A Data Capture and Gesture Recognition System to Enable Human-Robot Collaboration

A Technical Report submitted to the Department of Systems and Information Engineering

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Evan Smith

Spring 2025

Capstone Project Team Members

Camp Hagood

Sarah Naidu

Aramis Rolly

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature  Date 4/30/25

Evan Smith

Advisor: Tariq Iqbal, Department of Computer Science, Department of Systems and Information Engineering

A Data Capture and Gesture Recognition System to Enable Human-Robot Collaboration

Sarah Naidu, Evan Smith, Camp Hagood, Aramis Rolly, Sujan Sarker, Cory Hayes, and Tariq Iqbal

***Abstract*—Effective human-robot collaboration (HRC) relies on intuitive and reliable communication modalities, particularly in dynamic environments where traditional verbal or wearable sensor-based systems may be unreliable. While gesture-based communication offers a natural and non-intrusive alternative, it remains challenging due to limitations in current recognition systems, such as their dependence on large labeled datasets and lack of adaptability in various environmental conditions. Recent advances in vision-language models (VLMs) have shown promise in video understanding and general reasoning. However, they often lack the domain-specific context required for accurate classification in specialized applications. To address these challenges, we introduce a novel gesture recognition system that leverages a vision-language model (VLM) guided by retrieval-augmented generation (RAG) and chain-of-thought (CoT) prompting to introduce contextual understanding and reasoning. Our system captures upper-body gestures using an Azure Kinect, extracts sampled frames, and classifies them using GPT-4o enhanced by RAG from military gesture documentation and CoT reasoning strategies. Recognized gestures are encoded as ROS 2 messages and transmitted using a publisher-subscriber model to command a mobile robot to execute the corresponding actions. We validate our approach through controlled experiments using seven U.S. Marine Corps (USMC) gestures. The system achieved an accuracy of 80%, an F1 score of 89.9%, and demonstrated effective gesture-to-robot execution. Our results highlight the potential of VLMs for zero-shot gesture classification and robotic control, providing a foundation for robust, scalable, and field-deployable gesture-based HRC systems.**

I. Introduction

Collaborative robotic systems capable of understanding and safely operating alongside humans have the potential to significantly enhance operational effectiveness across various domains such as manufacturing, disaster response, search and rescue, and healthcare assistance [1]–[4]. Effective human-robot collaboration (HRC) relies on the robot's ability not only to anticipate but also to recognize and interpret human direction, and to act upon the perceived action [5]–[8]. In this process, nonverbal cues—such as hand, arm and leg gestures—serve as essential channels through which human intent can be communicated to robots [9], [10]. In particular, gestures are more useful than other cues in dynamic environments with noise. Enabling robots to interpret these signals accurately introduces unique technical challenges related to vision-based perception, motion disambiguation, and response in dynamic conditions.

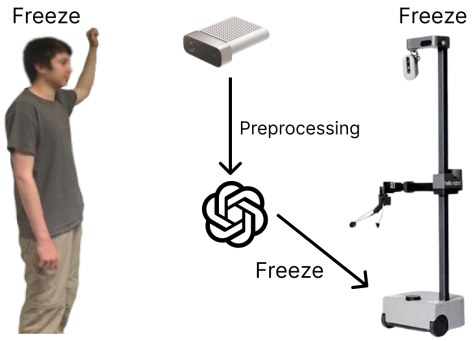

Fig. 1: Example interaction between a robot and person giving a "Freeze" command.

Gesture-based communication is a form of nonverbal communication in which a human performs a gesture to be recognized by a robot to perform an intended action. Establishing a robust gesture recognition system between humans and robots will facilitate intuitive, efficient communication in dynamic, fast-paced environments with limited verbal communication [11]. The goal is for the robot to autonomously perform the task based on the user's gesture.

Current gesture recognition or communication approaches include external sensors, vision-based machine learning, and vision-based deep learning [12]. There are limitations with existing approaches for gesture recognition. External sensor-based approaches can constrain user movement due to their fixed placement and connection wires, which is not ideal for prolonged use [12].

