Efficient Affordance Learning across Vision, Language, and Robotic Manipulation

Gen Li

Doctor of Philosophy

Institute of Perception, Action and Behaviour

School of Informatics

University of Edinburgh

2025

Abstract

Affordance refers to the action possibilities offered by the environment, indicating how objects can be interacted with. This concept is fundamental to embodied intelligent systems, facilitating a transition from passive perception to active interaction. Many existing approaches to affordance learning rely on large-scale datasets with extensive pixel-wise annotations, which are both time-consuming and resource-intensive to produce. Such methods, while effective, limit the scalability and applicability in real-world scenarios.

To enhance the efficiency of affordance learning, this thesis presents data- and resource-efficient approaches, spanning the domains of vision, language, and robotic manipulation. Our methods achieve highly efficient affordance learning by exploring weak supervision, data-limited learning paradigms, and human-object interactions. These include the use of action labels, a minimal set of annotated samples, and visual data such as images or videos depicting human-object interactions.

The thesis begins with the task of weakly supervised affordance grounding. Utilizing weak supervision, such as image-level action labels, greatly reduces the reliance on costly annotations. To this end, we introduce a novel pipeline, which extracts affordance knowledge from human-object interaction images and transfers it to object images in a localized scheme. This work achieves state-of-the-art results with far fewer parameters and faster inference speed compared to prior approaches.

While weak supervision is effective, it often requires substantial training data and struggles with unseen categories. To address these limitations, we leverage vision and language foundation models to explore generalizable affordance learning under data-scarce conditions. This data-limited vision-language affordance learning, which we call one-shot open affordance learning, aims to yield comparable performance using only one sample per object category. Building on a comprehensive analysis, we propose three simple yet effective modules, achieving leading results using only a fraction of the full training data.

The first two studies focus on visual affordance learning and demonstrate strong performance in vision-based evaluation; however, they lack real-world interactions, which are crucial for demonstrating practical applicability. To bridge this gap, we introduce a holistic affordance learning system that integrates affordance data collection, model learning, and robot deployment. This work begins with an automated pipeline for collecting and annotating data from egocentric human videos. It then presents

a highly effective model featuring an innovative depth injector to incorporate geometric priors. Finally, it includes a grasping pipeline to enable affordance-oriented robotic manipulation, including tool grasping, tool-object interaction, and robot-to-human handover.

In summary, this thesis tackles key challenges in affordance learning, including data efficiency, generalization, and real-world deployment. The proposed methods not only advance the state of the art but also demonstrate strong applicability in real-world scenarios, laying a foundation for future developments in affordance learning at the intersection of computer vision and robotics.

Lay Summary

Humans effortlessly understand how to use everyday objects, such as grasping a mug by the handle or cutting with the sharp edge of a knife. This intuitive sense of how objects can be used is called affordance. For robots and AI systems, affordance understanding is essential, as it allows them to interact with objects purposefully and safely in the real world.

Traditional approaches to teaching machines affordance knowledge rely on large, manually labeled datasets that specify where and how objects can be interacted with. However, creating such datasets is labor-intensive, costly, and difficult to scale, especially in real-world settings where variety and complexity are high. This thesis tackles this challenge by developing a series of novel, efficient approaches that reduce the reliance on expensive data annotations. The research spans three fundamental aspects of intelligent systems: vision, which allows robots to perceive and interpret their surroundings; language, which helps them understand descriptions of actions and object functions; and manipulation, which enables them to physically interact with and use objects.

The first part of the work introduces a method for learning object affordances using weak supervision, leveraging action labels from human-object interaction images instead of detailed pixel-level annotations. This significantly lowers the data requirements while still achieving strong performance in identifying how objects can be used. Building on this, the second part explores how vision and language foundation models, trained with massive amount of data, can help robots generalize affordance understanding from only a few examples. We propose a one-shot learning framework that allows affordance recognition for unseen objects after seeing just one sample per base object category, making the system highly adaptable to unfamiliar environments. The final part of the thesis focuses on applying affordance learning in the physical world. It presents a complete system that automatically extracts affordance knowledge from videos of people interacting with objects and transfers that knowledge to real robots. These robots can then choose tools based on task requirements, grasp them appropriately, and even hand them over to humans.

In conclusion, this thesis contributes to the development of efficient and generalizable affordance learning methods, taking an important step toward building intelligent machines that can operate safely and effectively in human-centered environments.

Acknowledgements

The PhD path is long and filled with challenges, and I am deeply grateful to have made it to the end. I would like to thank all those who accompanied and supported me along the way.

First and foremost, I would like to express my deepest gratitude to my supervisor, Laura Sevilla-Lara, for her unwavering support, trust, and encouragement throughout my PhD. I am also sincerely thankful to my collaborators, Varun Jampani and Deqing Sun. It was a privilege to work with such accomplished researchers, and their mentorship played a key role in helping me gain precious research experience. My gratitude also goes to my second supervisor, Timothy Hospedales, for his co-supervision and valuable input, and to my examiners, Anh Nguyen and Changjian Li, for their time and insightful feedback on my thesis and viva. I am truly grateful to Cheolkon Jung, under whose mentorship I began my research journey. I was also fortunate to receive invaluable guidance from Joongkyu Kim during my master's study, who supported me both academically and personally. A special thanks to my girlfriend, who has been by my side from the very beginning to the end of this journey. Her unwavering support, patience, and companionship have helped me navigate the pressures and difficulties of this demanding period. In addition, I owe my deepest gratitude to my parents, who have witnessed every step of my growth and supported me unconditionally in all my decisions. Their encouragement has always been my strongest pillar. Finally, I want to thank myself—for pushing through countless setbacks, for persevering in moments of self-doubt, and for finding the strength to keep going when the path was winding and unclear.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Gen Li*)

Table of Contents

1 Introduction **1**

1.1 Motivation and Challenges 1

1.2 Contributions and Outline 4

2 Related Work **6**

2.1 Affordance Learning in Vision 6

2.1.1 Fully Supervised Affordance Learning 6

2.1.2 Weakly Supervised Affordance Learning 7

2.1.3 Learning Affordance from Human-Object Interactions 8

2.2 Affordance Learning with Language 8

2.3 Affordance Learning for Robotic Manipulation 9

3 Affordance Grounding with Weak Supervision **11**

3.1 Introduction 12

3.2 Method 14

3.2.1 Locating Interaction Regions 15

3.2.2 Object-Part Embedding Selection 16

3.2.3 Part-Level Knowledge Transfer 18

3.3 Experiments 19

3.3.1 Experimental Setting 19

3.3.2 Comparison to State-of-the-Art Methods 20

3.3.3 Ablation Study 22

3.4 Conclusion and Limitations 26

4 Data-Limited Vision-Language Affordance Learning **27**

4.1 Introduction 28

4.2 One-Shot Open Affordance Learning 30

4.2.1 Problem Setting 30

4.2.2 Analysis of Foundation Models . . . 31
4.2.3 Motivation and Method . . . 33
4.3 Experiments . . . 36
4.3.1 Datasets . . . 36
4.3.2 Implementation Details . . . 37
4.3.3 Comparison to State-of-the-Art Methods . . . 37
4.3.4 Qualitative Results . . . 39
4.3.5 Ablation Study . . . 41
4.4 Conclusion and Limitations . . . 43

5 An Affordance Learning System for Robotic Manipulation 44
5.1 Introduction . . . 45
5.2 Method . . . 47
5.2.1 Data Collection from Egocentric Videos . . . 48
5.2.2 Geometry-guided Affordance Transformer . . . 50
5.2.3 Affordance-Oriented Robotic Manipulation . . . 52
5.3 Experiments . . . 54
5.3.1 Vision Experiments . . . 54
5.3.2 Robot Experiments . . . 57
5.3.3 Ablation Study . . . 60
5.4 Conclusion and Limitations . . . 62

6 Conclusion 65
6.1 Limitations . . . 65
6.2 Future Work . . . 66

A Affordance Grounding with Weak Supervision 69
A.1 Framework Comparison . . . 69
A.2 Additional Experimental Details . . . 70
A.3 Additional Experimental Results . . . 71
A.4 Additional Visualizations . . . 73

B Data-Limited Vision-Language Affordance Learning 76
B.1 Dataset Details . . . 76
B.2 Ablation Study on Hyperparameters . . . 77
B.3 Additional Visualizations . . . 79

C An Affordance Learning System for Robotic Manipulation **81**
C.1 Dataset Details 81
C.2 Implementation Details 82
C.3 Additional Experimental Results 86
C.4 Discussion 88

Bibliography **90**

List of Figures

1.1 Affordance understanding enables an agent to associate specific object regions with distinct actions, such as grasping a mug by its handle, drinking from its rim, or cutting with the blade of a knife. 3

3.1 Overview of LOCATE. It focuses on regions of exocentric interactions, and clusters visual embeddings into categories such as background, human, and object parts. The prototype of the object-part cluster is then selected to guide egocentric affordance grounding. 13

3.2 State-of-the-art methods in weakly supervised affordance grounding often fail to make accurate predictions for objects with complex structures, *e.g.*, chairs and bicycles. To address this, our model (LOCATE) focuses on localizing and transferring features of object parts, which is able to produce more accurate results. 14

3.3 Overview of the proposed LOCATE framework. It achieves part-level knowledge transfer in three steps: 1) locating interaction regions with ψ_{cam} (Sec. 3.2.1), 2) object-part embedding selection with PartSelect (Sec. 3.2.2), and 3) part-level knowledge transfer with L_{cos} (Sec. 3.2.3). Details for PartSelect are shown in Fig. 3.4. At test time, only the egocentric branch is maintained. 15

3.4 (a) PartSelect picks the object-part prototype through clustering and selection, where $\bigotimes$ denotes the calculation for PartIoU score. (b) The similarity maps between prototypes and exocentric features confirm our statement that each prototype represents the object part, background, and human. 17

3.5 Qualitative comparison between our approach and state-of-the-art affordance grounding methods (Hotspots [97], Cross-view-AG [85], and Cross-view-AG+ [87]). For the unseen setting, the objects presented are not exposed in the training set. For example, the model learns where a *motorcycle* can be ridden during training, and then locates rideable area for a *bicycle* at test time. 22
3.6 Visualization of the qualitative improvements. 23
3.7 Ablation study on the number of prototypes and exocentric images (lower is better). 24

4.1 The pipeline of one-shot open affordance learning. It uses one image per base object for training, and can perform zero-shot inference on novel objects and affordances. 29
4.2 **Analysis of vision-language foundation models** on text-based affordance grounding. The 1st and 3rd rows use affordance texts as input queries, and the 2nd and 4th rows use corresponding object parts as input text queries. Visualizations show that these models have limited ability to recognize fine-grained affordances and object parts. 32
4.3 **Analysis of visual foundation models** on affordance learning. Top row: visualizations of PCA components. We extract features from the last layer of each model and perform PCA on them. Bottom row: feature similarity maps between the yellow mark on the knife blade and the image of scissors. Qualitative results show that DINOv2 has clearer part-aware representations and better part-level semantic correspondence. 33
4.4 Proposed learning framework for OOAL. Our designs are highlighted in three color blocks, which are text prompt learning, multi-layer feature fusion, and CLS-guided transformer decoder. [CLS] denotes the CLS token of the vision encoder. 34
4.5 Qualitative comparison with LOCATE and ZegCLIP on AGD20K dataset. When multiple affordance predictions overlap, the one with higher value is displayed. Our predictions distinguish different object parts, while other methods often make overlapping predictions. 39

4.6 Qualitative comparison with SegFormer and ZegCLIP on UMD affordance dataset in OOAL setting. Images have been enlarged and cropped for better visualization. 40
4.7 Qualitative examples of novel affordance prediction on UMD dataset. The 1st and 2nd rows display results on base objects, and the 3rd and 4th rows show results for novel objects. 41

5.1 Illustration of affordance data collection and robot deployment. Existing work [4, 57, 77, 78] collects the graspable affordance as Gaussian heatmaps, whereas we extract both graspable and functional affordances with precise segmentation masks, enabling tool grasping, tool-object interaction, and robot-to-human tool handover. 46
5.2 Illustration of the data collection process from egocentric videos. First, graspable points (depicted in purple) are localized from clips of hand-object interaction and then projected to pre-contact frame by homography. Next, functional points (depicted in green) are identified from tool-object interactions and mapped to the pre-contact frame of hand-object interaction through point correspondence. Lastly, these points are used as prompts for the SAM to obtain affordance masks. 48
5.3 The architecture of GAT. It consists of a DINOv2 image encoder, a depth feature injector, an embedder, and LoRA layers. The model performs segmentation by computing cosine similarity between upsampled features and learnable / CLIP text embeddings. 50
5.4 The framework of Aff-Grasp. It first employs an open-vocabulary detector [79] to locate all objects within the scene, which are then sent to GAT to determine if they possess the corresponding affordance required for the task. Afterwards, a 6 DoF grasp generation model, Contact-GraspNet, leverages the object's graspable affordance and the depth map to generate dense grasp proposals. Finally, the robot executes affordance-specific sequential motion primitives to apply the functional part to the target. 53
5.5 Qualitative comparison between our approach and other segmentation models on the AED. 55
5.6 Qualitative comparison on unseen objects. 56
5.7 Qualitative examples on novel affordances. 57

5.8 Illustration of accuracy, robustness, and generalization evaluations. The accuracy evaluation requires the model to recognize the affordance of a single object and execute related task. The robustness evaluation involves accurately selecting a object in a cluttered scene to perform a specified affordance task. The generalization evaluation accesses if the model can reason about the graspable area of unseen objects. 58
5.9 Qualitative comparison of graspable affordance predictions on seen and unseen object categories. 60
5.10 Qualitative improvements with DFI and LoRA. 61
5.11 Screenshot of the collected data. Noisy annotations are highlighted with red bounding boxes. 63
5.12 Failure cases. The model fails to recognize objects with complex texture and confuses parts with similar shapes and appearances. 64

A.1 Comparison of LOCATE and previous work. Previous affordance grounding work [85, 97] performs knowledge transfer in a global and implicit manner. In contrast, LOCATE conducts knowledge transfer in a more localized and explicit manner. (GAP denotes global average pooling, and MAP is masked average pooling). 69
A.2 Localization maps for exocentric images. 73
A.3 Similarity maps computed between exocentric embeddings corresponding to the dot/cross and all egocentric features. Here dots and crosses are placed on positions of object parts and humans, respectively. . . . 74
A.4 Qualitative comparison between LOCATE and state-of-the-art affordance grounding methods (Hotspots [97], Cross-view-AG [85], and Cross-view-AG+ [87]) in both seen and unseen settings. 75

B.1 Different affordance annotation schemes. Dense affordance annotation is labeled as binary masks. Sparse affordance annotation is first labeled as keypoints, and then a gaussian kernel is performed over each point to produce pixel-wise ground truth. 77
B.2 Visualization of CLS-guided mask. 79
B.3 Qualitative examples of unseen affordance prediction on AGD20K dataset. The 2nd column shows the results on seen affordances, and the 3rd and 4th columns show results with unseen affordances. 80
B.4 Additional qualitative comparison on AGD20K dataset. 80

C.1 Examples from existing affordance datasets. 81
C.2 (a) Experimental setup. (b) Seen (left) and unseen (right) objects used in the experiments. 82
C.3 Illustrations of 7 tasks in the robot experiments. 83
C.4 Qualitative comparison of affordance prediction and final grasp pose for 3D point clouds in the cluttered scene. LOCATE fails to identify related objects for desired tasks, whereas Aff-Grasp can select the correct object with accurate affordance segmentation and is not affected by cluttered scenes. 85
C.5 Success rates of correct affordance predictions for each individual object from the accuracy and generalization evaluations. The total numbers of trials are 9 and 5, respectively. 85
C.6 The Aff-Grasp framework can perform the handover task by generating grasp poses within the functional parts of objects, and orienting the graspable parts towards the human hand. Green indicates all potential grasps, while red marks the final selected grasp. 86
C.7 Ablation study on hyper-parameters. 88

