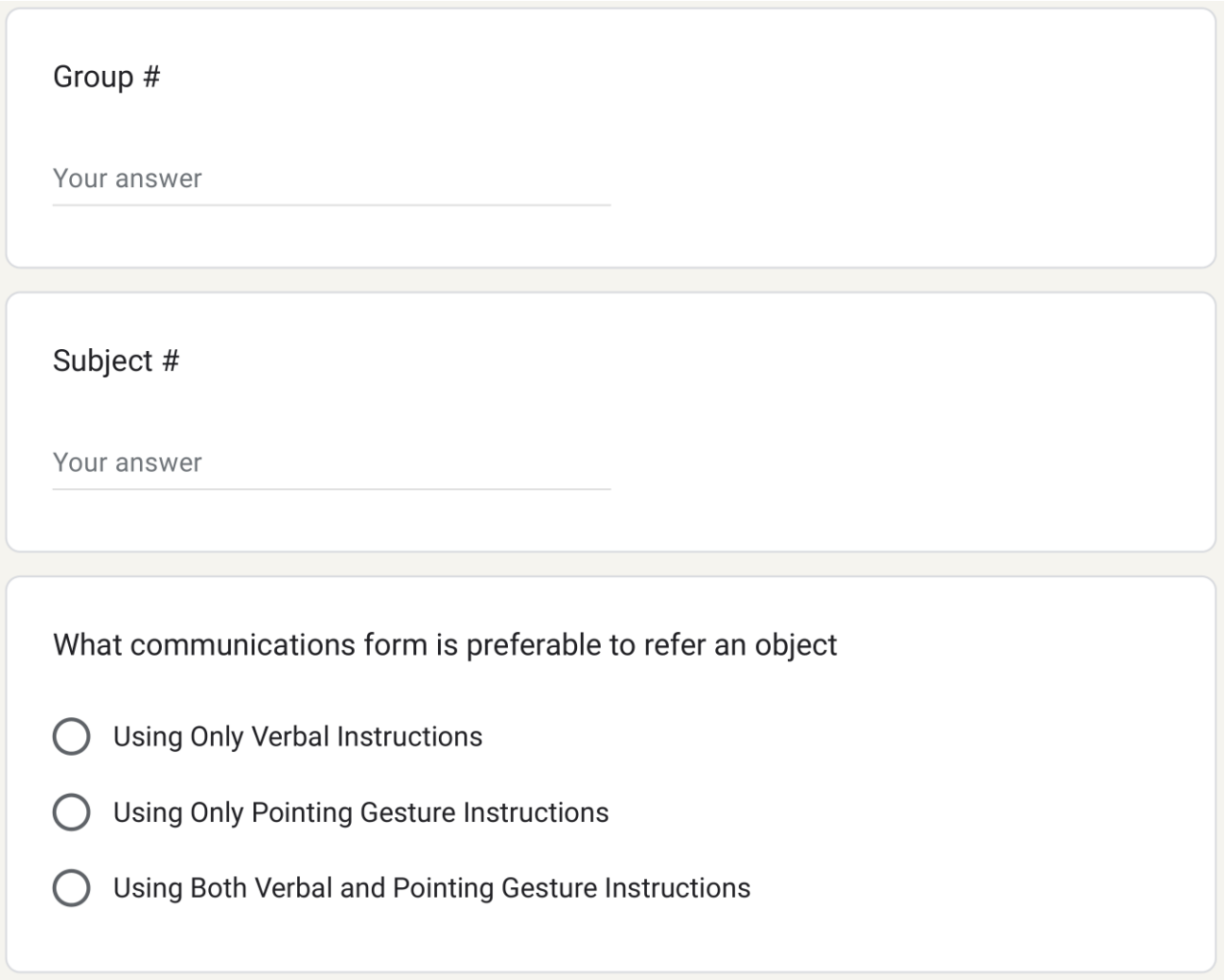


Figure A.2: Post Task Survey