A notable challenge with using gesture-based communication within HRC is the computational complexity behind the algorithms that interpret and classify the gesture for the robot [13]. Vision-based machine learning and deep learning approaches rely on supervised learning with large labeled datasets, which is often infeasible for dynamic field conditions [14]–[16]. Recent progress in vision-language models (VLMs) has introduced promising new pathways for gesture understanding, enabling more flexible, data-efficient systems for interpreting human behavior. Coupled with retrieval-augmented generation (RAG) and chain-of-thought (CoT)

S. Naidu, E. Smith, C. Hagood, A. Rolly, S. Sarker, and T. Iqbal are with the School of Engineering and Applied Science, University of Virginia, USA. Emails: {gjh5qw, ust5gc, nsh4ad, dzq7gw, zzr2hs, tiqbal}@virginia.edu.

C. Hayes is with the US Army Research Laboratory, USA. Email: cory.j.hayes4.civ@army.mil.

prompting, VLMs can generate descriptive interpretations of gestures and classify them effectively without prior training [17]. RAG improves the reliability of responses generated from VLMs by enhancing the contextual information it draws upon from an external knowledge base [17]. Chain-of-thought prompting instructs the VLM to execute reasoning when performing the instructed task [18]. The addition of chain-of-thought prompting has improved the performance of generative models, including VLMs [18]. The introduction and growth of VLMs offer novel approaches for gesture recognition. However, these models often suffer from hallucinations, which occur when generated responses incorporate information about images that were not supplied in the original image input, generate factually incorrect responses, or require greater computational resources to fine-tune the model for specific applications [17], [18].

To address this gap, we develop a gesture recognition system that maximizes the accuracy of identifying upper-body gestures by integrating RAG and CoT prompting. The combination of RAG and CoT improve the VLM's gesture recognition capability by introducing task-specific context and structured reasoning to facilitate informed decisions when generating a gesture label. Alongside our gesture classification pipeline, we have curated robot movements that correspond to each unique gesture (Fig. 3). We demonstrate how to use a gesture description-classification pipeline involving a VLM. The VLM is guided by RAG and chain-of-thought prompting to improve the identification accuracy. The system operates in conjunction with a publisher-subscriber architecture to send commands to the robot once the performed gesture has been classified. Our experimental results validate the system's ability to accurately recognize gestures and act upon them. Ultimately, this work presents a VLM-based system for a gesture-based communication system that can enhance human-robot collaboration in complex environments, laying the groundwork for more intuitive human-robot collaboration.

II. Background and Related Work

A. Modalities for Gesture Recognition

Vision-based and sensor-based approaches are the two primary ways to acquire data for gesture recognition. While vision-based approaches rely on the usage of a camera to capture images or videos of motion, sensor-based approaches use wearable sensors attached to the body to capture motion [19]. Sensor-based approaches are less likely to be affected by environmental conditions, such as lighting and sound [20]. However, sensor-based approaches can cause skin reactions, discomfort with prolonged use, and can restrict the user's movement due to wire connections or the weight of the sensor [21]. Vision-based approaches can provide additional information regarding texture and distance without restricting movement [22]. Although, vision-based approaches require that the gesture be visible to the camera and can be affected by occlusion and lighting [22].

B. Learning-based Gesture Recognition

Traditional machine learning techniques, such as Support Vector Machines (SVMs), K-Means, K-Nearest neighbors (K-NNs), and hidden Markov models (HMMs), are used to classify gestures based on extracted features from captured images or videos. In addition to machine learning, deep learning techniques, such as convolution neural networks (CNNs), recurrent neural networks (RNNs), and artificial neural networks (ANNs) are also used to classify gestures within vision-based approaches. SVMs are effective in binary classification tasks. However, they are computationally expensive and difficult to implement for multi-class classification [23], [24]. k-Means works by performing clustering to group similar gesture features, but does not perform well in the presence of outliers [25]. k-NN classifies gestures by voting based on the nearest neighbor (feature vector). While it is robust method, its computationally expense increases as the size of the dataset increases [26]. HMMs are beneficial for modeling temporal data, but they require predefined states for each gesture, limiting flexibility [27]. While CNNs are more robust in handling spatial features and RNNs are better at handling temporal data, both types of deep networks require large training datasets [28], [29].

C. VLM-based Gesture Recognition

The growth and advancements in VLMs, with models such as CLIP and GPT-4o, have led to their visual and textual understanding capabilities leveraged for gesture recognition tasks [30], [31]. Contrastive Language-Image Pretraining (CLIP) offers robust zero-shot learning, allowing for flexible classification without the need for training data pertaining to the task [30]. Pretrained VLMs, such as GPT-4o, have been leveraged for gesture classification tasks for medical imaging in computationally efficient manner, demonstrating the efficacy of VLMs without the computational intensiveness of fine-tuning [31]. However, VLMs have shown to hallucinate items not present in the initial video or image provided to it and face difficulty in differentiating hallucinated details compared to original details, raising concerns regarding the reliability of VLMs [32].