List of Tables

3.1 Comparison to state-of-the-art methods from relevant tasks on AGD20K dataset. The **best** and second-best results are highlighted in bold and underlined, respectively (↑/↓ means higher/lower is better). 20
3.2 Comparison of learnable parameters and inference time. The inference time is evaluated on a 3090Ti GPU. † denotes the adapted AffCorrs. . 21
3.3 Ablation results of the proposed LOCATE framework. GKT/RKT means global/regional knowledge transfer. $\mathcal{S}$ denotes PartSelect combined with $\mathcal{L}_{cos}$, and $\mathcal{L}_c$ is the concentration loss. 23
3.4 Ablation study on different feature extractors. 25

4.1 Comparison with state of the art on AGD20K dataset. OOAL setting uses 0.22% / 0.21% of the full training data. WSAG denotes weakly-supervised affordance grounding. The **best** and second-best results are highlighted in bold and underlined, respectively. 38
4.2 Comparison on UMD dataset. Fully-supervised methods are trained with 14,823 and 20,874 images with pixel-level labels for seen and unseen split, respectively. In contrast, OOAL setting uses 54 and 76 images, 0.36% of the full training data. 38
4.3 Ablation results of different visual foundation models. 42
4.4 Ablation results of proposed modules. TPL: text prompt learning. MLFF: multi-layer feature fusion. TD: transformer decoder. CTM: CLS-guided mask. 42

5.1 Quantitative comparison on the AED. 54
5.2 Success rates for accuracy evaluation. 58
5.3 Success rates for robustness evaluation. 59
5.4 Success rates for generalization evaluation and inference time for affordance prediction components. 59

5.5 Ablation results of embedder, loss functions, classifiers, and proposed modules. The baseline model is a DeiT III model with a linear layer and binary cross entropy loss. "w/o bg" means that there is no background classifier. "DFI-training only" denotes that the DFI is used during training, and discarded at inference. 60
5.6 Ablation study on DFI on inference efficiency. 61

A.1 Training and test object categories under the unseen setting. 71
A.2 Comparison to state-of-the-art methods on different object scales. The test set is divided into three subsets (Big, Middle and Small) based on the ratio of the mask to the image. The **best** and second-best results are highlighted in bold and underlined, respectively (↑/↓ means higher/lower is better). The symbol * indicates that we reproduce the results using the official code. 71
A.3 Ablation study on the choice of loss functions. Cos† denotes the cosine embedding loss without margin. 72
A.4 Ablation study on the localization map threshold τ. 72
A.5 Ablation study on the PartIoU threshold μ. 73

B.1 Affordance and object classes in the UMD and AGD20K dataset. The number of classes is shown in parentheses. 76
B.2 Object category division in the unseen split of UMD and AGD20K dataset. The number of categories is shown in parentheses. 77
B.3 Ablation study on the number of learnable token p in text prompt learning. 78
B.4 Ablation study on the number of fusion layers j in multi-layer feature fusion. 78
B.5 Ablation study on the number of transformer decoder layers t. 78

C.1 Statistics on the number of images for each object on the AED. 82
C.2 Ablation study on different depth representations in DFI. 87
C.3 Ablation study on different classification embeddings: learnable or CLIP text embeddings. 87

Chapter 1

Introduction

1.1 Motivation and Challenges

In recent years, there has been a growing convergence among computer vision, robotics, and embodied intelligence—fields aiming to build systems that can perceive, reason, and interact with the physical world. A fundamental challenge shared across these domains is enabling agents to go beyond simply recognizing what an object is, and instead understand how it can be used. This capability, known as affordance understanding, involves reasoning about the functional possibilities of objects in context. The term of affordance was originally coined by the psychologist James J. Gibson [43], who described it as the potential actions that arise from the interaction between an agent and its environment. In computer vision and robotics, affordances typically describe how objects or environments support specific actions, offering essential insights into how agents perceive and act in the world. For example, chairs afford sitting, knives afford cutting, and floors afford walking.

Affordances play a key role in bridging perception and action, a connection that is crucial for embodied AI agents. These agents must not only passively perceive their environment but also actively plan and execute complex interactions in dynamic, real-world settings. By understanding affordances, AI systems can infer the functional properties of objects and determine the most appropriate actions to accomplish specific goals. This capability allows agents to assess whether a task is feasible in a given environment and identify the tools or objects needed for execution. The concept of affordance is widely applied in computer vision and robotics because its dynamic, context-sensitive nature aligns well with the demands of real-world tasks such as scene understanding, action recognition, robotic manipulation, and navigation.

With the rapid development of deep learning and the rise of embodied AI, affordance learning has garnered increasing attention in recent years. Despite the notable progress made in this field, existing approaches to affordance learning still face several limitations across the domains of vision, language, and robotic manipulation. In computer vision, early methods [23, 113] have focused on scene affordances, which provide guidance to agents on how to behave within the current environment, but are often limited to a small number of categories. Subsequently, some methods [82, 84, 90, 121, 151] have shifted attention to the object-centric affordances by assigning functional labels to entire objects. While this provides finer granularity, it overlooks the intricacies involved in the actual use and manipulation of objects, which often require precise localization of the object parts. Meanwhile, the development of vision and language foundation models [70, 76, 106, 109] has opened up new possibilities for various tasks, yet few studies have explored their potential for affordance learning. In robotic manipulation, affordances are mainly defined as graspable areas on objects, and often applied to elementary tasks such as pick-and-place, neglecting the complex tool–object interactions that occur in everyday environments. Most importantly, the majority of affordance learning models are often trained in a fully supervised manner, relying heavily on labor-intensive data collection and annotation. This process is inefficient and costly, and the learned models may not generalize well to new tasks or environments.

To address the aforementioned limitations, this thesis presents affordance learning methods that require minimal resources and data, and are applicable across multiple domains, including vision, language, and robotic manipulation. We concentrate on fine-grained affordances, which refer to specific parts or regions of an object that afford particular actions (*e.g.*, the handle of a mug for grasping, the blade of a knife for cutting), as illustrated in Fig. 1.1. Understanding such nuanced affordances is crucial for enabling intelligent agents to interact with objects in a more human-like manner.

To effectively tackle the affordance learning research, we must first understand the key challenges involved. In general, developing a robust and generalizable model for affordance learning faces several challenges, including dataset scarcity, non-unified annotation standards, diverse representations, and complex correspondence, which are discussed in detail below.

Dataset Scarcity. A significant challenge in the field of affordance learning is the lack of large-scale, high-quality datasets. Many datasets [25, 64, 96] are often captured in controlled, laboratory-based tabletop settings, which hinder the scalability and gen-

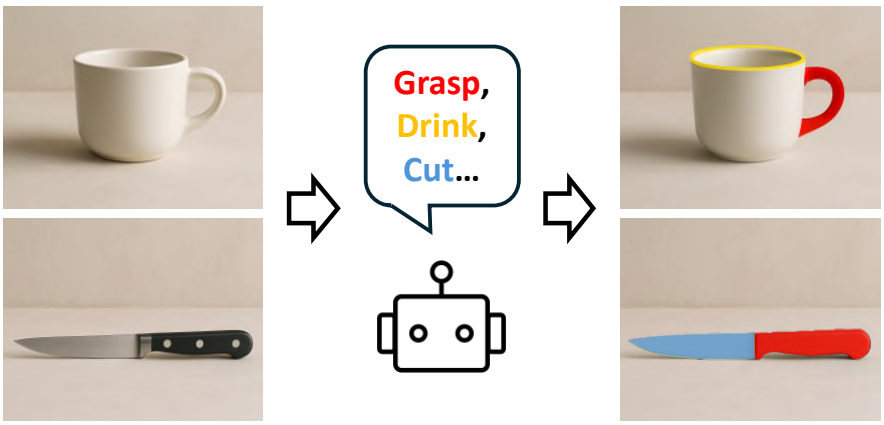


Figure 1.1: Affordance understanding enables an agent to associate specific object regions with distinct actions, such as grasping a mug by its handle, drinking from its rim, or cutting with the blade of a knife.

eralization to new tasks and environments. While some datasets [101, 113, 121] are collected in real-world environments, their annotations are often coarse and limited to a small number of object categories. More recent datasets, such as RGBD-Aff [58], are larger in scale but suffer from low resolution and monotonous backgrounds, making them unsuitable for real-world applications.

Non-unified Annotation Standards. The absence of unified annotation standards is a common issue in affordance learning, posing a challenge to maintain consistency across datasets. In contrast to semantic segmentation, which assigns labels to pixels based on well-defined semantic categories, the definition and labeling of affordances are often subjective and ambiguous, leading to inconsistent annotations. For instance, some datasets may label only the handle of a cup as "holdable", while others label both the cup body and handle.

Diverse Representations. Affordances can be represented in various ways, depending on the task and definition. For example, AffordanceNet [36] framed affordance detection as an instance segmentation task, where the entire object is first detected using a bounding box, followed by the segmentation of object parts with different affordances. Later, Luo et al. [151] and Zhai et al. [84] proposed the one-shot affordance detection task, with the objective of assigning a single affordance to the whole object. To better capture the dynamic nature of affordances, some methods [16, 39, 85, 97] represent affordance as Gaussian heatmaps, indicating potential interaction areas. This diversity in representation highlights the challenge of developing unified models that can effectively accommodate varied affordance definitions and tasks.

Complex Correspondence. The relationship between object parts and affordances is complex due to the many-to-many mapping. A single object part can offer multiple affordances (*e.g.*, a door handle can be used for both pulling and pushing), while different parts of various objects may share the same affordance, despite significant differences in appearance (*e.g.*, chopsticks, spoons, and whisks can all serve the purpose of mixing). This complexity introduces an additional layer of difficulty to the learning process.

1.2 Contributions and Outline

The main contributions of this thesis are the development of resource- and data-efficient affordance learning approaches across multiple domains, including vision, language, and robotic manipulation.

- In Chapter 3, we focus on the task of weakly supervised affordance grounding, proposing a framework that extracts affordance knowledge from human-object interaction images and transfers it to egocentric images in a localized manner. This framework features a novel module that selects affordance-specific cues from human-object interactions and achieves superior performance with far fewer trainable parameters and faster inference speed compared to previous methods.

- To further minimize the reliance on large amounts of training data, in Chapter 4, we propose a new task setting named one-shot open affordance learning, where we explore the potential of vision and language foundation models for data-limited affordance learning. This work introduces a framework that leverages the best of vision and language foundation models for affordance localization. It exhibits good generalization and open-vocabulary recognition capabilities for unseen object and affordance categories, requiring only one labeled sample per object category.

- Finally, in Chapter 5, we present a comprehensive affordance learning system that encompasses data collection, model learning, and deployment to real robots. Specifically, we propose an automated pipeline for collecting and annotating affordance data from egocentric human-object interaction videos; a model that incorporates geometric information to enhance affordance learning; and a manipulation framework that facilitates affordance-oriented robotic manipulation

such as tool grasping and robot-to-human handover. Given a task, the framework identifies the most suitable object, grasps the appropriate part, and utilizes its functional component to complete the task.

The thesis consists of six chapters. Chapter 1 introduces the motivation for our studies, discusses the challenges of affordance research, and presents the main contributions and outline of the thesis. Chapter 2 surveys related work in affordance learning across vision, language, and robotic manipulation. Chapter 3 presents a novel framework for weakly supervised affordance grounding, focusing on learning affordance knowledge from human-object interaction images. Chapter 4 proposes one-shot open affordance learning, which leverages the strengths of vision and language foundation models to facilitate affordance learning under data-limited conditions. Chapter 5 describes a comprehensive affordance learning system, covering data collection, model learning, and real-robot deployment. Finally, Chapter 6 concludes the thesis, discusses the limitations, and suggests potential future research directions.

Chapters 3 to 5 are based on peer-reviewed research papers, all published in leading conference proceedings:

- Chapter 3 is based on "LOCATE: Localize and Transfer Object Parts for Weakly Supervised Affordance Grounding", which has been accepted by the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2023.
- Chapter 4 is based on "One-Shot Open Affordance Learning with Foundation Models", which has been accepted by the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2024.
- Chapter 5 is based on "Learning Precise Affordances from Egocentric Videos for Robotic Manipulation", which has been accepted by the IEEE/CVF International Conference on Computer Vision (ICCV) 2025

Chapter 2

Related Work

In this chapter, we review related work on affordance learning. The chapter is structured into three sections, each focusing on a specific domain of affordance learning: vision, language, and robotic manipulation. Within each domain, we discuss representative research topics and methodologies.

2.1 Affordance Learning in Vision

There are several computer vision tasks relevant to visual affordance learning, such as affordance segmentation, affordance detection, and affordance grounding. While having different outputs and setups, these tasks share the common goal of localizing objects or object parts where specific actions can be performed. Depending on the learning paradigm, existing approaches to affordance learning primarily include fully supervised learning, weakly supervised learning, and learning from human–object interactions.

2.1.1 Fully Supervised Affordance Learning

Initial successes in the field of visual affordance learning were achieved through fully supervised methods [23, 36, 64, 96, 101] that employ convolutional neural networks (CNNs) to learn mapping relations between object functional regions and affordance labels. To enable affordance recognition, several datasets have been proposed. UMD [96] is one of the earliest datasets developed for affordance learning. It covers a variety of household tools annotated with affordance labels such as cutting, grasping, and pounding. Later, Nguyen et al. [101] introduced the IIT-AFF dataset, designed to sup-

port deep learning-based affordance detection. The images in this dataset are collected from diverse scenes and annotated with both bounding boxes and pixel-wise affordance masks. Unlike prior work that primarily focused on object-centric affordances, Chuang et al. [23] proposed the ADE-AFF dataset, which is annotated with scene-centric affordances, enabling affordance reasoning at the scene level. These fully supervised methods, while effective, rely heavily on large-scale annotated datasets, which are often costly and time-consuming to create.

2.1.2 Weakly Supervised Affordance Learning

To alleviate the burden of annotation, much subsequent research has shifted its focus to weak supervision, such as using keypoints [27, 114, 115] or image-level labels [85, 97], which are much easier to obtain compared to pixel-level annotations. In particular, Sawatzky et al. [115] proposed a convolutional network to tackle affordance detection using only a few keypoint annotations. Nagarajan et al. [97] introduced a method that infers object affordance regions by directly learning from human-object interaction videos with only coarse action labels. Subsequently, Luo et al. [85] modified this weakly supervised setting by leveraging action labels from exocentric interaction images. Building on advances in weakly supervised object localization [41, 60, 89, 107, 139, 154] and semantic segmentation [18, 20, 66, 137, 159], most weakly supervised affordance grounding methods rely on class activation mapping (CAM) [156] or its variants [14, 116] to generate activation maps for affordance prediction. However, previous work typically applies CAM at the inference stage, without incorporating explicit supervision during training. In contrast, to provide guidance at the training stage, we propose a method called LOCATE in Chapter 3 that produces localization maps in the forward pass and supervises them in a prototypical learning fashion. This line of research was further developed in subsequent work to improve weakly supervised affordance grounding. Xu et al. [140] extended it by replacing affordance labels with textual descriptions of performed actions, enabling multimodal affordance grounding. Jang et al. [55] leveraged contrastive learning to capture distinctive features of human-object interactions, while Xu and Yadong [142] developed a supervised training pipeline based on pseudo labels generated by advanced foundation models.

2.1.3 Learning Affordance from Human-Object Interactions

An emerging and promising alternative for affordance learning is the automatic extraction of affordance knowledge through the observation of Human-Object Interactions (HOI). The underlying motivation is that humans possess a remarkable ability to imitate diverse interactions and generalize their knowledge to unfamiliar objects. For example, watching someone swing a tennis racket allows us to infer that the handle is the appropriate region to grasp, even without prior experience playing tennis. Therefore, it is paramount to equip intelligent agents with this ability through HOI observations. In general, the objective is to observe where and how humans interact with objects when performing a specific action, and then transfer that knowledge to novel target objects.

