One-Shot 3D Object-to-Object Affordance Grounding with Semantic Feature Field for Generalizable Robotic Manipulation

---

A

Thesis

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

---

in partial fulfillment

of the requirements for the degree

Master of Science

by

Tongxuan Tian

May 2025

APPROVAL SHEET

This

Thesis

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

Master of Science

Author: Tongxuan Tian

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

Advisor: Yen-Ling Kuo

Advisor:

Committee Member: Zezhou Cheng

Committee Member: Henry Kautz

Committee Member:

Committee Member:

Committee Member:

Committee Member:

Accepted for the School of Engineering and Applied Science:

Jennifer L. West, School of Engineering and Applied Science

May 2025

Abstract

Affordance grounding—the identification of functional properties that indicate how objects can be manipulated—is fundamental to embodied intelligence and robotic manipulation. While previous research has made significant progress in single-object affordance prediction, it has largely overlooked the critical reality that most real-world tasks involve interactions between multiple objects. This thesis addresses the challenge of object-to-object (O2O) affordance grounding in 3D space under limited data constraints.

We introduce O^{3}Afford, a novel one-shot learning framework for object-to-object affordance grounding that leverages 3D semantic fields distilled from vision foundation models (VFMs). Our key insight is that by combining the rich semantic understanding capabilities of VFMs with the geometric information captured in 3D point clouds, we can enable effective generalization to unseen objects with minimal supervision. The framework projects multi-view features from vision foundation models onto point clouds of interacting objects, creating semantically-enriched representations that capture part-awareness critical for affordance prediction.

At the core of our approach is a transformer-based affordance decoder that explicitly models geometric relationships and semantic features between objects, considering how each object's geometry influences potential interaction regions on the other. This approach captures the geometry context of object-to-object affordances while maintaining awareness of the distinct functional roles in interactions such as pouring, cutting, and plugging.

We further integrate our affordance representations with large language models to enhance fine-grained spatial understanding for downstream tasks. Experimental evaluations demonstrate that O^{3}Afford significantly outperforms existing methods in both affordance prediction accuracy and generalization capabilities across unseen object instances, partial observation, and novel categories. Through experiments in both simulation and real-world environments, we validate that our approach facilitates more effective manipulation planning for complex interactive tasks.

This work bridges a critical gap in affordance learning by enabling robots to understand not just how humans interact with individual objects, but how objects functionally interact with each other—a fundamental capability for advanced robotic manipulation in everyday environments.

Acknowledgements

I would like to express my deepest gratitude to all the individuals who have supported and guided me throughout my Master's journey. First and foremost, I extend my heartfelt appreciation to my parents for their unwavering support and generous financial assistance, which allowed me to pursue my studies and research in the United States. Their belief in my potential and continuous encouragement have been invaluable, providing me with the strength and determination to follow my dreams.

My sincere gratitude also goes to my advisor, Professor Yen-Ling Kuo, who has profoundly shaped my academic and professional journey. Her mentorship introduced me to the fascinating world of robotics, guiding me seamlessly from my initial focus in computer vision to a new, exciting, and deeply fulfilling research area. Her patience, insight, and unwavering support have been crucial in helping me discover my true passion. The experiences and knowledge I gained under her guidance will undoubtedly shape my future endeavors.

I am equally thankful to the PhD students in our lab for their generous assistance, thoughtful advice, and camaraderie. Their guidance was instrumental in overcoming many technical challenges and significantly enriched my research experience. The collaborative and supportive lab environment they fostered allowed me to grow both intellectually and personally.

I would also like to sincerely thank the members of my thesis committee, Professor Zezhou Cheng and Professor Henry, for their insightful feedback, constructive criticism, and valuable guidance. Their expertise and thoughtful suggestions significantly enhanced the quality of my work.

Furthermore, I want to extend my appreciation to everyone in the Computer Science Department, especially Jai and Marion, for their invaluable assistance and support. Their genuine care and willingness to help made the department feel like a home away from home. Indeed, the Computer Science Department at our university stands out as one of the most compassionate and supportive communities I have ever been part of. The human-centered culture here made my academic experience enjoyable and deeply rewarding.

Finally, I wish to acknowledge all my friends and peers who offered emotional support and friendship throughout my graduate studies. Their presence made the highs more joyful and the challenges much more manageable.

Thank you all for making this journey unforgettable.

Table of Contents

Abstract **iv**

Acknowledgements **v**

List of Tables **viii**

List of Figures **ix**

1 Introduction **1**

1.1 What is Affordance Grounding? 1

1.2 Object-to-object Affordance Grounding 2

1.3 Vision Foundation Models 2

1.4 Motivation 3

1.5 Research Goal and Contribution 3

1.6 Thesis Overview 5

2 Related Work **6**

2.1 Affordance Grounding 6

2.2 Few-shot Learning with Foundation Models 6

2.3 Affordance-based Robotic Manipulation 7

3 Problem Formulation **9**

4 Methodology **11**

4.1 Overview 11

4.2 Semantic Point Cloud Construction 11

4.3 One-Shot Affordance Grounding 12

4.3.1 Point Cloud Encoder 12

4.3.2 Type Embedding Mechanism . . . 13
4.3.3 Cross-Attention Module . . . 13
4.3.4 Projection Head and Point-wise Prediction . . . 13
4.3.5 Training and Optimization . . . 14
4.4 Affordance-Based Planning with LLMs . . . 14
4.4.1 Affordance Region Identification . . . 14
4.4.2 Constraint Generation with LLMs . . . 15
4.4.3 Transformation Optimization . . . 17

5 Experiments **18**
5.1 Experiment Setup . . . 18
5.2 Affordance Grounding . . . 19
5.2.1 Comparison Results . . . 19
5.2.2 Generalization Experiments . . . 20
5.3 Affordance-based Manipulation . . . 22

6 Conclusion and Future Work **24**
6.1 Conclusion . . . 24
6.2 Future Work . . . 25

References **26**

A Prompt Template **30**

List of Tables

5.1 **Quantitative comparison on object-to-object affordance grounding.** 20

5.2 **Success rate comparison on robotic manipulation tasks.** 23

List of Figures

4.1 Framework pipeline. 12
4.2 LLM Planning Pipeline. 16
4.3 Example constraint function used for pouring interactions. 16

5.1 Affordance Qualitative Results. 21
5.2 Qualitative Results of Partial Point Cloud. 22
5.3 Qualitative Results of Unseen Object Category. 22
5.4 Manipulation Qualitative Results. 23

Chapter 1

Introduction

In robotics, understanding and interpreting the environment is essential for enabling robots to interact effectively with objects and perform meaningful tasks. Traditionally, robotic systems have relied heavily on explicit object models and predefined manipulation strategies. However, in real-world scenarios, objects often exhibit varying shapes, sizes, and physical properties, making it challenging for robots to apply fixed strategies. To overcome this limitation, researchers have increasingly turned towards concepts inspired by human perception, specifically, how humans intuitively understand and interact with their environment. Central to this shift is the concept of affordance, which describes the actionable possibilities that an environment or an object inherently provides to an agent.

1.1 What is Affordance Grounding?

Affordance grounding aims to identify the functional properties of objects or environments that indicate potential interactions, effectively communicating how objects can be manipulated and interacted with in meaningful ways. Accurate prediction of affordance maps for objects can significantly enhance numerous downstream tasks, including human-computer interaction [1], visual understanding [2], and robotic manipulation [3]. Several studies have investigated affordance prediction in 2D pixel space [4, 5, 6, 7, 8], focusing on predicting functional maps from input images with language conditions that describe the intended tasks. While effective, this approach may limit potential generalization capabilities due to its disregard of the geometric information inherent in object shapes, which is difficult to accurately estimate in 2D space. Our work investigates affordance grounding in the 3D space.