D. Gesture-based Human-Robot Collaboration

Gesture recognition provides an instinctive form of communication to enable effective human-robot collaboration [33]–[35]. Gesture recognition is most effective in dynamic, fast-paced environments where sensors and voice-based communication are not reliable. Vision-based hand gestures allow for intuitive and expressive communication, which is vital in conveying information for collaborative tasks [36]. Hand gestures have been used to issue commands for remote robotic operations, reducing the dependence on verbal commands and manual operation [37]. Upper-body gesture recognition is ideal in dynamic environments where the distance between the human and robot is further. It is more difficult for a

robot to parse out distinct hand movements compared to arm movements from a distance.

III. Approach

In this section, we discuss the different components of the developed gesture-based human-robot collaboration system (Fig. 2).

A. Gestures

The gestures used in our approach are sourced from the USMC Patrolling document, which includes common upper-body gestures [38]. The particular gestures used from this document are: change direction, freeze, halt, assemble, forward, decrease speed, and disregard previous command (Fig. 3). These gestures were chosen because they are currently in use by the U.S. Military and can be learned easily by the majority of the public. These gestures were also chosen because they can be interpreted as realistic actions that a robot may have to perform during human robot collaboration in a dynamic environment.

B. Gesture Capture and Preprocessing

The initial steps in our classification pipeline are video capture and preprocessing. We use a Microsoft Azure Kinect to collect RGB-D video of an individual performing one of the hand and arm gestures. After the video has been recorded, it is extracted into a series of frames. This is due to GPT-4o's ability to accept only image and text input. We then extract every 15th frame to be sent for classification, maintaining temporal information while reducing inference time.

C. VLM-RAG-based Gesture Classification

To perform gesture classification, we run GPT-4o on the same machine on which the preprocessing step was performed. Once the sampled frames are selected from the video, the model processes the series of images to describe and classify the gesture being performed. Although GPT-4o excels at generating descriptions, it requires domain knowledge to perform domain-specific tasks, such as descriptions or classification. To enhance GPT-4o's contextual understanding, we incorporate RAG to enrich the VLM's descriptions and align it with the performed gesture. RAG is fed into the VLM when it is initially informed by the prompt to generate an accurate description that aligns with one of the seven gestures (Fig. 3). To perform RAG, an external knowledge base—the USMC Patrolling Document—is supplied to the VLM, which then processes each page of the manual containing images and descriptions of hand and arm gestures, and incorporates it into the VLM [39]. The implementation of RAG has shown to improve the performance of generative models, such as VLMs, by incorporating additional contextual information through the form of a knowledge base. This results in less model hallucinations and incorrect responses [17].

After the VLM generates a description informed by RAG for the frames, it is then instructed to classify the description with one gesture label from the seven gesture options. During the classification stage, we employ chain-of-thought prompting to guide the VLM through structured reasoning when classifying the description [18]. Specifically, we instruct the VLM to think through step-by-step before settling on a final classification. The inclusion of CoT leads the VLM to methodically evaluate the generated description with the provided gesture descriptions in the prompt before settling on a classification. The combination of RAG and CoT introduces additional knowledge and reasoning, improving the performance of the gesture classification system.

TABLE I: Gestures and Corresponding Robot Actions

Gesture	**Robot Action**
Forward (F)	Moves forward at constant speed
Halt (H)	Stops immediately
Change Direction (CD)	Reverses direction
Decrease Speed (DS)	Halves current speed
Disregard Prev. Cmd (DPC)	Restores command before the last
Freeze (Fr)	Pauses for 3s, then resumes
Assemble (A)	Moves diagonally right, then left

D. Robot Execution

To enable communication between the gesture classification module and the Stretch 3 mobile robot, we implemented a publisher-subscriber architecture over a shared wireless local area network (WLAN). This setup was chosen because it is modular, scalable, and makes it easy for nearby computing agents to quickly share data with each other.