Due to the prosperity of HOI research, there are abundant image [12, 13, 49] or video [28–30, 44, 72, 80, 103, 145, 152] datasets that are available for this transfer learning [48]. One line of research [55, 85, 87, 140, 142, 147] focuses on learning affordances from HOI images, which are often efficient to collect and process. For example, methods like [85, 86] use paired exocentric interaction images and egocentric object images to perform affordance grounding. Another line of research [4, 39, 77, 78, 86, 94, 97, 148, 149] explores grounding affordances from HOI videos, which provide rich temporal context and support more detailed modeling of human-object interactions. In particular, methods like [78, 97] leverage annotations from egocentric videos, such as action labels or timestamped narrations, to achieve affordance grounding. In this thesis, we investigate both HOI images and videos for affordance learning. In Chapter 3, we propose LOCATE that utilizes DINO-ViT features to learn affordances from HOI images. In Chapter 5, we introduce an automated affordance data collection pipeline that produces graspable and functional affordance annotations from egocentric HOI videos.

2.2 Affordance Learning with Language

With the advent of vision foundation models (VFMs) [11, 62, 106], vision-language models (VLMs) [56, 70, 109], and large language models (LLMs) [9, 104, 105], there has been a growing interest in integrating visual perception and language understanding for various tasks. These models have demonstrated impressive zero-shot generalization and cross-modal understanding, enabling applications ranging from image retrieval to robotic perception. However, their potential for affordance learning remains

relatively underexplored, particularly in terms of grounding actions and interactions in real-world environments.

To address this gap, OpenAD [102] takes an initial step by leveraging text embeddings for affordance detection, which is achieved through learning a mapping between affordance text labels and point cloud features. WorldAfford [15] extends simple action labels to natural language instructions, enabling more expressive and flexible affordance grounding. Subsequently, AffordanceLLM [108] and RoboPoint [150] explore affordance grounding using multimodal large language models, which demonstrate stronger comprehension of language instructions and improved visual-linguistic alignment. More recently, LASO [74], 3D-AffordanceLLM [22], LMAffordance3D [161], and GREAT [119] have delved into grounding object affordances in 3D space, leveraging both geometric information and language priors to enhance spatial reasoning and interaction prediction.

Although these methods have made notable progress, substantial manual collection of training data is still needed. This reliance on supervision is further compounded by limitations in existing vision-language models, which, while effective at object-level localization, often struggle to capture fine-grained semantics related to object parts and affordances. To address this issue, we propose a novel approach in Chapter 4 to enhance vision-language affordance learning. Our method integrates the part-level semantic correspondences from DINOv2 [106] with the open-vocabulary recognition capabilities of CLIP [109] to associate affordance areas with corresponding action labels. Leveraging these foundation models, our method achieves strong generalization to unseen objects and affordances, requiring only one example per base object category for training. A recent study, UAD [126], also demonstrates the power of foundation models, introducing a pipeline that leverages DINOv2 [106] and GPT-4o [54] to automate affordance labeling.

2.3 Affordance Learning for Robotic Manipulation

Affordance understanding enables robots to interact effectively and intelligently with complex and dynamic environments [2, 146]. Prior work has often leveraged affordances to model relationships between objects, tasks, and manipulations for robotic grasping [3, 63, 125]. Another stream of research focuses on learning affordances from diverse data resources such as human teleoperated play data [8], image pairs [6], and egocentric video datasets [4, 57], which can be deployed on real robots. In contrast

to end-to-end affordance learning, recent studies [51, 52, 92, 110, 120, 127, 129, 132] have explored explicit reasoning about affordances to achieve task-oriented grasping using vision-language models (VLMs) and large language models (LLMs) [11, 104–106, 109, 112]. These methods use LLMs to infer which object parts should be grasped based on task instructions and VLMs to localize the relevant regions. While this pipeline can produce effective and accurate manipulation, the reasoning process often requires prompt engineering and can be time-consuming. By contrast, less attention has been given to affordance-oriented grasping, which aims to derive graspable and functional areas without explicitly specifying the corresponding object parts. For example, when tasked with slicing bread, an affordance-oriented system should deduce that the serrated edge of a bread knife is appropriate for slicing, while the handle is the correct part to grasp. One major barrier to this approach is the lack of large-scale affordance datasets, compounded by the difficulty of unifying existing ones due to inconsistent annotation standards.

To overcome data limitations in affordance-oriented grasping, one line of research such as VRB [4] and Robo-ABC [57] has explored learning affordances from human videos. However, VRB generates only coarse Gaussian heatmaps and requires additional policy learning to deploy the affordance model on real robots. Robo-ABC relies on point correspondences that can be noisy and sensitive to background variations, and may not yield reliable grasp poses. Another line of work [88, 143] manually collects and annotates datasets with affordance labels, then fine-tunes models to enable affordance grounding. Despite making notable progress, these methods focus primarily on the graspable affordances of objects. In contrast, we propose a pipeline in Chapter 5 that addresses both graspable and functional affordances. When combined with a small set of basic pre-recorded motion primitives (as in [34, 35]) and grasp pose detection models [37, 38, 123], our approach enables affordance-oriented robotic manipulation in complex scenarios such as tool grasping and robot-to-human handover.

Chapter 3

Affordance Grounding with Weak Supervision

Humans excel at acquiring knowledge through observation. For example, we can learn to use new tools by watching demonstrations. This skill is fundamental for intelligent systems to interact with the world. A key step to acquire this skill is to identify what part of the object affords each action, which is called affordance grounding. In this chapter, we address this problem and propose a framework called LOCATE that can identify matching object parts across images, to transfer knowledge from images where an object is being used (exocentric images used for learning), to images where the object is inactive (egocentric ones used to test). To this end, we first find interaction areas and extract their feature embeddings. Then we learn to aggregate the embeddings into compact prototypes (human, object part, and background), and select the one representing the object part. Finally, we use the selected prototype to guide affordance grounding. We do this in a weakly supervised manner, learning only from image-level affordance and object labels. Extensive experiments demonstrate that our approach outperforms state-of-the-art methods by a large margin on both seen and unseen objects.

This chapter first introduces the problem of affordance grounding and discusses the challenges associated with weak supervision in Sec. 3.1. The proposed framework LOCATE and its components are then detailed in Sec. 3.2. Extensive experiments are presented in Sec. 3.3 that demonstrate the effectiveness of LOCATE compared to state-of-the-art methods on both seen and unseen objects, and Sec. 3.4 summarizes the findings and contributions of the research and outlines its limitations.

3.1 Introduction

A fundamental skill of humans is learning to interact with objects just by observing someone else performing those interactions [10]. For instance, even if we have never played tennis, we can easily learn where to hold the racket just by looking at a single or few photographs of those interactions. Such learning capabilities are essential for intelligent agents to understand what actions can be performed on a given object. Current visual systems often focus primarily on recognizing *what* objects are in the scene (passive perception), rather than on *how* to use objects to achieve certain functions (active interaction). To this end, a growing number of studies [3, 36, 63, 135] have begun to utilize affordance [43] as a medium to bridge the gap between passive perception and active interaction. In computer vision and robotics [2, 46], affordance typically refers to regions of an object that are available to perform a specific action, *e.g.*, a knife handle affords holding, and its blade affords cutting.

In this chapter, we focus on the task of affordance grounding, *i.e.*, locating the object regions used for a given action. Previous methods [23, 36, 39, 96, 101] have often treated affordance grounding as a fully supervised semantic segmentation task, which requires costly pixel-level annotations. Instead, we follow the more realistic setting [85, 87, 97] where the task is learning object affordances by observing human-object interaction images. That is, given some interaction images, such as those in Fig. 3.1, along with the corresponding label (*e.g.*, "hold"), the aim is to learn affordance grounding on the novel instances of that object. This is a weakly-supervised problem setting where only the image-level labels are given without any per-pixel annotations. Concretely, given several third-person human-object interaction images (exocentric) and one target object image (egocentric), our goal is to extract affordance knowledge and cues from exocentric interactions, and perform affordance grounding in the egocentric view by using only affordance labels.

There are several key challenges underlying the problem of affordance grounding. The first is due to the nature of the supervision, where only image-level affordance labels are given, being a weakly supervised problem. Here, the system needs to automatically reason about affordance regions just from classification labels. Second, human-object interactions often introduce heavy occlusion of object parts by interacting humans. In other words, the object part that the system needs to predict for a particular affordance (*e.g.*, a mug handle for the "holding" affordance) in an exocentric image can often be the part that is occluded (*e.g.*, by hands). Third, interactions

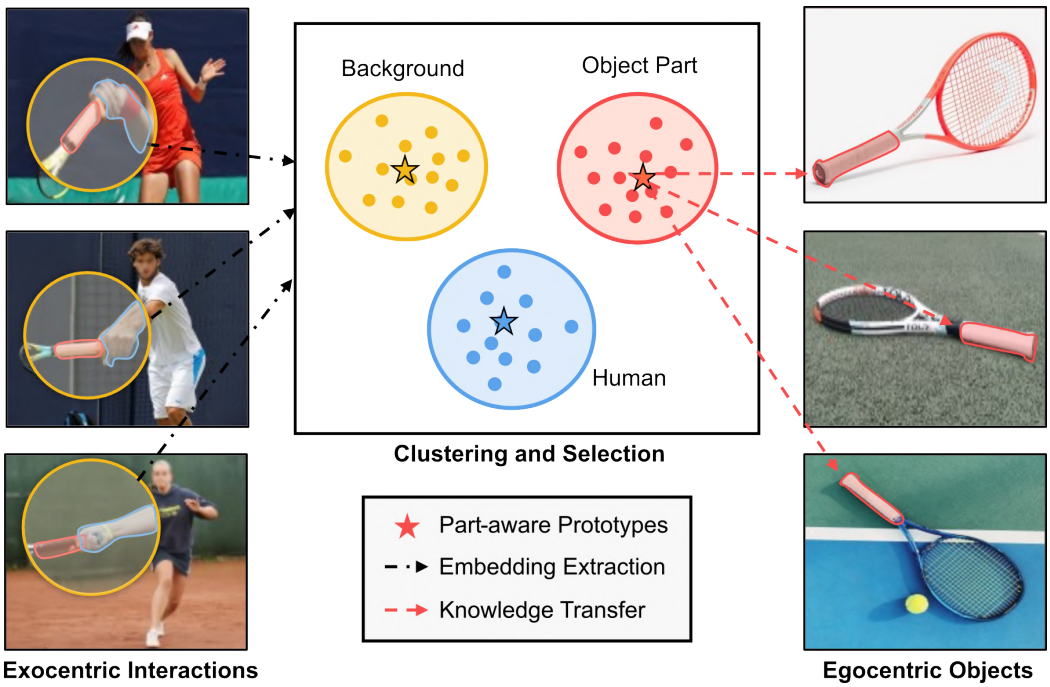


Figure 3.1: Overview of LOCATE. It focuses on regions of exocentric interactions, and clusters visual embeddings into categories such as background, human, and object parts. The prototype of the object-part cluster is then selected to guide egocentric affordance grounding.

are of great diversity. The way humans interact with objects varies across individuals resulting in diverse egocentric interaction images. Lastly, there is a clear domain gap between exocentric and egocentric images where the former have clutter, occlusion etc., and the latter are cleaner (*e.g.*, in Fig. 3.1). This makes affordance knowledge transfer particularly challenging.

In this chapter, we propose a framework called LOCATE that addresses these core challenges by locating the exact object parts involved in the interaction from exocentric images and transferring this knowledge to inactive egocentric images. Refer to Fig. 3.1 for the illustration. Specifically, we first use the class activation mapping (CAM) [156] technique to find the regions of human-object-interaction in exocentric images. Despite being trained for the interaction recognition task, we observe that CAM can generate good localization maps for interaction regions. We then segment this region of interest further into regions corresponding to human, object part, and background. We do this by extracting embeddings and performing k-means clustering to obtain several compact prototypes. Next, we automatically predict which of these prototypes corresponds to the object part relevant to the affordance. To this end, we

propose a module named PartSelect that leverages part-aware features and attention maps from a self-supervised vision transformer (DINO-ViT [11]) to obtain the desired prototype. Finally, we use the object-part prototype as a high-level pseudo supervision to guide egocentric affordance grounding.

Our contributions can be summarized as follows. (1) We propose a framework called LOCATE that extracts affordance knowledge from weakly supervised exocentric human-object interactions, and transfers this knowledge to the egocentric image in a localized manner. (2) We introduce a novel module termed PartSelect to pick affordance-specific cues from human-object interactions. The extracted information is then used as explicit supervision to guide affordance grounding on egocentric images. (3) LOCATE achieves state-of-the-art results with far fewer parameters and faster inference speed than previous methods, and is able to locate accurate affordance region for unseen objects. See Fig. 3.2 for examples of our results and comparison to state-of-the-art.

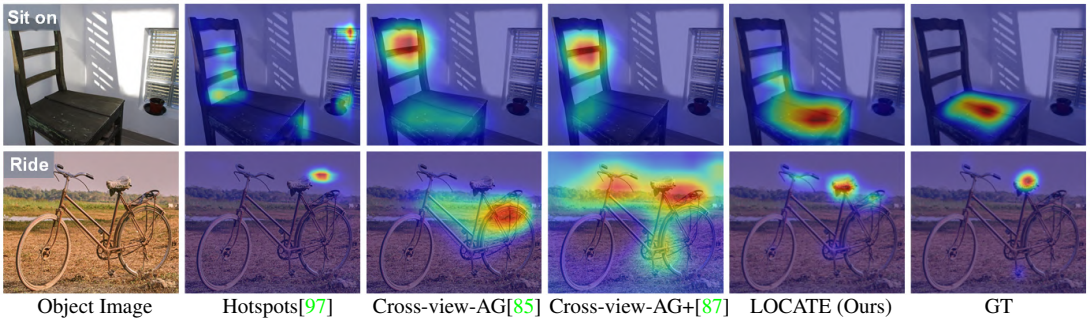


Figure 3.2: State-of-the-art methods in weakly supervised affordance grounding often fail to make accurate predictions for objects with complex structures, *e.g.*, chairs and bicycles. To address this, our model (LOCATE) focuses on localizing and transferring features of object parts, which is able to produce more accurate results.

3.2 Method

Given several exocentric interaction images and one egocentric object image $\{I_{exo}, I_{ego}\}$ ($I_{exo} = \{I_1, I_2, ..., I_N\}$), our goal is to extract affordance-related knowledge from exocentric interactions, and transfer it to egocentric images so that the affordance region can be located even for an inactive object. During training, the only supervision available are image-level affordance labels. In the inference stage, taking an egocentric image and an affordance label as input, the model needs to predict the corresponding affordance region.

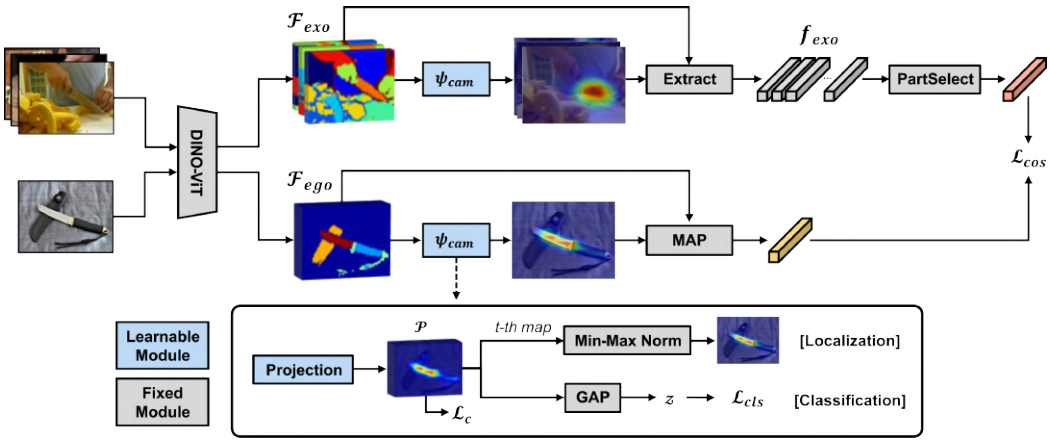


Figure 3.3: Overview of the proposed LOCATE framework. It achieves part-level knowledge transfer in three steps: 1) locating interaction regions with ψ_{cam} (Sec. 3.2.1), 2) object-part embedding selection with PartSelect (Sec. 3.2.2), and 3) part-level knowledge transfer with L_{cos} (Sec. 3.2.3). Details for PartSelect are shown in Fig. 3.4. At test time, only the egocentric branch is maintained.