1.2 Object-to-object Affordance Grounding

Although many works [9] have explored 3D object affordance grounding using object point clouds, they predominantly focus on single-object affordance prediction with the assumption that the affordance pertains solely to human interaction. This approach neglects the reality that many daily tasks involve object-to-object interactions. For instance, a pouring task typically requires two objects: a source container and a target container. Such an assumption limits the potential generalization ability to broader scenarios.

Object-to-object affordance grounding is essential for enabling robots to handle complex manipulation tasks that reflect real-world scenarios. In everyday life, tasks like cooking, assembling furniture, or tool usage inherently involve interactions between multiple objects. Accurate understanding of these interactions is crucial for robots to perform tasks autonomously and safely. Effective grounding of object-to-object affordances enhances robotic manipulation capabilities by providing contextual understanding of how different objects can functionally relate to one another, thus significantly improving robots' adaptability and efficiency in dynamic and unstructured environments.

However, a major challenge in object-to-object affordance grounding is the scarcity of annotated datasets. Unlike single-object affordance grounding, annotating data for object-to-object interactions is inherently more complex and time-consuming, as it requires specifying precise relational information between multiple objects. O2O-Afford [10] addressed this problem in an annotation-free manner through automatically extracting contacts in simulation, but this approach remains limited to simple affordances such as placing and fitting, extending to more complex ones require carefully craft the object interactions, which is as hard as solving O2O-afford. Additionally, many important interactions, like pouring liquids or cutting objects, involve sophisticated physical dynamics and subtle interactions, making simulation-based annotation approaches less effective. Our work aims to tackle these challenges by developing methodologies capable of inferring complex affordances such as pouring and cutting, which are currently unattainable by existing annotation-free techniques yet crucial for practical robotic manipulation applications.

1.3 Vision Foundation Models

Vision Foundation Models (VFMs) have their roots deeply embedded in the field of self-supervised learning. Early research on self-supervised learning focused primarily on designing pretext tasks, such as predicting image rotations, solving jigsaw puzzles, and image inpainting, to facilitate the

learning of meaningful representations without human annotations. As computational resources and data availability increased, more advanced self-supervised approaches emerged, notably contrastive learning frameworks such as SimCLR [11] and MoCo [12], which effectively leveraged large-scale unlabeled data to produce robust visual representations.

Following this, vision transformers emerged, revolutionizing the field by demonstrating superior scalability and generalization capabilities compared to traditional convolutional neural networks. Vision transformers, when pre-trained with self-supervised objectives on vast datasets, exhibited remarkable transferability to downstream tasks without extensive fine-tuning. This paradigm shift marked the rise of modern vision foundation models, which integrate large-scale training, transformer architectures, and self-supervised objectives to create universally applicable vision models capable of achieving state-of-the-art results across diverse vision tasks. Our exploration into VFMs for 3D affordance grounding builds upon this rich history, aiming to harness their powerful representational capacities to address complex, geometry-rich interactions encountered in object-to-object affordance grounding.

1.4 Motivation

We seek to develop a solution for generalizable object-to-object affordance grounding with minimal supervision. Recent advances in VFMs [13, 14, 15] have demonstrated impressive zero-shot generalization capabilities across various vision tasks. Being pre-trained on internet-scale datasets, these vision foundation models (VFMs) have been equipped with generalizable semantic understanding [16] and part awareness [17]. A series of works have explored leveraging the capabilities of VFMs for few-shot downstream vision tasks such as segmentation [18], detection [19], and visual correspondence [20]. Inspired by a prior work [4] in VFMs for 2D affordance, we further investigate VFMs for 3D affordance with the hypothesis that point clouds provide richer geometric information, thus enabling generalizable capabilities across different viewpoints, unseen object instances, and even entirely novel object categories.

1.5 Research Goal and Contribution

We introduce O^{3}Afford, a **O**ne-shot **O**bject-to-**O**bject **Affordance** Grounding framework with 3D feature fields distilled from VFMs. Specifically, we leverage pre-trained DINOv2 [13] to extract rich semantic features from multi-view observations. These semantic features are projected onto

the point clouds of both the source and target objects involved in manipulation tasks. The resulting enriched point clouds encapsulate part-aware semantic information, facilitating robust affordance inference. Our approach introduces a novel bi-directional affordance discovery module, which explicitly accounts for geometric and semantic contexts from both objects reciprocally. This bidirectional approach effectively captures nuanced interactions, enhancing the accuracy and generalizability of affordance predictions. Furthermore, we integrate our affordance grounding module with large language models (LLMs), where the predicted affordance maps serve as an interpretable spatial representation used by LLMs in their reasoning process. By implementing computational routines that load both object point clouds and their associated affordance values, the LLMs can perform enhanced spatial reasoning and planning. This integration significantly improves spatial understanding and enables comprehensive reasoning capabilities in robotic manipulation scenarios that surpass methods relying solely on visual or geometric inputs.

Our evaluation encompasses two primary aspects. First, we rigorously compare our affordance grounding pipeline against prior methods on various challenging object-to-object affordance tasks, demonstrating substantial improvement in generalization across diverse object instances, unseen categories, and varying viewpoints. Second, we validate the practical effectiveness and applicability of our pipeline in robotic manipulation scenarios by integrating it with LLMs for enhanced planning and policy training using our affordance representations. Extensive experiments conducted in both simulation environments and real-world robotic setups confirm the potential of our proposed approach in enabling robots to successfully perform a wide array of tasks that require sophisticated object interactions.

In summary, the key contributions of our proposed O^3Afford are three-fold:

- We propose a novel one-shot object-to-object affordance grounding framework, O^3Afford, which leverages 3D semantic features distilled from vision foundation models for effective affordance prediction.
- We introduce a bi-directional affordance discovery module that explicitly captures mutual geometric and semantic context between interacting objects, significantly enhancing generalization and prediction accuracy.
- We comprehensively evaluate our framework against state-of-the-art methods, demonstrating superior generalization and practical effectiveness in both simulated and real-world robotic manipulation tasks, highlighting the broad applicability and robustness of our approach.

1.6 Thesis Overview

We structure the remainder of this thesis as follows: Chapter **II** presents a comprehensive overview of related work, including affordance grounding techniques, few-shot learning methodologies utilizing vision foundation models, and affordance-based robotic manipulation approaches. Chapter **III** demonstrates the formulation of our method. Chapter **IV** provides an in-depth explanation of our proposed methodology, detailing our novel affordance grounding framework, its key modules, and the integration with vision-language models. Chapter **V** presents our systematic evaluation strategy, experimental setups, and extensive analyses, highlighting the robustness, generalization, and practical efficacy of our approach.

Chapter 2

Related Work

2.1 Affordance Grounding

Affordance grounding has attracted significant attention due to its pivotal role in enabling robots to interact effectively with their environment. Many studies predict a 2D affordance map for objects under specific application scenarios or language conditions [6, 8, 7, 5]. These 2D affordances can provide valuable guidance for robotic manipulation tasks by highlighting actionable regions [21, 22]. However, direct robot manipulation tasks inherently require a deeper understanding of the objects' geometry, necessitating richer, three-dimensional affordance representations.

Several recent studies have addressed this need by exploring 3D affordances, typically focusing on single-object scenarios matched with specific instructions or actions [23, 9, 24]. Despite their progress, these methods often fail to generalize to tasks involving interactions between multiple objects. To address this limitation, our method uniquely predicts 3D affordances for pairs of objects, enabling direct manipulation by leveraging spatial relationships and geometric context.