1) Gesture Classification System: The gesture recognition subsystem operates on a laptop equipped with an Azure Kinect DK camera. The system continuously captures a video stream and uses the Gesture Classification Program's distinct military gestures: forward, halt, change direction, decrease speed, disregard previous command, freeze, and assemble (Table I, Fig. 3). Upon successful recognition, each gesture is encoded as a plain-text command string (e.g., `decrease_speed`). These commands are intended to map directly to predefined robot behaviors.

2) Publisher-Subscriber Architecture: The architecture is structured as a publisher-subscriber model using a TCP/IP socket-based protocol over a local wireless network. The gesture recognition laptop serves as the client and publisher of command messages, while the Stretch robot hosts a lightweight server that acts as the subscriber and command interpreter. The publisher and subscriber are configured to monitor for gesture events in real time. When a gesture is recognized, the system transmits a text-based command string to the robot using a socket connection. On the robot's end, a subscriber program runs continuously, listening for incoming socket messages and interpreting them to determine the intended gesture. This subscriber is set up as a ROS2 node on the robot's onboard computer and works seamlessly with Hello Robot's Stretch 3 control system. When a gesture command comes in, the robot uses ROS2 service calls or action interfaces to convert it to a behavior.

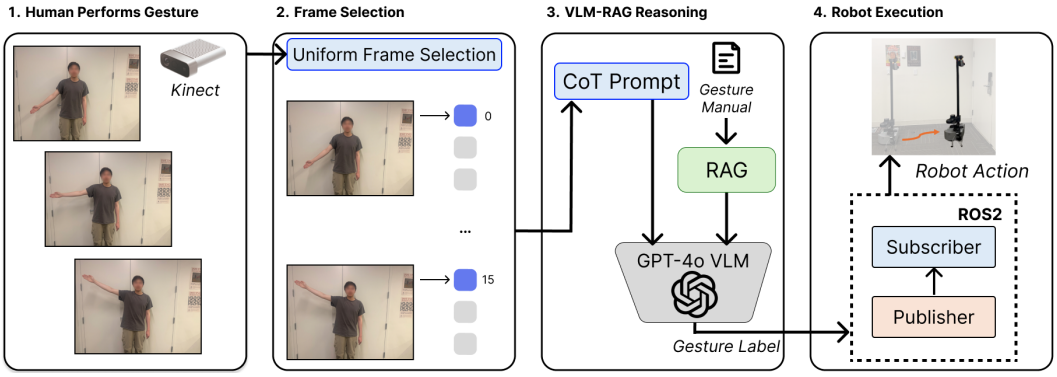

Fig. 2: Overview of the system components: 1. Kinect records the participant's gesture; 2. every 15th frame is sampled and sent to the VLM; 3. RAG provides context and CoT guides gesture classification 4. ROS 2 publisher-subscriber architecture triggers corresponding robot action.

3) Action Execution on the Robot: Each gesture command triggers a specific robot behavior, implemented using ROS 2 control primitives and Hello Robot's Stretch Body Python API. For example, the `forward` command moves the robot forward at a preset speed until a `halt` or `freeze` command is received, while `change_direction` reverses its motion. This hardware-agnostic setup supports flexible adaptation to new gesture models or robotic platforms with minimal modifications.

IV. Exploratory Study

We seek to validate our gesture recognition system through a preliminary study. We sought to observe the accuracy of the gesture classification pipeline, the delay induced by the gesture classification process, and the subsequent movement of the robot.

A. Study Procedure

Participants performed seven USMC gestures—assemble, change direction, decrease speed, disregard the previous command, forward, freeze, and halt—twice each (Fig. 3). They first completed a training session where they watched prerecorded videos demonstrating each gesture, and we informed them of the corresponding robot action. Following an errorless practice run, participants replicated the gestures in front of an Azure Kinect to trigger the robot's actions (Fig. 1). Recordings occurred under bright lighting, with participants standing before a white background. Once a gesture began, the experimenter initiated the Kinect recording, which continued until the gesture ended. After recording, we uniformly sampled a subset of frames and sent those to a VLM to classify the performed gesture. The VLM outputs the most probable gesture label, and we send it to the publisher program that publishes the detected gesture. The subscriber program receives the gesture label and triggers appropriate robotic action. Participants proceeded to the next gesture after the robot completed its movement. After the study, we briefed the participants on the study's purpose.

Fig. 3: U.S. Marine Corps gestures [39] used in the study.