The core idea of our approach is to exclude distracting information, *e.g.*, human and background, when extracting affordance-specific features from the exocentric view, and perform fine-grained part-level knowledge transfer from exocentric images to egocentric ones. To this end, we set up the framework LOCATE to transfer the knowledge in three steps (See Fig. 3.3). First, we utilize CAM to generate localization maps for exocentric images, and extract corresponding feature embeddings with high activation in the localization maps (Sec. 3.2.1). Then, we propose PartSelect that leverages part-aware deep features to remove irrelevant information while preserving embeddings that can represent affordance cues (Sec. 3.2.2). Finally, we use the output from PartSelect to supervise the egocentric affordance grounding in an explicit manner (Sec. 3.2.3).

3.2.1 Locating Interaction Regions

To determine where neural networks focus on for recognition, we adopt the technique of CAM [156] to generate class-aware localization maps, which has been widely used in weakly supervised tasks. The vanilla CAM generates localization maps as a post-processing step that cannot be guided during training. However, our goal is to extract affordance-specific cues from exocentric images, and use these cues as explicit supervision for the egocentric view. Therefore, in order to obtain localization maps during the training phase, we produce class-specific feature maps instead by adding a class-

aware convolution layer, which has proven to be identical to the generation process in CAM [154]. Specifically, we first extract deep features $\mathcal{F} = \{\mathcal{F}_{exo}, \mathcal{F}_{ego}\} \in \mathbb{R}^{D\times H\times W}$ from images using a network ϕ. In our case, ϕ is a self-supervised vision transformer (DINO-ViT), whose features are part-aware and provide good part-level correspondences. We then generate localization maps $\mathcal{P}$ and classification scores z as follows:

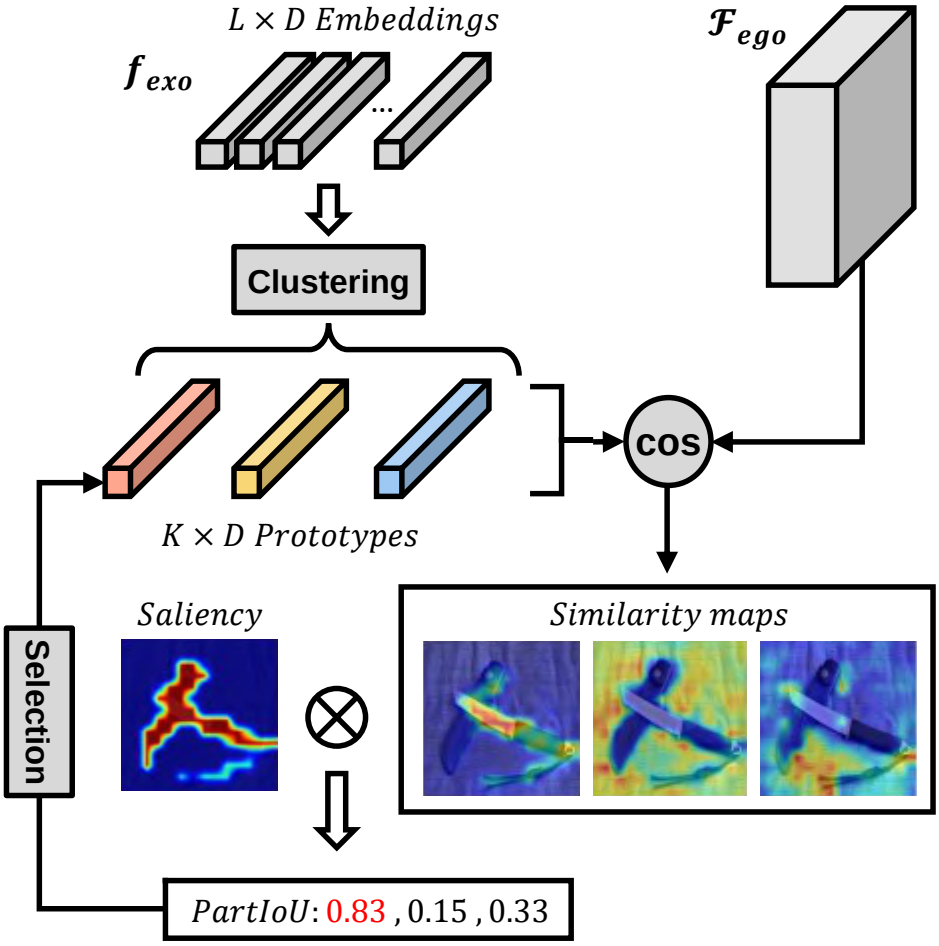
$$\mathcal{P} = \psi_{cam}(\mathcal{F}) \in \mathbb{R}^{C\times H\times W}, \quad z = \text{GAP}(\mathcal{P}) \in \mathbb{R}^{C}, \tag{3.1}$$

where ψ_{cam} starts with a projection layer consisting of a feed-forward layer followed by two convolutions to finetune features for the HOI recognition task, *i.e.*, recognizing actions shown in the exocentric images. Then a 1×1 class-aware convolution layer is added to yield localization maps, converting the number of channels to C, where C denotes the number of total interaction categories. Therefore, each map $\mathcal{P}_c \in \mathbb{R}^{H\times W}$ represents the network activation for the c-th interaction. Next, $\mathcal{P}$ is fed to a global average pooling (GAP) layer to obtain classification scores z, which are used to calculate cross-entropy loss $\mathcal{L}_{cls}$ for optimization.

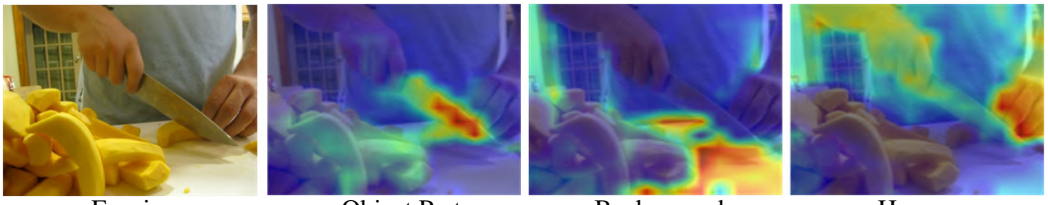
We notice that localization maps for the exocentric branch $\mathcal{P}_{exo}$ concentrate on the interaction areas, *i.e.*, where the action takes place. Since interaction areas carry strong affordance information, we therefore aim to collect embeddings from the high activation regions in exocentric localization maps. Specifically, we first extract the localization map corresponding to ground-truth class, and conduct min-max normalization to constrain activation values to $[0,1]$. After that, we set a threshold τ to control the number of extracted embeddings, therefore embeddings with activation value greater than τ in the localization map will be extracted from deep features. For multiple exocentric images, embeddings are extracted separately to produce $f_{exo} = [f_1, ..., f_N]$, each f_n containing a different number of embeddings. All embeddings are then concatenated together $f_{exo} \in \mathbb{R}^{L\times D}$, where L denotes the number of embeddings.

3.2.2 Object-Part Embedding Selection

In general, interaction areas are composed of human, object part, and background. Our objective is to eliminate the interference information, and purely deliver embeddings representing the object part to guide the egocentric branch. In consequence, we design PartSelect to choose affordance-related embeddings from exocentric branch. PartSelect is illustrated in Fig. 3.4(a). We first perform k-means clustering to get K compact prototypes $p \in \mathbb{R}^{K\times D}$ from extracted exocentric embeddings. Next, we compute the



(a) Illustration of PartSelect.



(b) Exocentric similarity maps.

Figure 3.4: (a) PartSelect picks the object-part prototype through clustering and selection, where ⊗ denotes the calculation for PartIoU score. (b) The similarity maps between prototypes and exocentric features confirm our statement that each prototype represents the object part, background, and human.

cosine distance between each prototype and the egocentric deep features $\mathcal{F}_{ego}$ to get similarity maps $\mathcal{S} \in \mathbb{R}^{K\times H\times W}$:

$$S^{k,u,v} = \frac{p_k \cdot \mathcal{F}_{ego}^{u,v}}{\|p_k\| \left\|\mathcal{F}_{ego}^{u,v}\right\|}. \tag{3.2}$$

Owing to the fine-grained semantic information of DINO-ViT deep features, embeddings of the same object parts bear high similarity.

To distinguish which prototype stands for the object part, we aggregate the self-attention maps from the last layer of the DINO-ViT to generate a saliency map $\mathcal{A} \in \mathbb{R}^{H\times W}$ for the egocentric image. Given the saliency map and similarity maps, we introduce a metric $\gamma \in [0,1]$ termed PartIoU to measure if a prototype carries object part information. The PartIoU for the k-th prototype is defined as follows:

$$\gamma = \frac{1}{2}\frac{\overline{\mathcal{S}_k} \cap \overline{\mathcal{A}}}{\overline{\mathcal{S}_k}} + \frac{1}{2}\frac{\overline{\mathcal{A}}}{\overline{\mathcal{S}_k} \cup \overline{\mathcal{A}}}, \tag{3.3}$$

where $\overline{\mathcal{S}_k}, \overline{\mathcal{A}} \in \{0,1\}^{H\times W}$ are binary masks, we set the threshold as the average of each map to perform binarization. The motivation of PartIoU is fairly straightforward, if $\overline{\mathcal{S}}$ belongs to a portion of $\overline{\mathcal{A}}$, then the intersection of $\overline{\mathcal{S}}$ and $\overline{\mathcal{A}}$ should equal $\overline{\mathcal{S}}$ itself, while the union of the two masks should be identical to $\overline{\mathcal{A}}$. Finally, when the maximum PartIoU among K prototypes is above a threshold μ, PartSelect will output the prototype with the largest PartIoU as the object-part representation. Otherwise, no prototype will be selected for the next step. In Fig. 3.4(b), we visualize the similarity maps between prototypes and exocentric features to demonstrate that the extracted embeddings are clustered into human, object part, and background.

3.2.3 Part-Level Knowledge Transfer

With the help of PartSelect, we find the prototype f_{op} that represents the object part. We then leverage it to perform supervision for egocentric localization maps $\mathcal{P}_{ego}$. Concretely, we first perform masked average pooling (MAP) between the normalized localization map and extracted deep features to aggregate into one embedding:

$$f_{ego} = \frac{\sum_{u=1,v=1}^{W,H} \mathcal{F}_{ego}^{u,v} \mathcal{P}_{ego}^{t,u,v}}{\sum_{i=1,j=1}^{W,H} \mathcal{P}_{ego}^{t,u,v}} \in \mathbb{R}^D, \tag{3.4}$$

where t denotes the ground-truth category. Then, a cosine embedding loss is applied to pull the embedding f_{ego} towards the direction of f_{op}:

$$\mathcal{L}_{cos} = \max(1 - \frac{f_{op} \cdot f_{ego}}{\|f_{op}\| \|f_{ego}\|} - \alpha, 0), \tag{3.5}$$

as the two embeddings come from different domains, we thereby add α as a margin to compensate the domain gap.

In addition, since the affordance region typically denotes a portion of an object, we can thus impose a geometry loss to regulate its distribution. Inspired by the co-part segmentation work [53], we add a concentration loss to encourage egocentric localization maps to form a concentrated and connected component. The concentration loss is formulated as

$$\bar{u}_c = \sum_{u,v} u \cdot \mathcal{P}_{ego}^{c,u,v} / z_k, \; \bar{v}_c = \sum_{u,v} v \cdot \mathcal{P}_{ego}^{c,u,v} / z_k, \tag{3.6}$$

$$\mathcal{L}_c = \sum_{c} \sum_{u,v} \| \langle u, v \rangle - \langle \bar{u}_c, \bar{v}_c \rangle \| \cdot \mathcal{P}_{ego}^{c,u,v} / z_c, \tag{3.7}$$

where $\bar{u}_c$ and $\bar{v}_c$ represents the center of the c-th localization map along axis u, v, and $z_c = \sum_{u,v} \mathcal{P}_{ego}^{c,u,v}$ is a normalization term. The concentration loss forces the high activation regions of the localization maps to be close to the geometric center.

Overall, we train the whole framework in an end-to-end manner, and use the following loss to optimize the model:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_{cos} \mathcal{L}_{cos} + \lambda_c \mathcal{L}_c, \tag{3.8}$$

where λ_{cos}, and λ_c are loss weights that balance the three terms. $\mathcal{L}_{cls}$ stands for the cross-entropy losses from the two branches. At test time, only the ego branch is maintained, taking an affordance label t and an egocentric image as input, the network extracts the t-th localization map as the prediction of affordance region.

3.3 Experiments

3.3.1 Experimental Setting

Dataset and Metrics. We evaluate our method in the Affordance Grounding Dataset (AGD20K) [85], which is the only large-scale image dataset with both exocentric and egocentric views. AGD20K is comprised of 20,061 exocentric images and 3,755 egocentric images, and is annotated with 36 commonly used affordances. Following prior affordance grounding work [39, 97], the ground truth of this dataset initially consists of densely annotated points in corresponding affordance regions, and a Gaussian blur is then applied over each point to get final heatmaps. Moreover, AGD20K can be evaluated in two different settings: 1) In the seen setting, object categories in training and test sets are identical. 2) In the unseen setting, there is no object category intersection

between training and test sets, *e.g.*, the model observes how humans hold a hammer and anticipates where to hold a knife.

As for the metrics, referring to previous affordance grounding work [39, 78, 85, 97], we adopt the commonly used Kullback-Leibler Divergence (KLD), Similarity (SIM), and Normalized Scanpath Saliency (NSS) to evaluate the similarity and correspondence of distributions between ground truth and prediction. Detailed calculation of each metric is shown in the appendix.

Implementation Details. We use the ImageNet [32] pretrained (without supervision) DINO-ViT-S [11] with patch size 16 to generate deep features. In each iteration, N exocentric images along with one egocentric image are taken as input (N is set to 3). Images are first resized to 256×256 and then randomly cropped to 224×224 followed by random horizontal flipping. SGD with learning rate 1e-3, weight decay 5e-4, and batch size 16 is used for parameter optimization. Loss weight coefficients $(\lambda_{cos}, \lambda_c)$ are set to (1, 0.07), and the margin α is set to 0.5. For the first epoch, we warm up the network without $\mathcal{L}_{cos}$, as initial localization maps are not accurate for supervision.

3.3.2 Comparison to State-of-the-Art Methods

State-of-the-Art from Relevant Tasks		Seen			Unseen		
		KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
Weakly Supervised Object Localization	EIL [89]	1.931	0.285	0.522	2.167	0.227	0.330
	SPA [107]	5.528	0.221	0.357	7.425	0.169	0.262
	TS-CAM [41]	1.842	0.260	0.336	2.104	0.201	0.151
Weakly Supervised Affordance Grounding	Hotspots [97]	1.773	0.278	0.615	1.994	0.237	0.577
	Cross-view-AG [85]	1.538	0.334	0.927	1.787	0.285	0.829
	Cross-view-AG+ [87]	1.489	0.342	0.981	1.765	0.279	0.882
	AffCorrs [45]	<u>1.407</u>	<u>0.359</u>	<u>1.026</u>	<u>1.618</u>	<u>0.348</u>	<u>1.021</u>
	LOCATE (Ours)	**1.226**	**0.401**	**1.177**	**1.405**	**0.372**	**1.157**

Table 3.1: Comparison to state-of-the-art methods from relevant tasks on AGD20K dataset. The **best** and <u>second-best</u> results are highlighted in bold and underlined, respectively (↑/↓ means higher/lower is better).