2.2 Few-shot Learning with Foundation Models

Foundation models have revolutionized few-shot learning paradigms by demonstrating remarkable capabilities in generalization and adaptation to novel tasks with limited training data [25, 26]. By pretraining on extensive, internet-scale datasets, these models acquire comprehensive semantic knowledge and common-sense reasoning abilities [14, 13, 27, 28], facilitating rapid adaptation through minimal demonstrations.

Recent works have further explored leveraging these capabilities in specialized vision tasks. For instance, ZegClip [18] extends CLIP's zero-shot capabilities to pixel-level predictions, effectively enabling zero-shot image segmentation. Similarly, [19] harnessed DINOv2's powerful representation capabilities, combined with large language models (LLMs), achieving robust few-shot object detection. Additionally, latent diffusion models have been utilized to enhance few-shot semantic segmentation tasks [29, 30, 31].

In the context of 3D data, several studies have pioneered pretraining on large-scale point cloud datasets, yielding encouraging results in tasks such as one-shot part segmentation and classification [32, 33]. Despite these advancements, the gap between 2D image-based foundation models and 3D domain-specific models remains substantial due to the limited availability of large-scale 3D datasets. Building upon insights from prior systematic studies like [4], which investigated optimal vision foundation models for one-shot 2D affordance grounding, we extend these methodologies into 3D. Our approach significantly improves geometric generalization and viewpoint invariance, addressing critical challenges in real-world robotic applications.

2.3 Affordance-based Robotic Manipulation

Affordance-based robotic manipulation has become increasingly prominent as robots are expected to perform complex interactions in dynamic environments. Previous works have explored diverse affordance frameworks to enhance robotic capabilities. For example, [34] identifies specific affordance locations for executing predetermined actions and trajectories, facilitating task-oriented manipulations. In [35], predicted keypoints are utilized to simplify subsequent manipulation planning and execution.

Recent advances have also integrated affordance predictions with powerful language models, enabling enhanced semantic and spatial reasoning. [36] demonstrates how large language models (LLMs) can predict affordances, subsequently guiding effective motion planning. Furthermore, affordance prediction has proven beneficial for policy training in reinforcement learning frameworks. [37] leverages detected affordance regions to streamline the reinforcement learning process, significantly improving training efficiency and task performance. Likewise, [38] integrates language goal grounding with affordance predictions to optimize policy training, enhancing robot performance across diverse scenarios.

Our work contributes to this rich body of research by focusing explicitly on object-to-object affordances. We utilize comprehensive 3D geometric reasoning to infer actionable affordances between

interacting object pairs directly. By employing these affordances to guide robotic motion planning, our framework achieves highly efficient and precise manipulations suitable for complex and realistic tasks.

Chapter 3

Problem Formulation

We formulate object-to-object affordance grounding as a problem of predicting functional interaction regions between two object point clouds. Given a source object point cloud $P_s \in R^{N_s \times (3+n)}$ and a target object point cloud $P_t \in R^{N_t \times (3+n)}$, where N_s and N_t represent the number of points in each cloud respectively, our goal is to predict affordance maps $A_s \in [0,1]^{N_s}$ and $A_t \in [0,1]^{N_t}$ that indicate the likelihood of interaction at each point. Each point is represented by its 3D coordinates (x, y, z) and an n-dimensional semantic feature vector extracted from vision foundation models.

Unlike previous approaches that require extensive training data across multiple instances for each affordance type, we tackle a more challenging one-shot training setting. Specifically, the training set consists of a set of K interacting object pairs, each corresponding to a distinct affordance type:

$$\mathcal{D}_{\text{train}} = \{(P_s^{(i)}, P_t^{(i)}, A_s^{(i)}, A_t^{(i)})\}_{i=1}^{K} \tag{3.1}$$

where each $(P_s^{(i)}, P_t^{(i)})$ represents a unique object pair exhibiting the i-th type of affordance, and each affordance type appears only once in the training set.

At test time, the model is evaluated on novel object pairs exhibiting the same set of affordance types, but with entirely unseen objects and geometries:

$$\mathcal{D}_{\text{test}} = \{(P_s^{(j)}, P_t^{(j)})\}_{j=1}^{K'} \tag{3.2}$$

where K' denotes the number of test pairs, possibly different from K, and the goal is to predict the corresponding affordance maps $(A_s^{(j)}, A_t^{(j)})$.

Formally, our model f_θ with parameters θ maps a pair of input point clouds to their respective affordance maps:

$$f_\theta : (P_s, P_t) \mapsto (A_s, A_t) \tag{3.3}$$

Chapter 4

Methodology

4.1 Overview

Our pipeline consists of three components. First, we construct 3D consistent semantic features from DINOv2 for object point clouds. Subsequently, our affordance grounding module takes these semantically-enriched point clouds as input and predicts the corresponding affordance maps. Last, we leverage LLM for constraint function generation, which will be optimized during planning.

4.2 Semantic Point Cloud Construction

To construct our 3D feature field for point cloud scenes, we draw inspiration from the approach proposed by Wang et al. [17] for projecting 2D semantic features into 3D space using DINOv2 [13]. In their framework, multi-view RGBD images are processed to extract DINOv2 features, which are then projected onto arbitrary 3D coordinates by mapping them to each camera's image space, interpolating features, and fusing them across views. Specifically, for a 3D point $\mathbf{x}$, D^3Field [17] compute its projection $\mathbf{u}_i$ in the i-th camera view, determine the truncated depth difference $d_i = r_i - r_i'$ (where r_i is sensor-captured depth and r_i' is the interpolated depth), and assign visibility v_i and weight w_i to prioritize points near the surface. These weights guide the fusion of semantic features $\mathbf{f}_i$ and instance masks $\mathbf{p}_i$ across K views, yielding a unified 3D descriptor field. We adapt this method to our point cloud representation by aligning multi-view RGB observations with the point cloud geometry, projecting DINOv2 features onto the 3D points, and fusing them to encode semantic information directly onto the point cloud structure. This approach enables efficient and generalizable

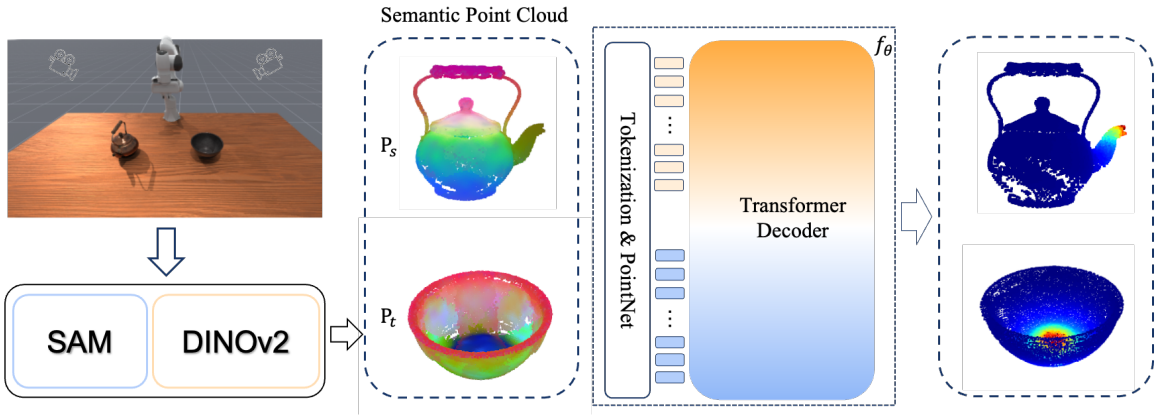

Figure 4.1: Framework pipeline.

feature representation without additional training, supporting robust scene understanding in our application.