B. Participants

4 adults participated in the study (75% male ($n = 3$), 25% female ($n = 1$); mean age = 22.19, SD = 0.17), all with engineering backgrounds. 1 participant reported limited experience with robots. None had a military background or prior familiarity with the USMC gestures (Fig. 3).

V. Results

Across all of the gestures, our approach achieved an 80% classification accuracy. It also achieved a weighted F1 score of 89.9%, a weighted precision of 96.4%, and a weighted recall of 86.5% (Fig. II). 4 unknown classifications out of 56 total performed gesture movements were omitted from the calculations for weighted F1, precision, and recall. Unknown was not a distinct gesture category provided to the VLM in the prompt. However, it was outputted by the VLM to indicate its inability to classify the gesture. Although each gesture was performed a total of 8 times, in certain instances

TABLE II: Overall Performance of VLM

Metric	F1 Score	Precision	Recall	Accuracy
VLM Performance	89.9%	96.4%	86.5%	80%

TABLE III: Gesture Classification Result

Gesture	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Assemble (A)	12.5	14.3	100.0	25.0
Change (C)	100.0	100.0	100.0	100.0
Decrease (DS)	100.0	100.0	72.7	84.2
Disregard (DPC)	100.0	100.0	100.0	100.0
Forward (F)	87.5	100.0	70.0	82.4
Freeze (Fr)	62.5	83.3	83.3	83.3
Halt (H)	100.0	100.0	100.0	100.0

some gestures were unable to be classified by the VLM. The assemble and forward gestures were classified as unknown once, while the freeze gesture was classified as unknown twice. Additionally, the assemble gesture was misclassified as the forward gesture 3 times (Fig. 5).

The correlation between video duration and inference delay is 0.927. The assemble gesture had the longest average video duration (5.95 seconds), longest average inference delay (18.3 seconds), and the lowest F1 score of 25%. The freeze gesture had the shortest average video duration (2.33 seconds), shortest average inference delay (7.88 seconds), and an F1 score of 83%. The change, disregard previous command, and halt gestures had the highset F1 scores of 100%. Their respective average video durations (in seconds) are 2.72, 2.46, and 2.67 and their respective average inference delays (in seconds) are 9.51, 9.11, and 8.60 (Table III).

VI. Discussion

A. Summary of Key Findings

The study highlights a strong correlation between video duration and inference delay, which is attributed to the uniform selection of frames. As the video duration increases, more frames are sent to the VLM, resulting in longer inference times (Fig. 4. This finding suggests that implementing key frame extraction could significantly reduce inference time for prolonged gestures by minimizing the number of frames analyzed by the VLM. The reduction in inference time would improve the robot's responsiveness, allowing it to act sooner after a gesture is performed.

Accurately classified gestures were characterized by distinct arm position and movement, even when analyzed using

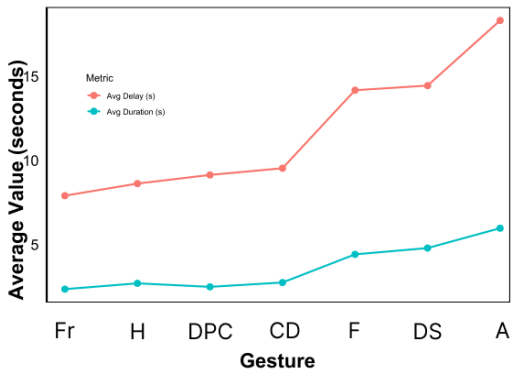

Fig. 4: Average Delay (seconds) and Average Duration (seconds) for each gesture.

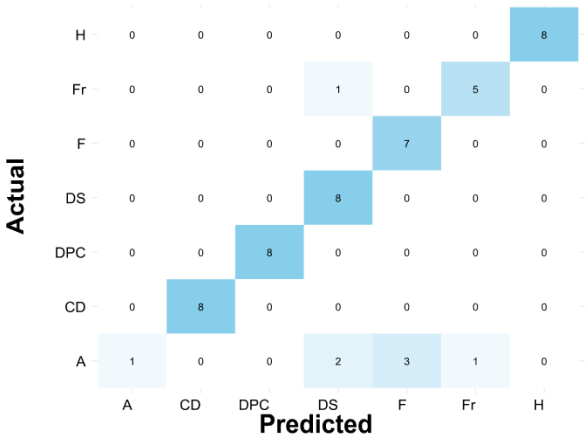

Fig. 5: Confusion matrix displaying the number of correct gesture classifications for each trial.

only every 15th frame of the recordings. This distinctiveness facilitated their correct classification relative to other gestures, except for the assemble gesture.