To conduct a comprehensive comparison, we also display the results of state-of-the-art methods from a relevant task, *i.e.*, weakly supervised object localization. As shown in Tab. 3.1, in both seen and unseen settings, LOCATE outperforms all other methods with a considerable margin on all metrics. In particular, compared to the state-of-

the-art affordance grounding method Cross-view-AG+ [87], we improve the KLD by 20.4%, SIM by 33.3%, and NSS by 31.2% in the unseen setting. Cross-view-AG+ is an extended version of Cross-view-AG, but still performs the knowledge transfer based on global pooled embeddings at the image level, thus bringing only minor improvement. AffCorrs [45] is a method that focuses on one-shot part affordance grounding, and it also uses the pretrained DINO-ViT features to do part matching. However, AffCorrs needs a pixel-level mask as a query, and there is no domain gap during the knowledge transfer. To make AffCorrs comparable in our problem setting, we adapt its structure by replacing the query annotated mask with our CAM estimator. The results verify that AffCorrs can also achieve good performance, but still considerably inferior to LOCATE.

Methods	Params (M)	Time (s)
EIL [89]	42.41	0.019
SPA [107]	69.28	0.081
TS-CAM [41]	85.86	0.023
Hotspots [97]	132.64	0.087
Cross-view-AG [85]	120.03	0.023
Cross-view-AG+ [87]	82.27	0.022
AffCorrs†[45]	**6.50**	0.205
LOCATE (Ours)	**6.50**	**0.011**

Table 3.2: Comparison of learnable parameters and inference time. The inference time is evaluated on a 3090Ti GPU. † denotes the adapted AffCorrs.

In Tab. 3.2, we make comparisons in terms of model parameters and inference time. Since our framework is built on a frozen small-sized vision transformer (ViT-small), the training process is efficient with a small number of parameters. For example, LOCATE only has 5.4% of learnable parameters in Cross-view-AG. Additionally, we use a large patch size 16 for the vision transformer, which constrains the input sequence length and greatly reduces computation cost. Therefore, the inference time of LOCATE is also faster than most other methods. By contrast, the adapted AffCorrs runs much slower than LOCATE, as it incorporates an additional CRF post-processing step.

We further visualize the qualitative comparisons with state-of-the-art affordance grounding methods. As shown in Fig. 3.5, we compare our results with Hotspots [97],

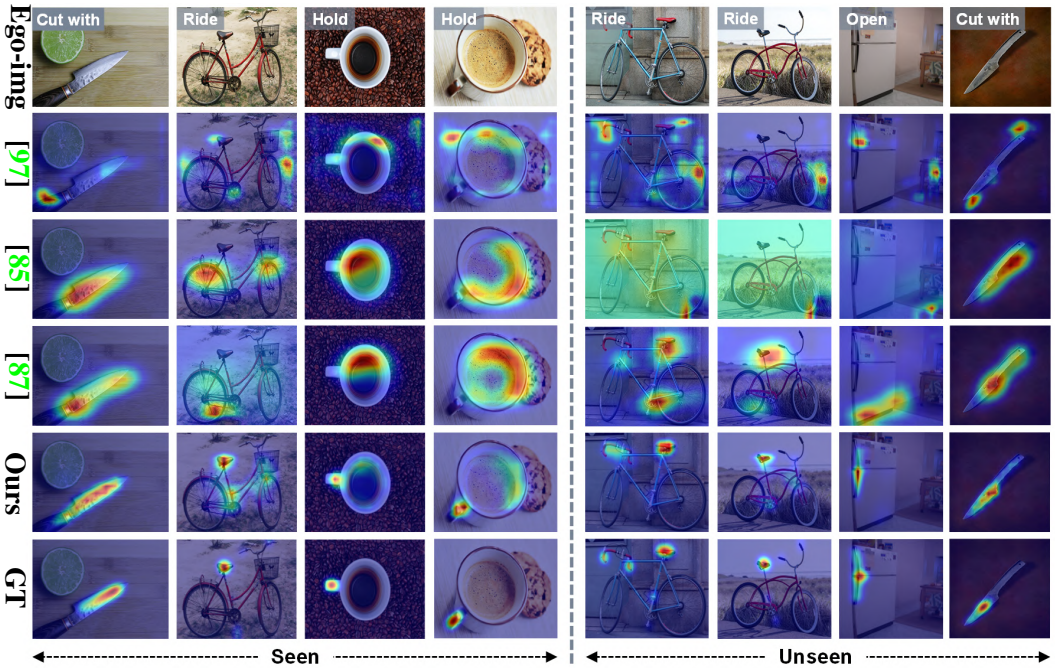


Figure 3.5: Qualitative comparison between our approach and state-of-the-art affordance grounding methods (Hotspots [97], Cross-view-AG [85], and Cross-view-AG+ [87]). For the unseen setting, the objects presented are not exposed in the training set. For example, the model learns where a *motorcycle* can be ridden during training, and then locates rideable area for a *bicycle* at test time.

Cross-view-AG [85] and Cross-view-AG+ [87]. We observe that the proposed LOCATE can make more concentrated and accurate predictions. Especially for complex objects like bicycles and refrigerators, even in the unseen setting, our method can still locate the saddle of bicycles for riding, and the handle of fridges for opening. In comparison, the results of Cross-view-AG for bicycles are quite noisy. More visualization results are in the appendix.

3.3.3 Ablation Study

Knowledge Transfer Manner. We first investigate the impact of knowledge transfer manner. Previous affordance grounding methods [85, 97] simply pull close the global embeddings (produced by global average pooling) of two branches to perform global knowledge transfer (GKT). In contrast, we set up an experiment to implement regional knowledge transfer (RKT), which generates the embeddings via masked average pooling between CAM-produced localization maps and feature maps. The results are shown in Tab. 3.3, regional knowledge transfer (RKT) outperforms global

Method	Seen			Unseen		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
GKT	1.732	0.267	0.810	1.971	0.221	0.626
RKT	1.516	0.320	1.074	1.823	0.259	0.850
+ $\mathcal{L}_c$	1.491	0.326	1.091	1.750	0.274	0.948
+ $\mathcal{S}$	1.236	0.397	1.178	1.439	0.358	1.130
+ $\mathcal{S}$ + $\mathcal{L}_c$	1.226	0.401	1.177	1.405	0.372	1.157

Table 3.3: Ablation results of the proposed LOCATE framework. GKT/RKT means global/regional knowledge transfer. $\mathcal{S}$ denotes PartSelect combined with $\mathcal{L}_{cos}$, and $\mathcal{L}_c$ is the concentration loss.

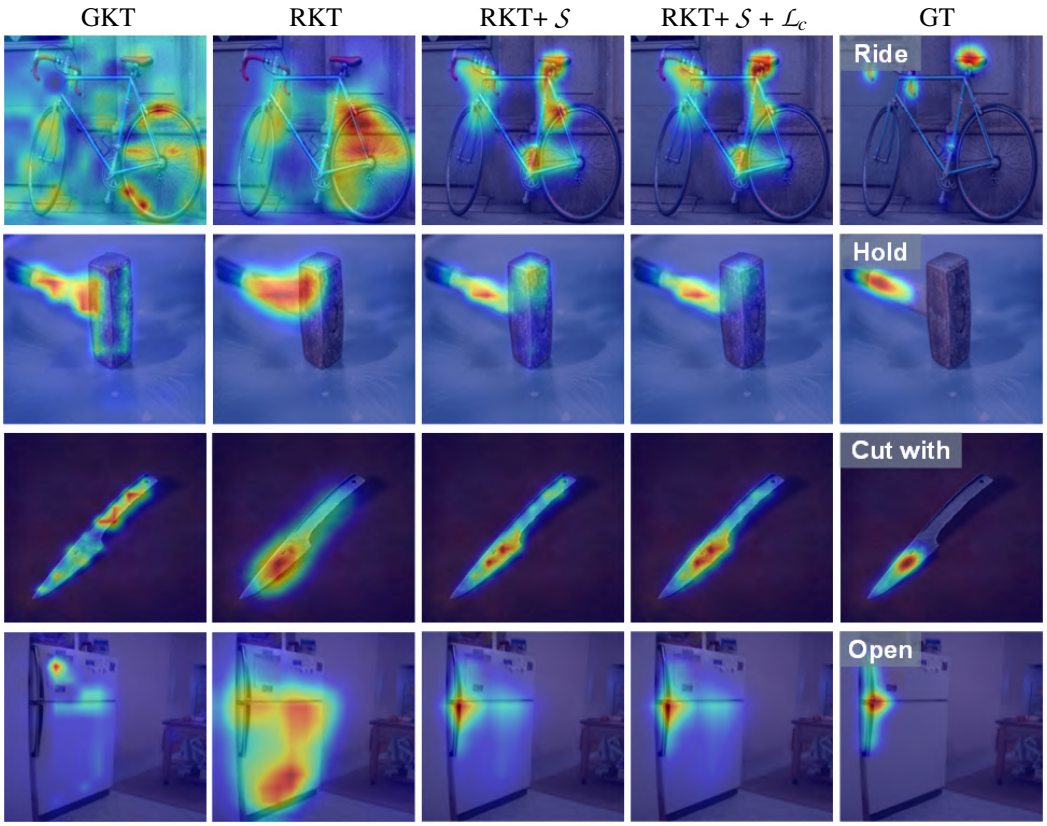
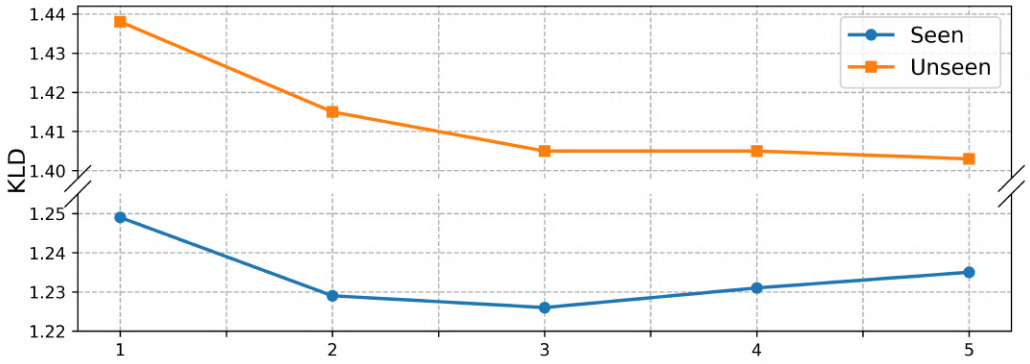


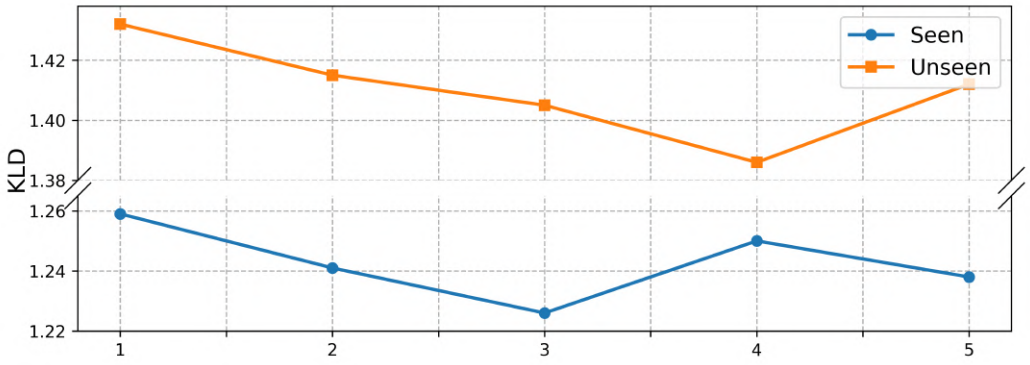
Figure 3.6: Visualization of the qualitative improvements.

knowledge transfer (GKT) on all metrics, demonstrating the effectiveness of filtering irrelevant information.

PartSelect and Concentration Loss. Based on the regional knowledge transfer, we analyze the effect of PartSelect and concentration loss. As shown in Tab. 3.3, directly applying the concentration loss $\mathcal{L}_c$ can only bring marginal improvement. The reason



(a) Number of prototypes N



(b) Number of exocentric images K

Figure 3.7: Ablation study on the number of prototypes and exocentric images (lower is better).

is $\mathcal{L}_c$ can make egocentric predictions more concentrated, but fail to guide it to focus on the right affordance area. Nonetheless, when adding PartSelect and using cosine embedding loss $\mathcal{L}_{cos}$ as explicit supervision, the performance is greatly boosted, which proves the effectiveness of the part-level knowledge transfer scheme. In addition, to check the qualitative improvement, we visualize the affordance grounding results in Fig. 3.6. It is clear that GKT tends to locate the wrong affordance area, while RKT can sometimes find the right region, but only give coarse grounding results. After adding PartSelect, the results become much more part-focused, and the concentration loss further makes the grounding maps more robust.

Number of Prototypes/Exocentric Images. We then explore the impact from the number of prototypes K and exocentric images N. From Fig. 3.7(a), we observe that the model yields the best performance with three prototypes in seen setting, which is consistent with our statement that interaction areas typically consist of human, object part, and background information. While for the unseen setting, $N = 5$ achieves the

	RN-50	ViT-S/16	PartSelect	KLD↓	SIM↑	NSS↑
Seen	✓			1.482	0.334	1.005
		✓		1.491	0.326	1.091
	✓		✓	1.449	0.340	1.021
		✓	✓	1.226	0.401	1.177
Unseen	✓			1.701	0.287	0.962
		✓		1.750	0.274	0.948
	✓		✓	1.707	0.287	0.949
		✓	✓	1.405	0.372	1.157

Table 3.4: Ablation study on different feature extractors.

best results, but improvement is minor. One reason lies in that more prototypes segment objects into more small parts, which boosts the generalization ability. As for the number of exocentric images, we find that more exocentric images can alleviate the impact of interaction diversity and occlusion, thus providing more robust knowledge for the egocentric branch. As shown in Fig. 3.7(b), the model gets largely improved when increasing the number of exocentric images from 1 to 3 in both seen and unseen settings. Finally, we set both N and K to 3.

Different Feature Extractors. In LOCATE, we employ pretrained DINO features based on ViT, which have been proven to encode high-level semantic information [1]. To investigate the impact of different backbones, we conducted experiments with DINO features trained on the ResNet-50 [47]. From the results in Tab. 3.4, we observe that using DINO-ViT features directly (without the proposed PartSelect) can only obtain similar or even inferior results to their ResNet counterpart. After incorporating PartSelect, the results of both backbone features can be improved under the seen setting, but ViT features show better potential in enhancing the performance due to their part-aware property. In the unseen setting, PartSelect does not yield improvement for ResNet-based features, while ViT features obtain consistent gains.

3.4 Conclusion and Limitations

In this chapter, we propose LOCATE to address the weakly supervised affordance grounding task by observing human-object interaction images. Specifically, we first localize where the interaction happens for the exocentric interactions, and then design a module called PartSelect to pick the affordance-specific information from the interaction regions. Finally, we transfer the learned knowledge to the egocentric view to perform affordance grounding with only image-level affordance labels.

While LOCATE achieves state-of-the-art results with far fewer parameters and faster inference speed, several limitations remain: (1) The performance on small objects with the "holding" affordance is generally poor, as the corresponding object parts in exocentric images are often fully occluded, such as the handle of a knife. This issue could be potentially addressed by learning from human-object interaction videos, which can better mitigate the effects of interaction diversity and occlusion. (2) LOCATE identifies matching object parts based on DINO-ViT features, which can be sensitive to factors such as texture, shadow, and lighting, leading to inconsistent clustering results. To enhance the clustering stability, future work could introduce random crop and flip augmentations, as suggested in [1]. (3) Egocentric images in the AGD20K dataset typically focus on a single instance and are primarily object-centric, while real-life images can be more cluttered and complex. (4) LOCATE relies on the class activation mapping (CAM) to localize object parts. However, CAM often falls short in localizing the accurate interaction areas for different actions, and it tends to produce identical heatmaps for different affordances. As a result, LOCATE often fails to ground different affordances for the same object.