4.3 One-Shot Affordance Grounding

To enable effective affordance prediction from point clouds, we propose a neural network architecture that integrates geometric encoding, cross-object attention mechanisms, and explicit type embeddings. Our network processes paired point clouds representing a source object ($\mathbf{P}_{\text{src}} \in R^{B\times N\times 3}$) and a target object ($\mathbf{P}_{\text{tgt}} \in R^{B\times N\times 3}$), alongside corresponding visual features extracted from DINOv2 [13], denoted by $\mathbf{F}_{\text{src}}, \mathbf{F}_{\text{tgt}} \in R^{B\times N\times 1024}$. The output consists of per-point affordance scores optimized via binary cross-entropy (BCE) loss against ground-truth annotations during training.

Our model comprises four primary components: (1) a **Point Cloud Encoder**, (2) a **Type Embedding Mechanism**, (3) a **Cross-Attention Module**, and (4) a **Projection Head**.

4.3.1 Point Cloud Encoder

The *PointCloudEncoder* first jointly processes concatenated point clouds and their DINOv2-derived features. Given the combined input $\mathbf{P} = [\mathbf{P}_{\text{src}}, \mathbf{P}_{\text{tgt}}] \in R^{2B\times N\times 3}$ and features $\mathbf{F} \in R^{2B\times N\times 1024}$, the encoder employs a hierarchical structure with hidden dimensions [784, 512] to aggregate local geometric information into compact patch-level feature representations. This produces tokenized features $\mathbf{Z} \in R^{2B\times T\times 512}$ and corresponding patch centroids $\mathbf{C} \in R^{2B\times T\times 3}$, where $T = 256$ denotes the number of patches per object.

The encoded features are then separated into source and target object representations: $\mathbf{Z}_{\text{src}}, \mathbf{Z}_{\text{tgt}} \in R^{B\times T\times 512}$.

4.3.2 Type Embedding Mechanism

To explicitly distinguish between source and target objects, we introduce a learnable type embedding vector. Specifically, we define a fixed one-hot style type embedding $\mathbf{E}_{\text{type}} \in R^{2\times 512}$, assigning $\mathbf{E}_{\text{type}}[0] = +\mathbf{1}$ for source objects and $\mathbf{E}_{\text{type}}[1] = -\mathbf{1}$ for target objects. The embeddings are replicated across tokens and added to the respective features to clearly encode object roles:

$$\mathbf{Z}_{\text{src}} \leftarrow \mathbf{Z}_{\text{src}} + \mathbf{E}_{\text{type}}[0], \quad \mathbf{Z}_{\text{tgt}} \leftarrow \mathbf{Z}_{\text{tgt}} + \mathbf{E}_{\text{type}}[1].$$

This embedding significantly enhances the model's understanding of inter-object affordance semantics.

4.3.3 Cross-Attention Module

We apply a bidirectional cross-attention mechanism to enable dynamic feature interaction between source and target objects. Specifically, a multi-head attention module with 8 heads captures contextual dependencies:

$$\mathbf{A}_{\text{src}} = \text{CrossAttention}(\mathbf{Z}_{\text{src}}, \mathbf{Z}_{\text{tgt}}, \mathbf{Z}_{\text{tgt}}),$$

$$\mathbf{A}_{\text{tgt}} = \text{CrossAttention}(\mathbf{Z}_{\text{tgt}}, \mathbf{Z}_{\text{src}}, \mathbf{Z}_{\text{src}}).$$

Residual connections are then employed to integrate attention results back into original feature representations:

$$\mathbf{Z}_{\text{src}}^{\text{final}} = \mathbf{Z}_{\text{src}} + \mathbf{A}_{\text{src}}, \quad \mathbf{Z}_{\text{tgt}}^{\text{final}} = \mathbf{Z}_{\text{tgt}} + \mathbf{A}_{\text{tgt}}.$$

This module effectively captures nuanced affordance interactions between objects.

4.3.4 Projection Head and Point-wise Prediction

To generate per-point predictions, we interpolate the final patch-level embeddings back to individual points in the original point clouds using nearest-neighbor interpolation based on patch centroids $\mathbf{C}$. This yields dense point-level embeddings $\mathbf{E}_{\text{src}}, \mathbf{E}_{\text{tgt}} \in R^{B\times N\times 512}$.

These dense embeddings are concatenated and passed through a lightweight projection head, comprising:

- A linear transformation from 512 to 256 dimensions.

- Layer normalization and GELU activation.
- Dropout regularization (rate 0.1).
- Final linear projection from 256 dimensions to scalar values.

A sigmoid activation function produces affordance scores $\mathbf{S} \in [0,1]^{2B\times 6\times N}$, representing predicted interaction probabilities, where the dimension 6 corresponds to separate channels for each affordance type prediction.

4.3.5 Training and Optimization

During training, we optimize our network using the binary cross-entropy (BCE) loss:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{2BN}\sum_{i=1}^{2B}\sum_{j=1}^{N}\left[L_{ij}\log(S_{ij}) + (1-L_{ij})\log(1-S_{ij})\right],$$

where L_{ij} represents ground-truth affordance labels.

During inference, the network directly outputs per-point affordance predictions, facilitating effective robotic manipulation planning and execution in real-world scenarios.

4.4 Affordance-Based Planning with LLMs

In this section, we introduce an approach for affordance-based robotic planning by leveraging LLMs. Our method systematically translates affordance data embedded in 3D point clouds into explicit geometric and semantic constraints. Specifically, given a source object point cloud $\mathbf{P}_{\text{src}}$ and a target object point cloud $\mathbf{P}_{\text{tgt}}$, along with their respective affordance maps $\mathbf{A}_{\text{src}}$ and $\mathbf{A}_{\text{tgt}}$, we optimize a 6-DoF transformation $\mathbf{T} \in R^{4\times 4}$ that aligns objects appropriately for the intended task.

Our pipeline includes three key steps which can be summarized in Fig 4.2.

4.4.1 Affordance Region Identification

To efficiently utilize affordance information, we first extract significant interaction regions from each object's point cloud. We cluster points using DBSCAN [39] based on affordance scores, selecting points above a specified percentile (typically top 25%) to form distinct regions, denoted as $\mathcal{R}_{\text{src}}$ and $\mathcal{R}_{\text{tgt}}$. Each region is characterized by its centroid, constituent points, and averaged affordance scores, effectively condensing essential geometric and semantic details.

4.4.2 Constraint Generation with LLMs

We leverage the spatial reasoning capabilities of LLMs to translate object affordance and geometry representation into explicit constraints through code generation. This approach abstracts semantic affordance instructions (e.g., “pouring” or “cutting”) into concrete geometric constraints such as alignment, orientation, and spatial relations. The generated constraints include:

- **Affordance Alignment Constraint**: Ensures optimal alignment between source and target high-affordance regions.
- **Positional Constraint**: Enforces spatial relationships such as "above," "inside," or "aligned," based on task semantics.
- **Orientational Constraint**: Ensures objects maintain specific orientations (e.g., tilted or perpendicular) appropriate for the manipulation.
- **Collision Avoidance Constraint**: Prevents unrealistic penetrations or collisions.
- **Stability Constraint**: Ensures physically feasible object placement to maintain stability post-manipulation.

Each constraint is guided by task-specific semantic reasoning provided by the LLM, making our approach robust across varied manipulation scenarios. I provide an example contraint function generated by LLM in Fig 4.3. For detailed prompt template, we provide it in Appendix.