The low classification accuracy for assemble may stem from its movement similarities with other gestures when analyzed frame-by-frame. Its prompt description may also lack sufficient detail, contributing to misclassification.

B. Implications

Our preliminary results show that careful prompt engineering can significantly improve gesture classification accuracy without extensive labeled training data. By providing multiple example descriptions and instructing the VLM to dissect gestures frame-by-frame, we have guided the models toward more accurate interpretations of complex hand signals.

C. Future Research

Our work is a foundational project for future gesture recognition-based HRC. We identify several future directions for the proposed gesture classification system. The incorporation of key frame extraction techniques is expected to enhance classification accuracy while simultaneously reducing processing times for gestures that have a long movement duration, allowing the system to focus on critical moments within a gesture. Additionally, implementing a skeleton overlay on top of the subject may improve the VLM's ability to recognize gestures, emphasizing critical movements and arm positions, which could improve recognition accuracy.

Further validation through expanded studies is essential. Involving a larger participant base and a more diverse range of upper-body gestures would improve the validity of our findings. Moreover, conducting field tests in varied environments–such as forests, mountains, and deserts–with robots specifically designed for these conditions will provide crucial insights into the practical application of this work.

VII. Conclusion

In this work, we developed a system to recognize complex, domain-specific gestures (such as USMC hand and arm signals). The system uses a VLM guided by retrieval-augmented generation and chain of thought to control a robot. Our findings indicate that this approach has the potential to accurately classify gestures through zero-shot classification to enable a robust, field-deployable HRC system. As part of our future work, we will incorporate more modalities, such as skeleton overlay, depth, and speech, and evaluate the system across various gesture domains. Such a system can enhance HRC in industries like manufacturing, healthcare, and education, where intuitive, gesture-based interaction can streamline operations.

Acknowledgment

The team would like to thank Haley Green for her assistance in working with the robot.

References

[1] S. Sarker, M. T. Arafat, A. Lameesa, M. Afrin, R. Mahmud, M. A. Razzaque, and T. Iqbal, "Fold: Fog-dew infrastructure-aided optimal workload distribution for cloud robotic operations," *Internet of Things*, 2024.

[2] S. Sarker, H. N. Green, M. S. Yasar, and T. Iqbal, "Cohrt: A collaboration system for human-robot teamwork," *arXiv preprint arXiv:2410.08504*, 2024.

[3] M. S. Yasar, M. M. Islam, and T. Iqbal, "Posetron: Enabling close-proximity human-robot collaboration through multi-human motion prediction," in *HRI*, 2024.

[4] S. Ali, H. N. Green, and T. Iqbal, "What am i? evaluating the effect of language fluency and task competency on the perception of a social robot," in *ROMAN*. IEEE, 2024.

[5] M. S. Yasar and T. Iqbal, "Coral: Continual representation learning for overcoming catastrophic forgetting," in *AAMAS*, 2023.

[6] M. S. Yasar, M. M. Islam, and T. Iqbal, "Imprint: Interactional dynamics-aware motion prediction in teams using multimodal context," *THRI*, 2024.

[7] M. M. Islam, A. Gladstone, R. Islam, and T. Iqbal, "Eqa-mx: Embodied question answering using multimodal expression," in *ICLR*, 2023.

[8] M. S. Yasar and T. Iqbal, "Vader: Vector-quantized generative adversarial network for motion prediction," in *IROS*. IEEE, 2023.

[9] M. M. Islam, A. Gladstone, and T. Iqbal, "Patron: Perspective-aware multitask model for referring expression grounding using embodied multimodal cues," in *AAAI*, 2023.

[10] M. S. Yasar and T. Iqbal, "Robots that can anticipate and learn in human-robot teams," in *HRI*. IEEE, 2022.

[11] O. Kobzarev, A. Lykov, and D. Tsetserukou, "Gestllm: Advanced hand gesture interpretation via large language models for human-robot interaction," 2025.

[12] B. K. Chakraborty, D. Sarma, M. K. Bhuyan, and K. F. MacDorman, "Review of constraints on vision-based gesture recognition for human–computer interaction," *IET Computer Vision*, 2018.