Furthermore, the LOCATE pipeline involves several hyperparameters, including various thresholds, the number of exocentric images, and the number of clustered prototypes. While the performance of the pipeline is generally robust to variations in these hyperparameters, we recommend specific adjustments based on the characteristics of the dataset. In particular, when working with noisy data, increasing the number of exocentric images can help improve stability and generalization. Similarly, employing a larger number of prototypes is beneficial for more fine-grained affordance grounding, enabling better differentiation between subtle interaction cues. Regarding the clustering approach, we adopt k-means clustering for its simplicity and computational efficiency, but more advanced methods can be explored to further enhance the performance.

Chapter 4

Data-Limited Vision-Language Affordance Learning

We introduce One-shot Open Affordance Learning (OOAL), where a model is trained with just one example per base object category, but is expected to identify novel objects and affordances. While vision-language models excel at recognizing novel objects and scenes, they often struggle to understand finer levels of granularity such as affordances. To handle this issue, we conduct a comprehensive analysis of existing foundation models, to explore their inherent understanding of affordances and assess the potential for data-limited affordance learning. We then propose a vision-language framework with simple and effective designs that boost the alignment between visual features and affordance text embeddings. Experiments on two affordance segmentation benchmarks show that the proposed method outperforms state-of-the-art models with less than 1% of the full training data, and exhibits reasonable generalization capability on unseen objects and affordances.

This chapter begins by highlighting the limitations of current vision-language models in understanding fine-grained affordances and emphasizes the necessity for data-efficient affordance learning approaches in Sec. 4.1. Subsequently, Sec. 4.2 details the problem setting, performs analyses of existing foundation models, and introduces the proposed vision-language framework. Sec. 4.3 provides extensive evaluations on two affordance segmentation benchmarks, demonstrating the efficiency and generalization ability of the framework. Finally, Sec. 4.4 summarizes the key contributions and findings of the research, and discusses its limitations.

4.1 Introduction

Affordances are the potential "action possibilities" regions of an object [43, 46], which play a pivotal role in various applications, including robotic learning [42, 63, 98], scene understanding [23, 83, 113], and human-object interaction [49, 97]. In particular, affordance is crucial for embodied intelligence, since it facilitates agents' understanding of the associations between objects, actions, and effects in dynamic environments, thus bridging the gap between passive perception and active interaction [26, 93].

Learning to recognize object affordances across a variety of scenarios is challenging, since different objects can vary significantly in appearance, shape, and size, yet have the same functionality. For instance, a chef's knife and a pair of office scissors share common affordances of cutting and holding, but their blades and handles look different.

A large portion of the work [23, 33, 36, 95, 100, 101] has focused on learning a mapping between visual features and affordance labels, utilizing diverse resources as inputs, such as 2D images, RGB-D data, and 3D point clouds. This mapping can be established through a labeled dataset with predefined objects and affordances. However, large-scale affordance datasets are scarce, and most of them have a small number of object categories, making it difficult to apply the learned mapping to novel objects and scenes. To reduce the reliance on costly annotation, some recent studies perform affordance learning from sparse key points [27, 114, 115], videos of humans in action [39, 78, 97], or human-object interaction images [67, 85]. While alleviating the need for dense pixel labeling, these methods still require a large amount of training data. In addition, they often struggle to generalize to unseen objects and cannot identify novel affordances.

To tackle the above limitations, we are interested in learning an affordance model that does not rely on extensive datasets, and can comprehend unseen object and affordance classes. For example, after a model is trained with the knowledge that scissor blades afford cutting, it should generalize to related objects such as knives and axes, inferring that their blades can cut objects too. Moreover, the model should be able to reason about semantically similar vocabularies, *e.g.*, "hold" and "grasp", "cut" and "slice", instead of knowing only predefined affordance categories.

In this chapter, we target the extreme case of using merely one example from each base object category and term this research problem as One-shot Open Affordance Learning (OOAL), where the model is trained with very little data, and is expected to

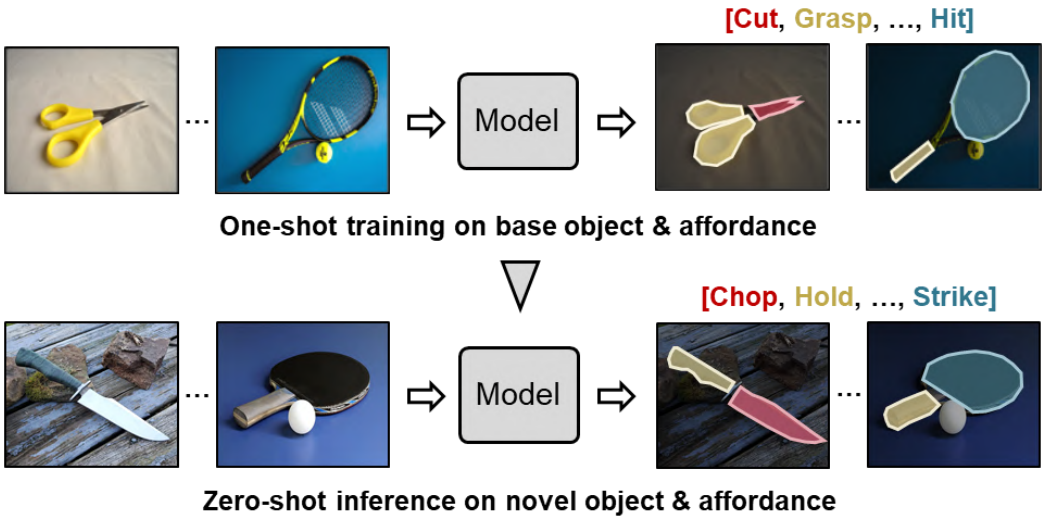


Figure 4.1: The pipeline of one-shot open affordance learning. It uses one image per base object for training, and can perform zero-shot inference on novel objects and affordances.

recognize unseen objects and affordances during inference. The illustration of OOAL pipeline is shown in Fig. 4.1. Compared with the typical affordance learning that requires numerous training samples and can only reason within a closed affordance vocabulary, OOAL alleviates the need of large-scale datasets and broadens the scope of inference.

To this end, we note that foundation Vision-Language Models (VLMs) can be a potential solution, which have recently emerged as powerful tools for a wide array of computer vision tasks. The open vocabulary nature of these VLMs like CLIP [109] that are trained on a large corpus of image-text data enables reasoning of previously unseen objects, scenes, and concepts. However, we observe that these models often fail to understand nuanced vocabularies such as affordances or object parts. One hypothesis is that object parts and affordances appear much less frequently in image captions compared with objects. Therefore, the following question naturally arises: *Can we teach foundation models to comprehend more subtle, fine-grained aspects of objects, such as affordances, with very few examples?* In this way, the generalization capability of foundation models can be inherited with minimum annotation effort.

To achieve this, we first conduct a thorough analysis of several representative foundation models. The objective is to delve into their inherent understanding of affordances, and figure out what visual representation is suitable for data-limited affordance learning. Based on the analysis, we then build a learning architecture and propose sev-

eral methods, including text prompt learning, multi-layer feature fusion, and a CLS-token-guided transformer decoder, that can facilitate the alignment between visual representation and affordance text embeddings. Lastly, we select a dense prediction task, affordance segmentation, for evaluation and comparison with a variety of state-of-the-art models, where we find that our methods can achieve higher performance with less than 1% of the complete training data.

Overall, our contributions can be summarized as follows: (1) We introduce the problem of OOAL, aiming to develop a robust affordance model that can generalize to novel object and affordance categories without the need of massive training data. (2) We conduct a comprehensive analysis of existing foundation models to explore their potential for OOAL. Following the analysis, we build a learning architecture with vision-language foundation models, and design several methods to improve the alignment between visual features and affordance text labels. (3) We implement extensive experiments with two affordance segmentation datasets to demonstrate the effectiveness of our learning pipeline, and observe noticeable gains over baselines with strong generalization capability.

4.2 One-Shot Open Affordance Learning

4.2.1 Problem Setting

One-shot Open Affordance Learning (OOAL) aims to learn a model to predict affordance with one example per base object class and can generalize to novel object classes. In this chapter, we focus on the dense prediction task of affordance segmentation. Specifically, objects are first divided into N_b base classes and N_o novel classes without intersection. The model receives only N_b samples during training, one for each base object category, which is a pair of image $I \in \mathbb{R}^{H\times W\times 3}$ and pixel-wise affordance annotation $M \in \mathbb{R}^{H\times W\times N}$ (N is the number of affordance categories in the dataset). After training, evaluation is performed on the combination of base and novel object categories to measure the generalization ability of the model. Also, affordance labels can be replaced with novel vocabularies that share similar semantics, such as “chop”, “slice”, and “trim” to represent affordance akin to “cut”.

It is worth noting that OOAL is different from one-shot semantic segmentation (OSSS) [117] and one-shot affordance detection (OS-AD) [84]. Both OSSS and OS-AD receive one-shot sample during training. However, the sample keeps changing

in each iteration, so the model has access to a large set of image-mask pairs. Additionally, a support image is required at inference to provide prior information. In comparison, OOAL performs one-shot training and zero-shot inference, which poses additional challenges. The model needs to generalize to previously unseen objects, necessitating the ability to understand and recognize semantic relationships between seen and unseen classes with very limited data.

The field of computer vision has recently witnessed a surge in the prevalence of large foundational models, such as CLIP [109], Segment Anything [62], and DINO [11, 106]. These models exhibit strong zero-shot generalization capabilities for several computer vision tasks, making them seem like a great option to tackle the problem of OOAL. To this end, we perform analysis of several existing foundation models which we split into three parts: ❶ Do current vision-language foundation models and their variants have the ability to detect affordances via affordance/part-based prompting? ❷ Can the features of visual foundation models discriminate affordance regions in images? and ❸ Can these models generalize affordance recognition to novel objects and perform well in the low-shot setting?

4.2.2 Analysis of Foundation Models

Driven by question ❶, we select four representative models, *i.e.*, the vanilla CLIP, a CLIP-based explainability method CLIP Surgery [73], an open-vocabulary segmentation method CAT-Seg [21], and an open-vocabulary detection method GroundingDINO [79]. For vanilla CLIP, we employ the method proposed in MaskCLIP [157] that directly extracts dense predictions without fine-tuning. We use the text prompt template of "somewhere to [*affordance*]" to query visual features to find corresponding areas. As illustrated in Fig. 4.2, we note that most models cannot understand affordance well, except the detection model GroundingDINO, but its predictions mainly focus on the whole object rather than parts. As for dense prediction models, CAT-Seg often recognizes affordance regions as background, and CLIP gives high activation on both foreground and background. In comparison, CLIP Surgery fails to localize the "holding" area for a knife, but manages to associate the phrase "sit on" with a chair. Furthermore, even when the affordance text is replaced with corresponding object parts, predictions from CLIP and GroundingDINO remain biased toward objects, while CLIP Surgery and CAT-Seg tend to activate the wrong parts. This is consistent with recent findings [122, 133] that CLIP has limited part recognition ability.

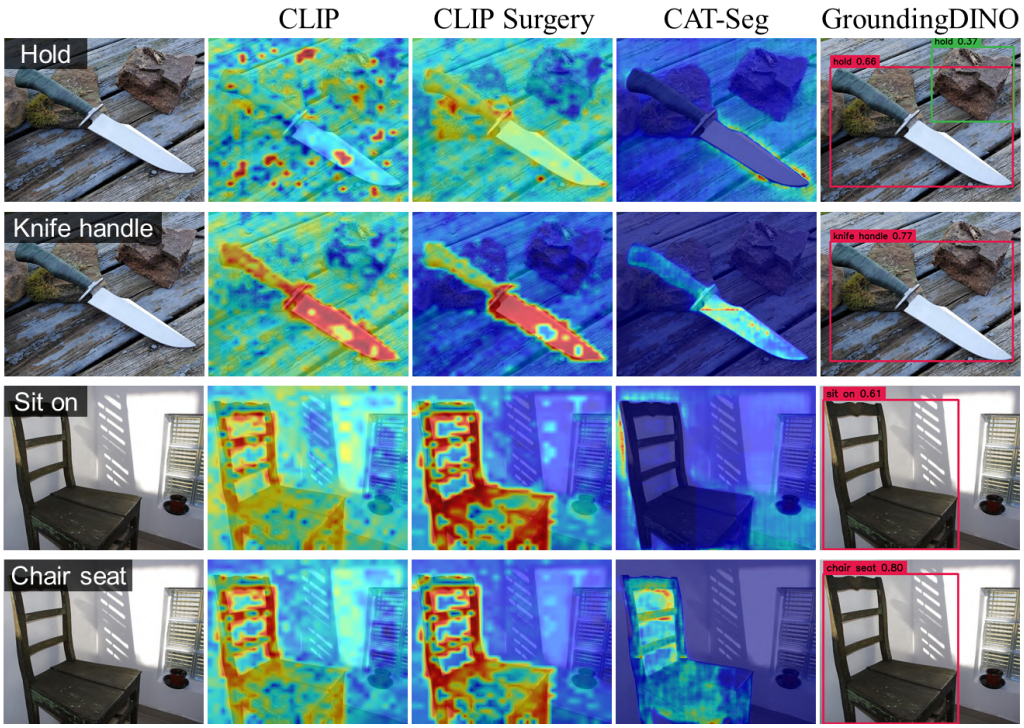


Figure 4.2: **Analysis of vision-language foundation models** on text-based affordance grounding. The 1st and 3rd rows use affordance texts as input queries, and the 2nd and 4th rows use corresponding object parts as input text queries. Visualizations show that these models have limited ability to recognize fine-grained affordances and object parts.

To answer questions ❷ and ❸, we consider two essential characteristics of a good affordance model in the low-shot setting: (1) Part-aware representation. The visual representation should exhibit awareness of object parts, given that affordance often denotes small and fine-grained regions, *e.g.*, a bicycle saddle to sit on or a knife handle to hold. (2) Part-level semantic correspondence. This property is critical for generalization, since the model requires the understanding of semantic relations to make reasonable predictions on novel objects. In addition, good correspondence proves advantages in scenarios with limited data, as the model can be more robust to intra-class recognition, and less susceptible to changes in appearance. We then analyze the features from three representative and powerful visual foundation models, *i.e.*, vision-language contrastive learning CLIP, fully-supervised learning DeiT III [128], and self-supervised learning DINOv2. First, we perform the principal component analysis (PCA) on the extracted patch features of each model to investigate the part awareness. Visualization of PCA components in the top row of Fig. 4.3 shows that all three models have part-

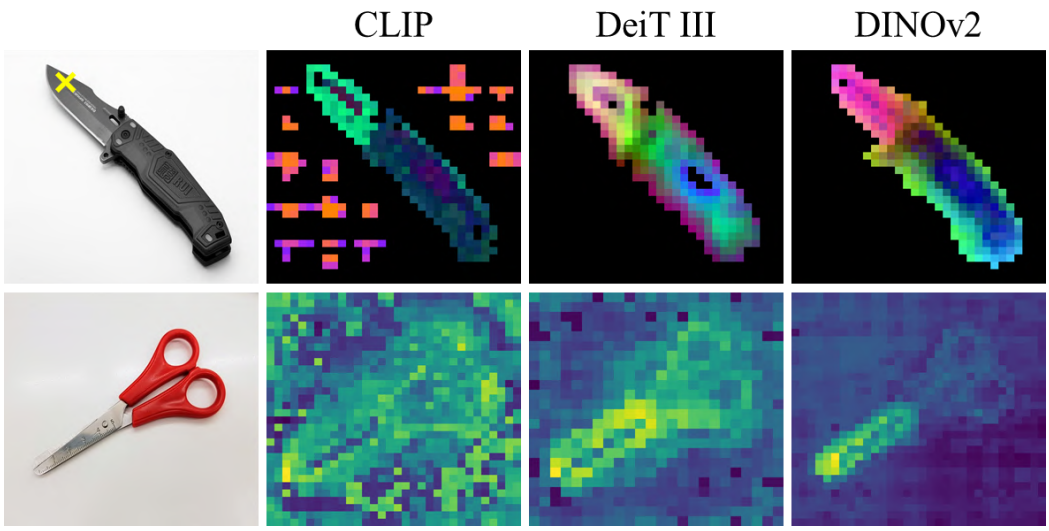


Figure 4.3: **Analysis of visual foundation models** on affordance learning. Top row: visualizations of PCA components. We extract features from the last layer of each model and perform PCA on them. Bottom row: feature similarity maps between the yellow mark on the knife blade and the image of scissors. Qualitative results show that DINOv2 has clearer part-aware representations and better part-level semantic correspondence.

aware features to some extent, yet CLIP cannot well distinguish the background, and features of DeiT III are not discriminative enough for different parts. Next, we choose a different object that has equivalent affordances, *i.e.*, knife and scissors, to assess the semantic correspondence. The bottom row of Fig. 4.3 shows the feature similarity maps computed as the cosine similarity between one patch representation on the knife blade and an image of scissors. It is obvious that DINOv2 shows finer correspondence between blades of knife and scissors. By contrast, CLIP produces messy correspondences in both foreground and background, and feature correspondences of DeiT III are only discriminative at the object level, but not specific to the affordance part region. From the above analysis, we conclude that DINOv2 is well suited for affordance learning due to its fine-grained part-aware representation and superior part-level semantic correspondence. Quantitative comparisons are shown in Sec. 4.3.5.