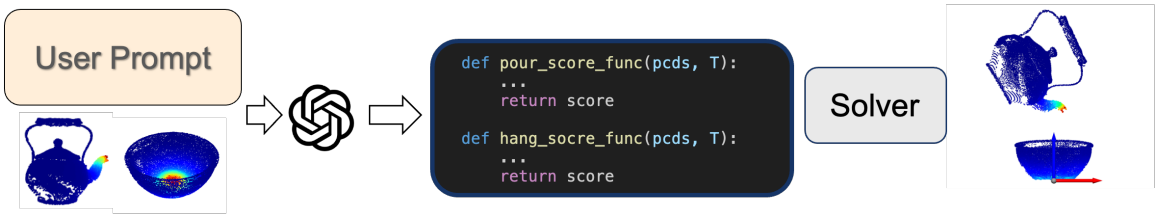

Figure 4.2: LLM Planning Pipeline.

```python
def evaluate_pour_interaction_score(src_pcd, tgt_pcd, transform):
    # 1. Transform source point cloud using the given transformation
    src_transformed = transform_pointcloud(src_pcd, transform)

    # 2. Extract high-affordance regions (mouth and receiver)
    src_mouth_region = extract_high_affordance_region(src_transformed)
    tgt_receiver_region = extract_high_affordance_region(tgt_pcd)

    # 3. Compute affordance alignment score
    score_aff = compute_alignment_score(src_mouth_region, tgt_receiver_region)

    # 4. Check vertical positioning: source should be above target
    score_pos = compute_position_penalty(src_mouth_region, tgt_receiver_region)

    # 5. Check tilt angle: pouring requires the source to be tilted
    score_ori = compute_orientation_penalty(transform)

    # 6. Check collision: ensure source does not intersect target
    score_clear = compute_clearance_penalty(src_transformed, tgt_pcd)

    # 7. Weighted sum of all constraints
    total_score = weighted_sum(score_aff, score_pos, score_ori, score_clear)

    return total_score
```

Figure 4.3: Example constraint function used for pouring interactions.

4.4.3 Transformation Optimization

We formulate the transformation planning problem as a constrained optimization over the rigid-body transformation $\mathbf{T} \in SE(3)$ applied to the source object. Our goal is to find the optimal $\mathbf{T}$ that minimizes a composite cost, which evaluates the plausibility and feasibility of the resulting interaction configuration. Formally, we solve:

$$\begin{aligned} \min_{\mathbf{T} \in SE(3)} \quad & w_{\text{align}} S_{\text{align}}(\mathbf{T}) + w_{\text{pos}} S_{\text{pos}}(\mathbf{T}) + w_{\text{orient}} S_{\text{orient}}(\mathbf{T}) \\ & + w_{\text{coll}} S_{\text{coll}}(\mathbf{T}) + w_{\text{stab}} S_{\text{stab}}(\mathbf{T}) \\ \text{s.t.} \quad & \mathbf{T} \text{ satisfies task-specific constraints (e.g., reachability, visibility)} \end{aligned} \tag{4.1}$$

Each term in the objective corresponds to a soft constraint defined as follows:

- **Affordance Alignment Score** $S_{\text{align}}(\mathbf{T})$: Measures the proximity between transformed source affordance regions $\mathcal{R}'_{\text{src}} = \mathbf{T} \cdot \mathcal{R}_{\text{src}}$ and target regions $\mathcal{R}_{\text{tgt}}$.
- **Positional Score** $S_{\text{pos}}(\mathbf{T})$: Penalizes deviation from desired relative positions, e.g., vertical offset between functional regions.
- **Orientational Score** $S_{\text{orient}}(\mathbf{T})$: Quantifies angular deviation of transformed axes from task-specific reference directions.
- **Collision Score** $S_{\text{coll}}(\mathbf{T})$: Penalizes intersections between transformed source point cloud and the target scene.
- **Stability Score** $S_{\text{stab}}(\mathbf{T})$: Evaluates physical stability based on the placement of the object's centroid over the support base.

This optimization is solved using gradient-based methods (e.g., Adam or L-BFGS) with numerical gradients computed over the transformed geometry.

Chapter 5

Experiments

Here, we evaluate our method for both affordance prediction and robotic manipulation tasks. We aim to answer three key research questions:

- How effectively does our method perform in object-to-object affordance grounding tasks?
- To what extent can our method generalize when trained with only a single example for each affordance type?
- How effectively can our method improve downstream LLM's spatial planning and policy training?

We first demonstrate our experiment settings in Sec 5.1, then we address the above three questions through two stages: evaluating the accuracy and generalization capability of affordance grounding (Sec 5.2) and validating our approach through simulation manipulation experiments (Sec 5.3)

5.1 Experiment Setup

Given the absence of high-quality object-to-object affordance grounding datasets, we annotate and construct our own dataset in simulation using SAPIEN [40]. The affordance map is generated in two steps: first we have several user-assigned contact points on the point cloud and then propagate the affordance label to other points based on distance following [9].

We conduct manipulation experiments in simulation environments using SAPIEN , we position four stereo-depth sensor from different viewpoints around the workspace and employ GPT-4o [41] from OpenAI as the vision-language model for planning. We design six tasks that require two

objects to interact meaningfully (e.g., with correct contact poses) with each other: *pouring from teapot into bowl*, *inserting pen into penholder*, *knocking button with hammer*, *hanging mug onto mug tree*, *cutting apple with knife*, and *plugging in charger* with the goal of evaluating our system's performance in varying contact geometries, force applications, and spatial relationships between objects. Our training dataset consists of a single pair of interacting objects for each affordance type.

5.2 Affordance Grounding

5.2.1 Comparison Results

Baselines We evaluate our affordance grounding module against 3 baselines :

i. **O2O-Afford [10]**: O2O-Afford addresses data limit issue through extracting contact area in simulation.

ii. **IAGNet [24]**: IAGNet addressed the task of grounding 3D object affordance from 2D interactions in images.

iii. **RoboPoint [42]**: RoboPoint is a VLM that can predict image keypoints affordance given language.

We adopt four metrics during evaluation: **aIOU** [43], **SIMilarity** [44], **MAE** [45], and **AUC** [46], computed as follows:

- **Average Intersection-over-Union (aIOU)** measures the overlap between predicted affordance maps (A_{pred}) and ground-truth maps (A_{gt}):

$$\text{aIOU} = \frac{1}{N} \sum_{i=1}^{N} \frac{|A_{\text{pred}}^{(i)} \cap A_{\text{gt}}^{(i)}|}{|A_{\text{pred}}^{(i)} \cup A_{\text{gt}}^{(i)}|}$$

- **Similarity (SIM)** quantifies similarity between prediction and ground-truth affordance distributions:

$$\text{SIM} = \frac{\sum_j \min(A_{\text{pred},j}, A_{\text{gt},j})}{\sum_j A_{\text{gt},j}}$$

- **Mean Absolute Error (MAE)** evaluates the pixel-wise average absolute difference between predicted and ground-truth affordances:

$$\text{MAE} = \frac{1}{M} \sum_{j=1}^{M} |A_{\text{pred},j} - A_{\text{gt},j}|$$

- **Area Under the ROC Curve (AUC)** measures the discriminative capability of predicted affordances over multiple thresholds, where TPR and FPR denote true-positive and false-positive rates, respectively:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR}^{-1}(x))\, dx$$

Results Quantitative results are presented in Tab 5.1. We also present qualitative results of our method against all baselines in Fig 5.1. Each row in Fig 5.1 represents a distinct affordance type, with three different affordances illustrated. For each affordance, we present two corresponding rows: the upper showing the source object and the lower displaying the target object. O2O-Afford demonstrates the poorest performance, attributable to its contact-based affordance data collection strategy, which results in highly unpredictable model predictions as evident in our qualitative results. RoboPoint, as a vision-language model, demonstrates capability in object localization but lacks the precision to infer fine-grained affordance regions on objects. IAGNet exhibits the strongest performance among all baselines but suffers significantly from overfitting due to the one-shot training paradigm. Qualitative results reveal that while it predicts effectively for objects with similar geometry, it fails to generalize to more complex unseen instances, much less to novel categories. Our method significantly outperforms all baselines and demonstrates robust generalization capabilities.