[13] A. Bonarini, "Communication in human-robot interaction," *Current Robotics Reports*, 2020.

[14] S. Samyoun, M. M. Islam, T. Iqbal, and J. Stankovic, "M3sense: Affect-agnostic multitask representation learning using multimodal wearable sensors," July 2022.

[15] S. Hasan, M. S. Yasar, and T. Iqbal, "M2rl: A multimodal multi-interface dataset for robot learning from human demonstrations," in *Proceedings of the 26th International Conference on Multimodal Interaction*. New York, NY, USA: Association for Computing Machinery, 2024.

[16] T. Iqbal, S. Li, C. Fourie, B. Hayes, and J. A. Shah, "Fast online segmentation of activities from partial trajectories," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019.

[17] H. Yu, A. Gan, K. Zhang, S. Tong, Q. Liu, and Z. Liu, "Evaluation of retrieval-augmented generation: A survey," in *CCF Conference on Big Data*. Springer, 2024.

[18] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, 2022.

[19] D. Sarma and M. K. Bhuyan, "Methods, databases and recent advancement of vision-based hand gesture recognition for hci systems: A review," *SN Computer Science*, 2021.

[20] T. Zhao, J. Liu, Y. Wang, H. Liu, and Y. Chen, "Ppg-based finger-level gesture recognition leveraging wearables," in *INFOCOM*. IEEE, 2018.

[21] M. Oudah, A. Al-Naji, and J. Chahl, "Hand gesture recognition based on computer vision: a review of techniques," *journal of Imaging*, 2020.

[22] K. Gao, H. Zhang, X. Liu, X. Wang, L. Xie, B. Ji, Y. Yan, and E. Yin, "Challenges and solutions for vision-based hand gesture interpretation: A review," *Computer Vision and Image Understanding*, 2024.

[23] N. H. Dardas and N. D. Georganas, "Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques," *IEEE Transactions on Instrumentation and Measurement*, 2011.

[24] L. Liu, J. Xing, H. Ai, and X. Ruan, "Hand posture recognition using finger geometric feature," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. IEEE, 2012.

[25] D. K. Ghosh and S. Ari, "A static hand gesture recognition algorithm using k-mean based radial basis function neural network," in *ICICS*. IEEE, 2011.

[26] T. Marasović and V. Papić, "Feature weighted nearest neighbour classification for accelerometer-based gesture recognition," in *SoftCOM*. IEEE, 2012.

[27] H.-K. Lee and J. Kim, "An hmm-based threshold model approach for gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1999.

[28] Y. Kim and B. Toomajian, "Hand gesture recognition using micro-doppler signatures with convolutional neural network," *IEEE Access*, 2016.

[29] X. Chai, Z. Liu, F. Yin, Z. Liu, and X. Chen, "Two streams recurrent neural networks for large-scale continuous gesture recognition," in *ICPR*. IEEE, 2016.

[30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[31] K. Bimbraw, Y. Wang, J. Liu, and T. Koike-Akino, "Gpt sonograpy: Hand gesture decoding from forearm ultrasound images via vlm," 2024.

[32] N. Jiang, A. Kachinthaya, S. Petryk, and Y. Gandelsman, "Interpreting and editing vision-language representations to mitigate hallucinations," *arXiv preprint arXiv:2410.02762*, 2024.

[33] M. M. Islam, R. Mirzaiee, A. Gladstone, H. Green, and T. Iqbal, "Caesar: An embodied simulator for generating multimodal referring expression datasets," in *NeurIPS*, 2022.

[34] T. Iqbal and L. D. Riek, "Temporal anticipation and adaptation methods for fluent human-robot teaming," in *ICRA*. IEEE, 2021.

[35] M. S. Yasar and T. Iqbal, "Improving human motion prediction through continual learning," *arXiv preprint arXiv:2107.00544*, 2021.

[36] H. Liu and L. Wang, "Latest developments of gesture recognition for human–robot collaboration," in *Advanced Human-Robot Collaboration in Manufacturing*. Springer, 2021.

[37] N. Mendes, J. Ferrer, J. Vitorino, M. Safeea, and P. Neto, "Human behavior and hand gesture classification for smart human-robot interaction," *Procedia Manufacturing*, 2017.

[38] "Land Navigation," U.S. Marine Corps, n.d.

[39] "FMST 205," U.S. Marine Corps, n.d.