4.2.3 Motivation and Method

Through a systematic analysis, we identify DINOv2 as a powerful tool for addressing the OOAL problem. However, there are still fundamental issues that hinder perfor-

mance in this challenging setting. The first is that DINOv2 is a vision-only model, and lacks the ability to identify unseen affordances. One potential solution involves integrating a text encoder like CLIP, but it is recognized that the input text is sensitive to prompts. This is particularly problematic in the case of affordances, which combine both an object and a verb, making manual prompt design a complex task. The second issue is that while features of DINOv2 are part-oriented, the level of granularity varies across layers. Determining the appropriate granularity level is crucial when handling affordances associated with diverse objects. The third issue arises due to the absence of alignment between the DINOv2 vision encoder and CLIP text encoder, as they are trained separately and independently of each other. Building upon these observations, we establish a vision-language framework based on DINOv2 and CLIP, and propose three modules to resolve each of the three fundamental bottlenecks mentioned above.

In this section, we first describe the overview of our proposed learning framework that builds on the powerful foundation models. Then, we elaborate on the three proposed designs that help in the challenging OOAL problem. Finally, we discuss the framework's capability to identify unseen objects and affordances at inference.

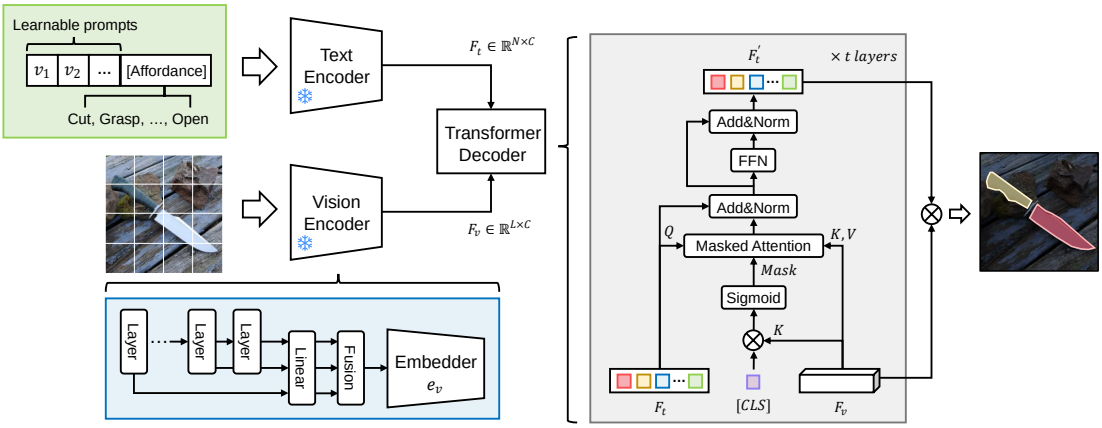


Figure 4.4: Proposed learning framework for OOAL. Our designs are highlighted in three color blocks, which are text prompt learning, multi-layer feature fusion, and CLS-guided transformer decoder. [CLS] denotes the CLS token of the vision encoder.

Overview. The proposed learning framework is presented in Fig. 4.4, which consists of a vision encoder, a text encoder, and a transformer decoder. First, the pretrained vision encoder DINOv2 is used to extract dense patch embeddings $\hat{F}_v \in \mathbb{R}^{L\times C_v}$, where L is the number of tokens or patches. Then, affordance labels are processed by the CLIP text encoder to obtain text embeddings $F_t \in \mathbb{R}^{N\times C}$. To cope with inconsistent dimensions between visual and text embeddings, an embedder $e_v : \mathbb{R}^{C_v} \rightarrow \mathbb{R}^{C}$ with a single MLP

layer is employed to transform $\hat{F}_v$ to F_v. In the end, the lightweight transformer decoder takes both visual and text embeddings as input, and outputs the affordance prediction.

Text Prompt Learning. Manually designing prompts for affordances can be a complicated work, especially considering that CLIP has difficulty in recognizing affordance (see Fig. 4.2). Thus, we adopt the Context Optimization (CoOp) [158] method to introduce automatic text prompt learning. Instead of finetuning the CLIP text encoder, the inclusion of learnable prompts is an effective strategy that can alleviate overfitting and retain the inherent text recognition ability of CLIP. Specifically, p randomly initialized learnable context vectors $\{v_1, v_2, ..., v_p\}$ are inserted in front of the text CLS token, and are shared for all affordance classes.

Multi-Layer Feature Fusion. Different layers of DINOv2 features often exhibit different levels of granularity [1]. Since affordance may correspond to multiple parts of an object, a diverse set of granularities can be beneficial. For this purpose, we aggregate the features of the last j layers. Each layer of features is first processed by a linear projection, and then all features are linearly combined with a weighted summation:

$$\hat{F}_v = \sum_{i=1}^{j} \alpha_i \cdot \phi(F_{n-i+1}), \quad \alpha_1 + \alpha_2 + ... + \alpha_j = 1, \tag{4.1}$$

where F_n denotes the last layer, α is a learnable parameter that controls the fusion ratio of each layer, and ϕ indicates the linear transformation. This straightforward fusion scheme enables adaptive selection among different granularity levels, allowing the model to handle affordance recognition across diverse scenarios.

CLS-Guided Transformer Decoder. To deal with the lack of alignment between visual and text features, we propose a lightweight transformer decoder that applies a masked cross-attention mechanism to promote the mutual communication between two branches. Since the [CLS] token of a foundation model is used in the computation of objective function, it often carries rich prior information of the whole image, such as salient objects or regions. Consequently, we utilize the [CLS] token to produce a guidance mask that constrains the cross-attention within a foreground region.

The decoder receives three inputs, *i.e.*, text embeddings F_t, visual features F_v, and the [CLS] token L_{cls}. Firstly, linear transformations are performed to yield query, key, and value:

$$Q = \phi_q(F_t),\ K = \phi_k(F_v),\ V = \phi_v(F_v). \tag{4.2}$$

Here we use text embeddings as query, and visual features as key and value, allowing the model to focus on the update of text embeddings by retrieving relevant visual

information that corresponds to the affordance text. Next, the CLS-guided mask is calculated between the [CLS] token and key via matrix multiplication:

$$M_{cls} = \text{sigmoid}(\frac{\phi_c(L_{cls})K^T}{\sqrt{d_k}}), \tag{4.3}$$

where d_k is a scaling factor that equals the dimension of the keys. The masked cross-attention is then computed as:

$$\hat{F}_t = \text{softmax}(QK^T/\sqrt{d_k}) \cdot M_{cls}V + F_t. \tag{4.4}$$

After that, the updated text embeddings $F_t^{'}$ are obtained by sending $\hat{F}_t$ through a feed-forward network (FFN) with a residual connection. The decoder comprises t layers of transformers, and the ultimate prediction is generated by performing matrix product between the output of the last transform layer and original visual features F_v, thereby ensuring the maximum retention of part-aware representations from DINOv2. Lastly, binary cross entropy is employed as loss function to optimize parameters of linear layers, embedder, and decoder.

Inference on Unseen Objects and Affordances. During the training process, the decoder learns to establish an alignment between visual features and affordance text embeddings. When encountering a novel object at inference, the aligned affordance text embeddings can locate corresponding object regions, leveraging the part-level semantic correspondence property inherent in DINOv2. Similarly, as the model processes unseen affordance text inputs, the generated text embeddings can also retrieve the aligned visual features, which are based on the semantic similarities to the base affordances seen in the training.

4.3 Experiments

4.3.1 Datasets

We choose two typical datasets, AGD20K [85] and UMD part affordance [96], both of which include a large number of object categories that help in the evaluation of novel objects. AGD20K is a large-scale affordance grounding dataset with 36 affordances and 50 objects, containing 23,816 images from exocentric and egocentric views. It aims to learn affordance from human-object interaction images, and perform affordance localization on egocentric images. As it is a dataset for weakly-supervised

learning, images in the training set only have image-level labels. Therefore we manually annotate 50 randomly selected egocentric images from each object category for training. AGD20K also has two train-test splits for seen and unseen settings, and we follow their splits to evaluate the performance. Note that AGD20K uses sparse annotation, where ground truth consists of keypoints within affordance areas, and then a gaussian kernel is applied over each point to produce dense annotation.

UMD dataset consists of 28,843 RGB-D images with 7 affordances and 105 kitchen, workshop, and gardening tools. It has two train-test splits termed category split and novel split. We use the category split to evaluate base object categories and novel split to evaluate performance on novel object classes. Due to its small number of object categories, we take one example from each base object instance to form the training set. Specific affordance categories and object class splits can be found in the appendix.

4.3.2 Implementation Details

Experiments are implemented on two GeForce RTX 3090 GPUs. All visual foundation models use the same base-sized vision transformer (ViT-base). We train the model using SGD optimizer with learning rate 0.01 for 20k iterations. For experiments on AGD20K, images are first resized to 256 × 256 and randomly cropped to 224 × 224 with horizontal flipping. Experiments for UMD dataset are conducted on the opensource toolbox MMSegmentation [24] with the default training setting. The hyperparameters p, j, and t are set to 8, 3, and 2, respectively.

Following previous work, we adopt the commonly used Kullback-Leibler Divergence (KLD), Similarity (SIM), and Normalized Scanpath Saliency (NSS) metrics to evaluate the results on AGD20K. For UMD dataset, we use the metric of mean intersection-over-union (mIoU), and also incorporate the harmonic mIoU as a balanced measure that accounts for both seen and unseen settings.

4.3.3 Comparison to State-of-the-Art Methods

AGD20K dataset is benchmarked with weakly supervised affordance grounding approaches (WSAG), which use image-level object and affordance labels to do affordance segmentation. Note that results from WSAG methods are not directly comparable to our setting, as training labels are different. Despite using only image-level labels, the training data required are more than 460 times of ours. The results in Tab. 4.1 demonstrate that our performance exceeds all WSAG counterparts in an

Task	Training Data seen / unseen split	Method	Seen			Unseen		
			KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
WSAG	23,083 / 15,543 images image-level labels	Hotspots [97]	1.773	0.278	0.615	1.994	0.237	0.577
		Cross-view-AG [85]	1.538	0.334	0.927	1.787	0.285	0.829
		Cross-view-AG+ [87]	1.489	0.342	0.981	1.765	0.279	0.882
		LOCATE [67]	<u>1.226</u>	<u>0.401</u>	<u>1.177</u>	<u>1.405</u>	<u>0.372</u>	<u>1.157</u>
OOAL	50 / 33 images keypoint labels	MaskCLIP [157]	5.752	0.169	0.041	6.052	0.152	0.047
		SAN [141]	1.435	0.357	0.941	1.580	0.351	1.022
		ZegCLIP [160]	1.413	0.387	1.001	1.552	0.361	1.042
		Ours	**0.740**	**0.577**	**1.745**	**1.070**	**0.461**	**1.503**

Table 4.1: Comparison with state of the art on AGD20K dataset. OOAL setting uses 0.22% / 0.21% of the full training data. WSAG denotes weakly-supervised affordance grounding. The **best** and <u>second-best</u> results are highlighted in bold and underlined, respectively.

Setting	Method	Seen	Unseen	hIoU
Fully Supervised	DeepLabV3+ [17]	70.5	57.5	63.3
	SegFormer [136]	<u>74.6</u>	57.7	65.0
	PSPNet [155]	72.0	**60.8**	<u>66.0</u>
OOAL	PSPNet [155]	56.7	46.6	51.1
	DeepLabV3+ [17]	56.8	48.4	52.3
	SegFormer [136]	64.6	51.4	57.3
	MaskCLIP [157]	4.25	4.24	4.25
	SAN [141]	45.1	32.2	37.5
	ZegCLIP [160]	47.4	36.0	40.9
	Ours	**74.6**	<u>59.7</u>	**66.4**

Table 4.2: Comparison on UMD dataset. Fully-supervised methods are trained with 14,823 and 20,874 images with pixel-level labels for seen and unseen split, respectively. In contrast, OOAL setting uses 54 and 76 images, 0.36% of the full training data.

easy and realistic setting. We also benchmark open-vocabulary segmentation methods of MaskCLIP, SAN, and ZegCLIP for further comparison. We find that these CLIP-based methods have a large performance gap with ours, and are also inferior to the state-of-the-art WSAG method LOCATE.

The comprehensive comparison on UMD dataset is displayed in Tab. 4.2, where

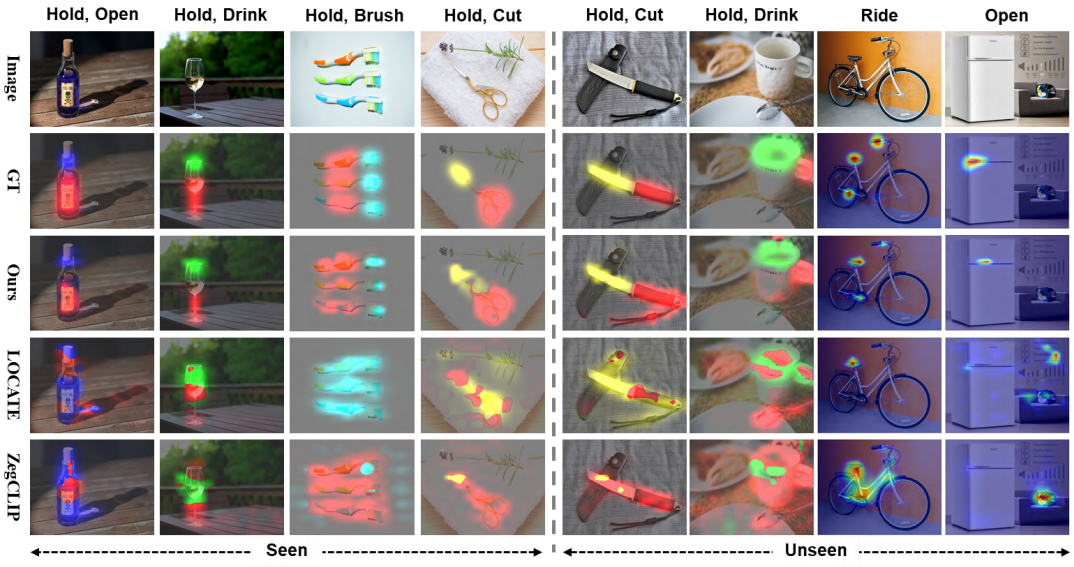


Figure 4.5: Qualitative comparison with LOCATE and ZegCLIP on AGD20K dataset. When multiple affordance predictions overlap, the one with higher value is displayed. Our predictions distinguish different object parts, while other methods often make overlapping predictions.

we benchmark the results with several representative semantic segmentation methods (PSPNet, DeepLabV3+, SegFormer) and open-vocabulary semantic segmentation methods. For fair comparison, the classical segmentation methods are trained with the full training set, while foundation-model-based methods like ZegCLIP and SAN are evaluated in the OOAL setting. It is clear that our proposed model is quite effective, which can be comparable to fully-supervised methods with only 0.36% of their training data. To explore how fully-supervised methods are affected by the limited data, we further train these models in the OOAL setting. Results in Tab. 4.2 show that the performance of these models degrades by around 10% in both seen and unseen settings when given only one-shot example. Additionally, under the same OOAL setting, we observe a more apparent gain over other CLIP-based open-vocabulary segmentation methods, showing that CLIP is not suitable for data-limited affordance learning. The poor performance of MaskCLIP from both tables also verifies that vanilla CLIP has limited understanding on affordances.