Method	**↑ IOU**	**↑ SIM**	**↓ MAE**	**↑ AUC**
IAGNet	14.81	0.5574	0.1402	73.30
RoboPoint	11.84	0.4376	0.3344	59.78
Ours	**17.82**	**0.6387**	**0.0612**	**96.00**

Table 5.1: **Quantitative comparison on object-to-object affordance grounding.**

5.2.2 Generalization Experiments

We present additional results demonstrating our model's generalization capabilities across two distinct dimensions: partial point cloud generalization and unseen category generalization. We evaluate partial point cloud generalization because occlusions frequently occur in real-world scenarios, making complete point cloud observations difficult to obtain. The ability to generalize to partial point clouds significantly enhances system robustness for real-world robotic applications. We also evaluate unseen category generalization because real-world manipulation tasks often involve functionally similar but categorically distinct objects—for instance, substituting a stick for a pen when inserting

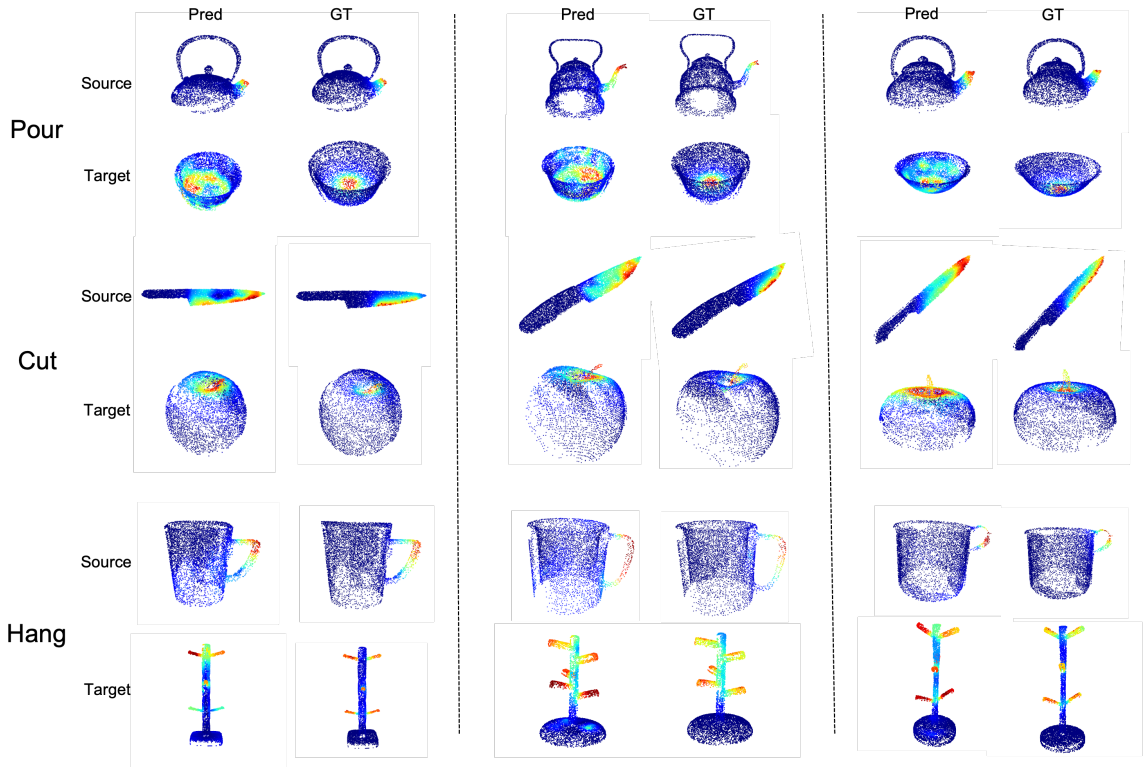

Figure 5.1: Affordance Qualitative Results.

into a holder, where the manipulated object belongs to a category entirely absent from the training data. Generalization to such cases substantially contributes to developing more versatile robotic manipulation systems with broader applicability.

Partial Point Cloud Generalization We test our model's generalization ability by removing two views of observation, leaving only two camera in the workspace which is a more common case in real world setup. We present qualitative results on pouring action in Fig 5.2. We hypothesize that the generalizable semantic features extracted from DINOv2 facilitate effective performance even with partial point clouds that cannot provide complete geometric information.

Unseen Category Generalization We evaluate our model's generalization capability on unseen source object categories. We test three novel objects: multi-tool knife, scissors, and spray bottle. Qualitative results are presented in Fig 5.3. Despite the geometry of multi-tool knife and scissors differing substantially from the knife in our training distribution, our model effectively generalizes to these unseen categories due to the semantic feature preservation. Similarly, although the semantic properties of spray bottles differ from teapots in our training set, our model successfully generalizes based on geometric similarities. These results demonstrate that our model's generalization

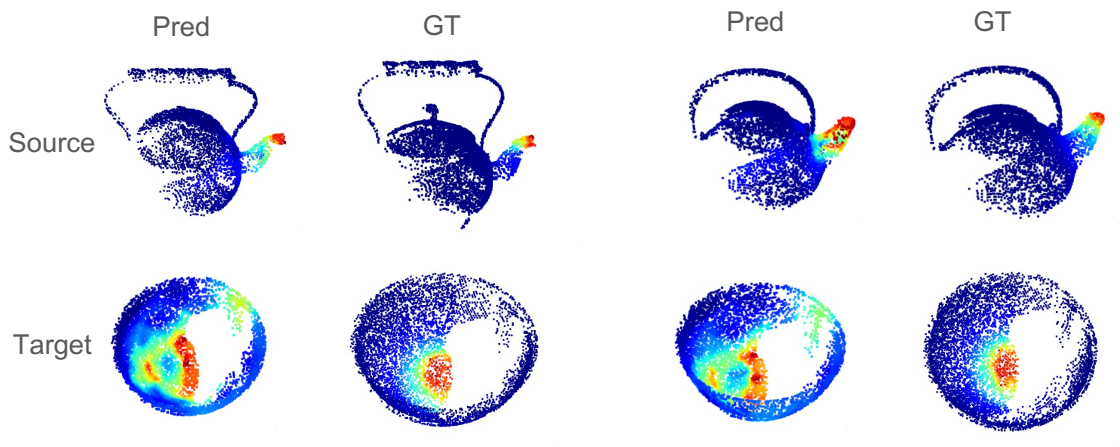

Figure 5.2: Qualitative Results of Partial Point Cloud.

capabilities derive from both semantic feature extraction and geometric similarity analysis—two complementary mechanisms that underpin its robust transfer performance across novel objects.

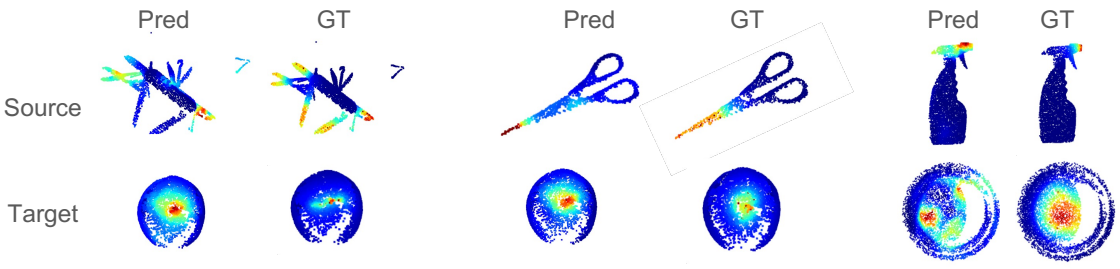

Figure 5.3: Qualitative Results of Unseen Object Category.

5.3 Affordance-based Manipulation

Baselines We evaluate LLM-based planning against an ablated version of our method that performs planning directly from object point clouds. We present quantitative results in Tab 5.2. We evaluate our approach against the baseline across six manipulation tasks, conducting ten trials per task and recording success rate as the primary evaluation metric.

Results According to qualitative results, affordance, as a mid-level representation, significantly enhances manipulation success rates. In common tasks requiring two-object interaction, our method achieves approximately 80% success rate, while the baseline exhibits considerably poorer performance due to its inability to recognize functional properties from object point clouds. In more complex and extended-horizon tasks such as hanging and plugging, the baseline fails in all trials, whereas our

method maintains approximately 50% success rate. We present qualitative results in Fig 5.4. Each row depicts a distinct manipulation task, while each column illustrates the sequential progression of the task execution. As illustrated in Fig 5.4, our tasks involve extended time horizons and require precise contact for successful completion. Our results demonstrate that affordance representation significantly enhances LLMs' spatial reasoning capabilities, as evidenced by the resulting coherent and purposeful manipulation sequences. The integration of predicted affordance maps enables the LLMs to make more informed spatial decisions, leading to substantially improved task performance.

Method	**Pouring**	**Hanging**	**Pressing**	**Putting**	**Cutting**
Baseline	2/10	0/10	5/10	5/10	4/5
Ours (%)	**8/10**	**5/10**	**9/10**	**8/10**	**9/10**

Table 5.2: **Success rate comparison on robotic manipulation tasks.**

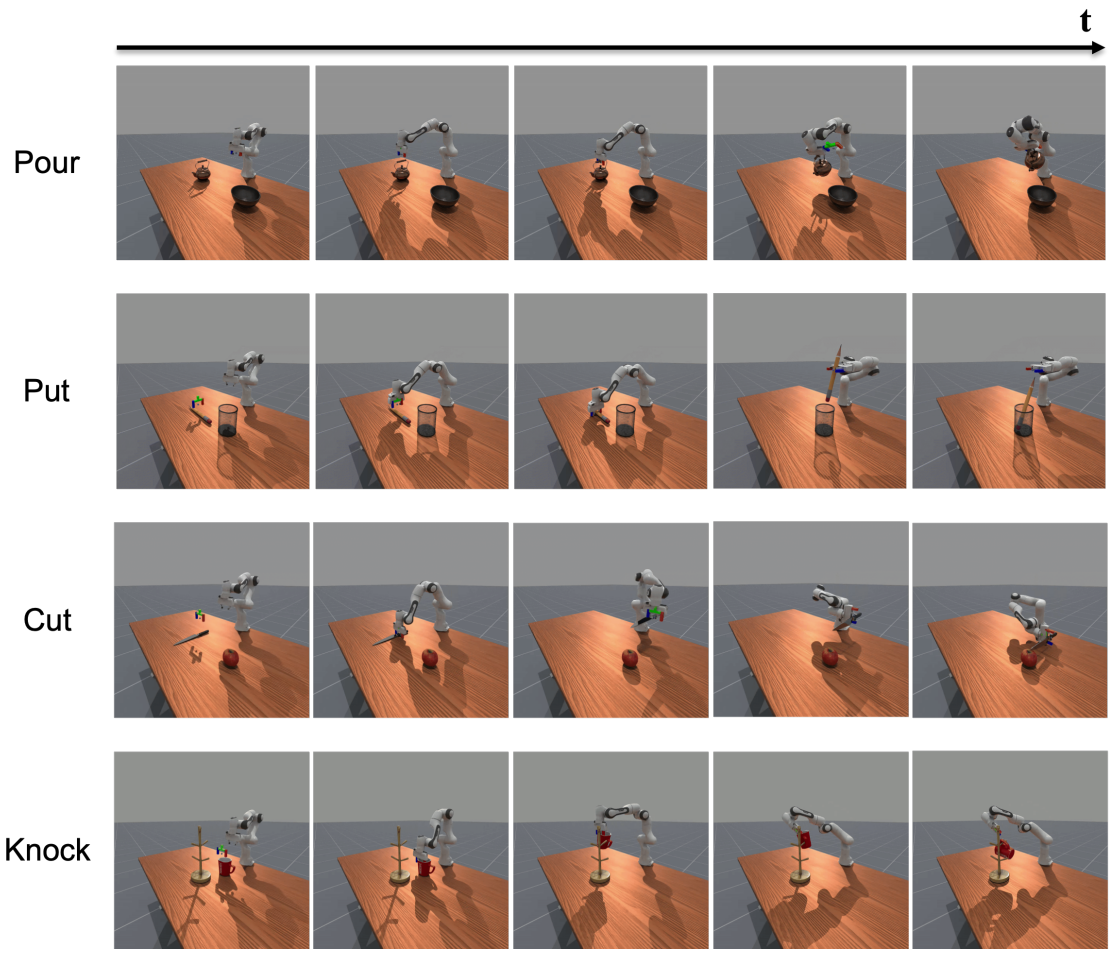

Figure 5.4: Manipulation Qualitative Results.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

In this thesis, we presented **O^{3}Afford**, a novel framework designed for one-shot object-to-object affordance grounding using 3D semantic features distilled from vision foundation models (VFMs). Motivated by the limitations of traditional affordance grounding approaches—which either neglect essential geometric contexts or predominantly focus on single-object interactions—we explicitly addressed complex, real-world manipulation tasks involving multiple interacting objects. Our method leverages semantic features extracted from pre-trained VFMs and integrates them effectively into 3D point cloud representations, enriching geometric data with robust semantic understanding. We further introduced a bi-directional affordance discovery module that captures reciprocal geometric and semantic relationships between source and target objects, significantly enhancing the accuracy and generalization capabilities of affordance predictions. Additionally, we demonstrated the practical effectiveness of integrating our framework with large language models, which provided superior spatial reasoning and robust manipulation planning in both simulated and real-world robotic scenarios. Comprehensive evaluations across multiple challenging manipulation tasks—including pouring, hanging, pressing, putting, cutting, and plugging—highlighted our method's consistent improvement over existing state-of-the-art methods in both prediction accuracy and robotic manipulation success rates. Notably, our approach exhibited remarkable generalization across different viewpoints, unseen object instances, and novel object categories, demonstrating its broad applicability. Looking forward, our proposed method presents numerous exciting avenues for further research, such as incorporating physics-informed neural models to capture complex dynamic interactions, and extending

to manipulation scenarios involving deformable objects or multi-step tasks. Ultimately, by bridging the semantic richness of vision foundation models and 3D geometric reasoning, our work advances robots' capability to intuitively interpret and autonomously interact with their environments, marking a significant step toward achieving human-like manipulation intelligence.

6.2 Future Work

While our proposed **O^{3}Afford** framework has demonstrated strong performance in one-shot object-to-object affordance grounding, several promising directions remain to be explored. First, establishing a comprehensive benchmark specifically dedicated to object-to-object affordance grounding would greatly benefit the research community. Such a benchmark should encompass diverse scenarios, tasks, and interactions, clearly defining evaluation metrics and datasets to facilitate consistent and fair comparisons among emerging methods. Second, to further validate the practical applicability and robustness of our method, more extensive real-world robotic experiments across diverse environmental conditions and manipulation tasks should be conducted. Expanding real-world validations would not only confirm the robustness of our method but also identify potential challenges and limitations inherent to practical robot deployments. Third, improvements in neural network architecture and the one-shot learning paradigm hold considerable potential. Future efforts could explore integrating meta-learning or advanced few-shot learning strategies to further enhance generalization from minimal supervision. Additionally, designing neural models that more effectively capture subtle physical interactions and dynamic properties between objects could significantly boost affordance prediction accuracy, paving the way toward more sophisticated, human-level robotic manipulation capabilities.

References

[1] Victor Kaptelinin and Bonnie Nardi. Affordances in hci: toward a mediated action perspective. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 967–976, 2012.

[2] Alexandros Delitzas, Ayca Takmaz, Federico Tombari, Robert Sumner, Marc Pollefeys, and Francis Engelmann. Scenefun3d: fine-grained functionality and affordance understanding in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14531–14542, 2024.

[3] Yuxuan Kuang, Junjie Ye, Haoran Geng, Jiageng Mao, Congyue Deng, Leonidas Guibas, He Wang, and Yue Wang. Ram: Retrieval-based affordance transfer for generalizable zero-shot robotic manipulation. *arXiv preprint arXiv:2407.04689*, 2024.

[4] Gen Li, Deqing Sun, Laura Sevilla-Lara, and Varun Jampani. One-shot open affordance learning with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3086–3096, 2024.

[5] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2252–2261, 2022.

[6] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Leverage interactive affinity for affordance learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6809–6819, 2023.

[7] Wei Zhai, Hongchen Luo, Jing Zhang, Yang Cao, and Dacheng Tao. One-shot object affordance detection in the wild. *International Journal of Computer Vision*, 130(10):2472–2500, 2022.

[8] Xue Zhao, Yang Cao, and Yu Kang. Object affordance detection with relationship-aware network. *Neural Computing and Applications*, 32(18):14321–14333, 2020.

[9] Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3d affordancenet: A benchmark for visual object affordance understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1778–1787, June 2021.

[10] Kaichun Mo, Yuzhe Qin, Fanbo Xiang, Hao Su, and Leonidas Guibas. O2o-afford: Annotation-free large-scale object-object affordance learning. In *Conference on robot learning*, pages 1666–1677. PMLR, 2022.

[11] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[13] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[17] Yixuan Wang, Mingtong Zhang, Zhuoran Li, Tarik Kelestemur, Katherine Driggs-Campbell, Jiajun Wu, Li Fei-Fei, and Yunzhu Li. D^3fields: Dynamic 3d descriptor fields for zero-shot generalizable rearrangement. In *8th Annual Conference on Robot Learning*, 2024.

[18] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[19] Guangxing Han and Ser-Nam Lim. Few-shot object detection with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28608–28618, June 2024.

[20] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36:45533–45547, 2023.

[21] Soroush Nasiriany, Sean Kirmani, Tianli Ding, Laura Smith, Yuke Zhu, Danny Driess, Dorsa Sadigh, and Ted Xiao. Rt-affordance: Affordances are versatile intermediate representations for robot manipulation. *arXiv preprint arXiv:2411.02704*, 2024.

[22] Olivia Y Lee, Annie Xie, Kuan Fang, Karl Pertsch, and Chelsea Finn. Affordance-guided reinforcement learning via visual prompting. *arXiv preprint arXiv:2407.10341*, 2024.

[23] Cuiyu Liu, Wei Zhai, Yuhang Yang, Hongchen Luo, Sen Liang, Yang Cao, and Zheng-Jun Zha. Grounding 3d scene affordance from egocentric interactions. *arXiv preprint arXiv:2409.19650*, 2024.

[24] Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Grounding 3d object affordance from 2d interactions in images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10905–10915, October 2023.

[25] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle,

M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[26] Haoyu Song, Li Dong, Wei-Nan Zhang, Ting Liu, and Furu Wei. Clip models are few-shot learners: Empirical studies on vqa and visual entailment. *arXiv preprint arXiv:2203.07190*, 2022.

[27] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[29] Ri-Zhao Qiu, Yu-Xiong Wang, and Kris Hauser. Alignddiff: aligning diffusion models for general few-shot segmentation. In *European Conference on Computer Vision*, pages 384–400. Springer, 2024.

[30] Muzhi Zhu, Yang Liu, Zekai Luo, Chenchen Jing, Hao Chen, Guangkai Xu, Xinlong Wang, and Chunhua Shen. Unleashing the potential of the diffusion model in few-shot semantic segmentation. *arXiv preprint arXiv:2410.02369*, 2024.

[31] Weimin Tan, Siyuan Chen, and Bo Yan. Diffss: Diffusion model for few-shot semantic segmentation. *arXiv preprint arXiv:2307.00773*, 2023.

[32] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. In *International Conference on Learning Representations (ICLR)*, 2024.

[33] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. *arXiv preprint arXiv:2112.02413*, 2021.

[34] Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6813–6823, 2021.

[35] Zengyi Qin, Kuan Fang, Yuke Zhu, Li Fei-Fei, and Silvio Savarese. Keto: Learning keypoint representations for tool manipulation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7278–7285. IEEE, 2020.

[36] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. In Jie Tan, Marc Toussaint, and Kourosh Darvish, editors, *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 540–562. PMLR, 06–09 Nov 2023.

[37] Jessica Borja-Diaz, Oier Mees, Gabriel Kalweit, Lukas Hermann, Joschka Boedecker, and Wolfram Burgard. Affordance learning from play for sample-efficient policy learning. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6372–6378. IEEE, 2022.

[38] Oier Mees, Jessica Borja-Diaz, and Wolfram Burgard. Grounding language with visual affordances over unstructured data. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, 2023.

[39] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. Density-based spatial clustering of applications with noise. In *Int. Conf. knowledge discovery and data mining*, volume 240, 1996.

[40] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. SAPIEN: A simulated part-based interactive environment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[41] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[42] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: A vision-language model for spatial affordance prediction for robotics. *arXiv preprint arXiv:2406.10721*, 2024.

[43] Md Atiqur Rahman and Yang Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International symposium on visual computing*, pages 234–244. Springer, 2016.

[44] Michael J Swain and Dana H Ballard. Color indexing. *International journal of computer vision*, 7(1):11–32, 1991.

[45] Cort J Willmott and Kenji Matsuura. Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance. *Climate research*, 30(1):79–82, 2005.

[46] Jorge M Lobo, Alberto Jiménez-Valverde, and Raimundo Real. Auc: a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography*, 17(2):145–151, 2008.

Appendix A

Prompt Template

Prompt Template Certain Contraint

You are given two 3D objects represented as point clouds. Each point is associated with an affordance score predicted by a perception model. Your task is to propose a constraint function for the specified affordance type that evaluates how well the source object can interact with the target object.

Inputs:

- `Source Object Name`: [SRC_OBJECT_NAME]
- `Target Object Name`: [TGT_OBJECT_NAME]
- `Interaction Type`: [AFFORDANCE] (e.g., pour, hang, press, cut, put, plugin)
- `Source Point Cloud`: $\{(x_i, y_i, z_i)\}_{i=1}^{N}$, with affordance scores $\{a_i\}_{i=1}^{N}$
- `Target Point Cloud`: $\{(x_j, y_j, z_j)\}_{j=1}^{M}$, with affordance scores $\{b_j\}_{j=1}^{M}$

Task: Generate a constraint function that evaluates the quality of an affordance-specific interaction between source and target objects. The function should consider high-affordance regions, interaction-specific spatial constraints, and physical plausibility.

Code Skeleton:

```
def compute_alignment_score(src_aff, tgt_aff, src_pcd, tgt_pcd):
    score = 0
    """
    # TODO: Implement affordnace alignment contraint

    return score
```

Constraints:

- The function should use the information from high-affordance regions of both objects
- The evaluation must reflect the semantic meaning of the specified affordance type
- Consideration should be given to physical feasibility of the interaction
- All constraints should be combined into a single cost value (lower is better)