4.3.4 Qualitative Results

Qualitative comparisons on AGD20K dataset are shown in Fig. 4.5. We note that WSAG methods like LOCATE often make overlapping predictions for examples with

Figure 4.6: Qualitative comparison with SegFormer and ZegCLIP on UMD affordance dataset in OOAL setting. Images have been enlarged and cropped for better visualization.

multiple affordances, while our results show a clear separation between different affordance regions. ZegCLIP can make reasonable predictions to some extent, but it mostly focuses on the whole object and the accuracy is far from satisfactory, whereas our results are more part-focused, especially for the unseen objects. For example, the prediction for the unseen object of bicycle show that our model can handle the complex affordance (ride) with multiple separated affordance areas (saddle, handlebar, and pedal). In Fig. 4.6, we display the results for UMD dataset. We observe that SegFormer and ZegCLIP often fail to recognize affordances of objects whose parts are similar in appearance. Also, they tend to misclassify metallic object parts as cuttable affordance,

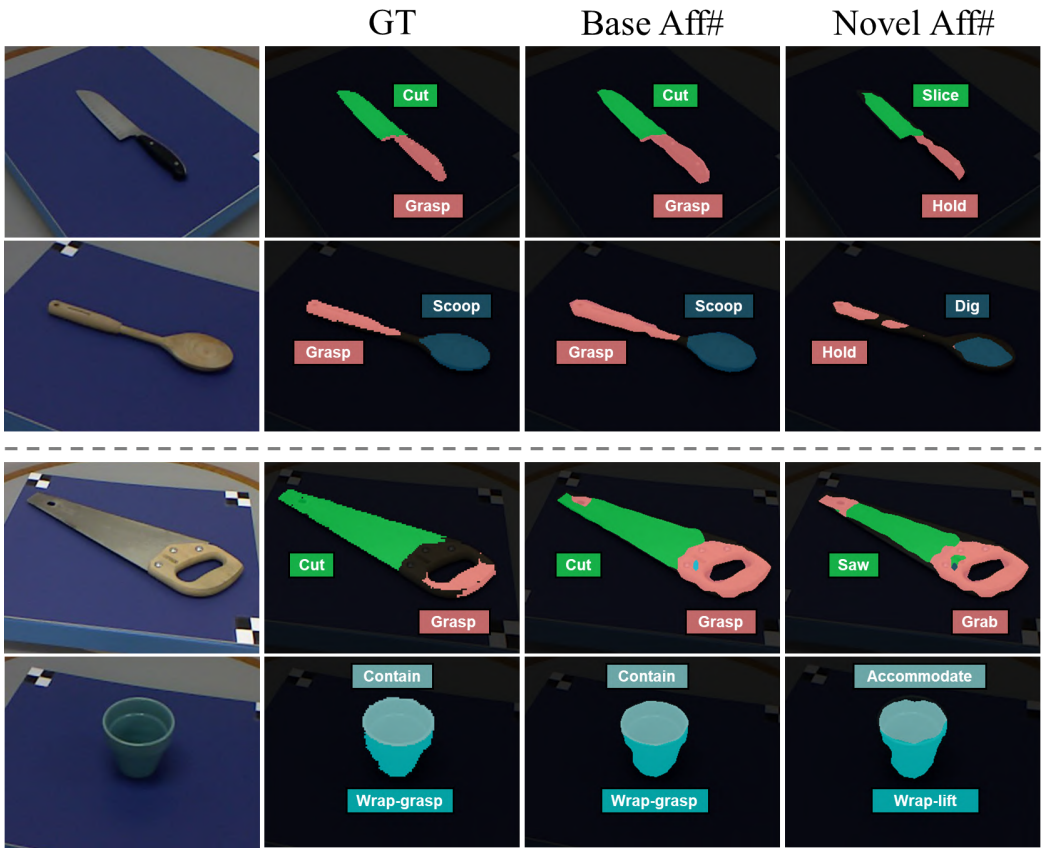


Figure 4.7: Qualitative examples of novel affordance prediction on UMD dataset. The 1st and 2nd rows display results on base objects, and the 3rd and 4th rows show results for novel objects.

suggesting that inferring affordances with only appearance features can be misleading. In comparison, our predictions are more accurate due to the utilization of DINOv2's part-level semantic correspondences.

One particular feature of our model is that it can recognize novel affordances not shown during training. To demonstrate this, we replace the original affordance labels with semantically similar words and check if the model can still reason about corresponding affordance areas. As shown in Fig. 4.7, the model manages to make correct predictions for novel affordances, such as "hold" and "grab" for base affordance "grasp", "saw" for "cut", and "accommodate" for "contain".

4.3.5 Ablation Study

The ablation study is performed on the more challenging AGD20K dataset due to its natural images with diverse backgrounds. Ablations on hyperparameters are left in the

Model	Seen			Unseen		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
CLIP	1.294	0.384	1.107	1.556	0.327	0.966
DeiT III	1.301	0.378	1.140	1.535	0.321	1.049
DINOv2	1.156	0.425	1.297	1.462	0.360	1.105

Table 4.3: Ablation results of different visual foundation models.

Method	Seen			Unseen		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
Baseline	1.156	0.425	1.297	1.462	0.360	1.105
+ TPL	1.060	0.455	1.422	1.338	0.390	1.302
+ MLFF	0.846	0.537	1.622	1.115	0.447	1.440
+ TD	0.749	0.578	1.738	1.131	0.443	1.408
+ CTM	0.740	0.577	1.745	1.070	0.461	1.503

Table 4.4: Ablation results of proposed modules. TPL: text prompt learning. MLFF: multi-layer feature fusion. TD: transformer decoder. CTM: CLS-guided mask.

appendix.

Different Vision Encoders. To complement the qualitative analysis in Sec. 4.2.2, we conduct quantitative experiments on CLIP, DeiT III, and DINOv2. Specifically, we simply process the visual features with the embedder, and perform matrix multiplication with pre-computed affordance text embeddings to output segmentation maps. As shown in Tab. 4.3, CLIP and DeiT III exhibit comparable performance, whereas DINOv2 achieves much better results in both seen and unseen settings, which are consistent with the analysis that DINOv2 is more suitable for affordance learning.

Proposed Methods. We use the DINOv2 with a simple embedder as baseline, and gradually integrate our methods to analyze the effect of each proposed design. The results in Tab. 4.4 reveal that each module can consistently deliver notable improvements. In particular, we notice that the inclusion of a transformer decoder can enhance the performance in the seen setting, but yield inferior results for the unseen setting. With the integration of the CLS-guided mask, results of both settings can be improved, suggesting that restricting the cross-attention space is an effective strategy for unseen object affordance recognition.

4.4 Conclusion and Limitations

In this chapter, we propose the problem of one-shot open affordance learning that uses one example per base object category as training data, and has the ability to recognize novel objects and affordances. We first present a detailed analysis into different foundation models for the purpose of data-limited affordance learning. Motivated by the analysis, we build a vision-language learning framework with several proposed designs that better utilize the visual features and promote the alignment with text embeddings. Experiment results on two affordance segmentation datasets demonstrate that we achieve comparable performance with less than 1% of the full training data.

Our framework exhibits several limitations: First, while text prompt learning enhances the performance within unseen objects, it diminishes the framework's generalization capacity to unseen affordances. This issue arises because an excess of learnable tokens may weaken the intrinsic word similarities within the CLIP text encoder. A viable solution to this limitation involves combining the learnable prompts with manually designed ones. Second, the performance is notably influenced by the selection of the one-shot example. Instances with heavy occlusion or poor lighting conditions can adversely affect learning performance. Given the inherent challenges in learning from merely one-shot example, this limitation appears reasonable and logical. Lastly, we did not employ large language models, which, as demonstrated in recent work [15, 22, 65, 161], can enhance semantic understanding of natural language instructions and facilitate affordance reasoning.

Chapter 5

An Affordance Learning System for Robotic Manipulation

Affordance, defined as the potential actions that an object offers, is crucial for embodied AI agents. For example, such knowledge directs an agent to grasp a knife by the handle for cutting or by the blade for safe handover. While existing approaches have made notable progress, affordance research still faces three key challenges: data scarcity, poor generalization, and real-world deployment. Specifically, there is a lack of large-scale affordance datasets with precise segmentation maps, existing models struggle to generalize across different domains or novel object and affordance classes, and little work demonstrates deployability in real-world scenarios. In this chapter, we address these issues by proposing a complete affordance learning system that (1) takes in egocentric videos and outputs precise affordance annotations without human labeling, (2) leverages geometric information and vision foundation models to improve generalization, and (3) introduces a framework that facilitates affordance-oriented robotic manipulation such as tool grasping and robot-to-human tool handover. Experimental results show that our model surpasses state-of-the-art methods by 13.8% in mIoU, and the framework achieves 77.1% successful grasping among 179 trials, including evaluations on seen, unseen classes, and cluttered scenes.

This chapter begins by identifying three key challenges in current affordance learning research in Sec. 5.1, followed by a detailed explanation of the proposed affordance learning system in Sec. 5.2. Extensive vision and robotic experiments are presented in Sec. 5.3 to evaluate the effectiveness of the system. Finally, Sec. 5.4 summarizes the findings and discusses the limitations.

5.1 Introduction

Understanding affordances of an object means knowing the possible actions it enables as well as locating specific object parts for those actions. This knowledge serves as a key link between perception and action, enabling an intelligent system to transition from observing how an object is used to actually performing the action itself. Despite its importance and recent progress, current affordance-related research often suffers from three major challenges: data scarcity, poor generalization, and real-world deployment. First, affordance datasets are scarce and expensive to create, as they require detailed annotations of object parts, which are small, low-resolution, and often occluded. For example, accurately segmenting the handle of a spoon for holding is challenging due to its thin structure and potential occlusions in cluttered scenes. Second, current models often struggle to generalize to diverse new objects, affordances, or environments, as they are typically trained from scratch on a limited set of data. Finally, few affordance studies deploy methods in real robots, where the models need to be robust to noise, novel scenes, and other real-world factors. These three challenges are in fact deeply interconnected: a lack of large-scale and diverse data limits model generalization, which is essential for reliable real-world deployment.

To tackle each of these challenges, we propose a comprehensive affordance learning system that encompasses data collection, model learning, and robot deployment. The system starts with an automatic pipeline for collecting training data from egocentric videos, which are rich in human-object interactions. While a number of studies [4, 57, 78, 97] have explored extracting affordances from egocentric videos, two limitations persist in their pipeline as illustrated in Fig. 5.1: (1) The focus is primarily on how humans grasp objects (graspable affordance), rather than on which part of the tool is being used (functional affordance). (2) Affordances are learned and represented as probabilistic distributions, which are coarse and noisy, making them difficult to apply in real-life situations and susceptible to distractions. To resolve these limitations, we aim to jointly annotate both graspable (*e.g.*, object handles) and functional affordances (*e.g.*, knife blades, hammerheads), focusing on generating precise segmentation maps rather than coarse heatmaps. Concretely, given an egocentric video, our pipeline first extracts graspable points on objects from hand-object interactions and functional points from tool-object interactions. To cope with occlusions, we identify the pre-contact frame, where contact is about to occur, and project the extracted points to this frame through homography or point correspondence. Finally, these points act

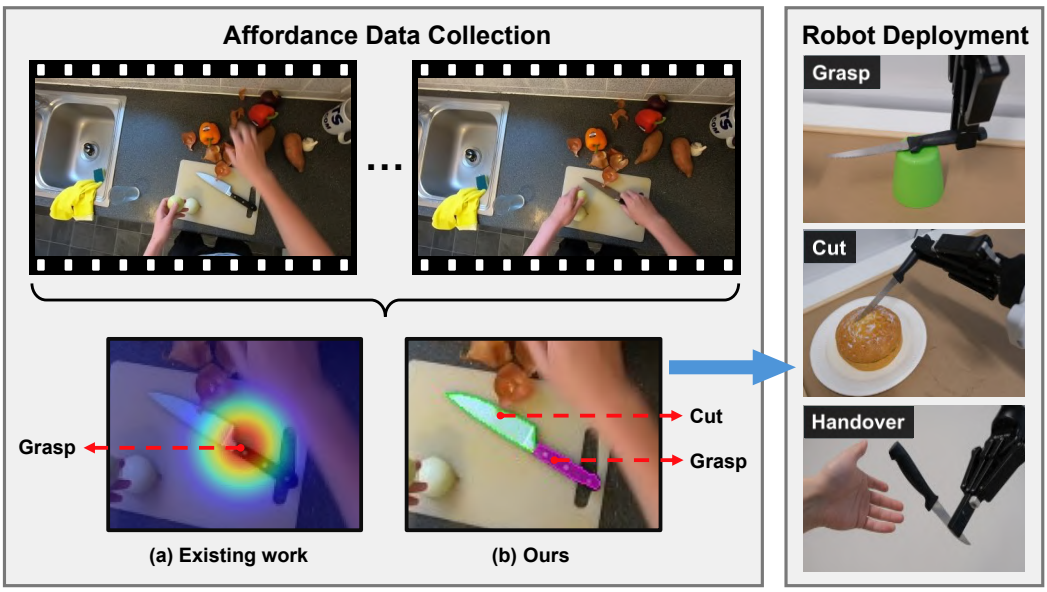


Figure 5.1: Illustration of affordance data collection and robot deployment. Existing work [4, 57, 77, 78] collects the graspable affordance as Gaussian heatmaps, whereas we extract both graspable and functional affordances with precise segmentation masks, enabling tool grasping, tool-object interaction, and robot-to-human tool handover.

as prompts for the segment anything model (SAM) [62] to obtain precise part segmentation.

Although the previous step enables collection of images and annotations at a low cost, the resulting data pose challenges for training. Most samples have quite low resolution, with cropped areas often comprising only 5% of the original frame. To address this issue, we propose Geometry-guided Affordance Transformer (GAT), which enriches and constrains the prediction process by leveraging the geometric features, rather than relying solely on blurry or low-resolution appearance. Moreover, we observe that the model yields inferior performance when evaluated on data that largely differs from the training source domain. To tackle this domain gap, we use the visual foundation model DINOv2 [106] as the image encoder, which has been trained on data from various domains. Lastly, to enable affordance-oriented manipulation, we introduce Aff-Grasp, a framework that combines the GAT with a grasp generation model for robotic manipulation. Building on the accurate predictions of graspable and functional affordances by GAT, Aff-Grasp can handle tasks beyond simple grasping, including tool-object interaction and robot-to-human tool handover (see examples in Fig. 5.1).

To comprehensively demonstrate the effectiveness of our methods, we perform evaluations from two perspectives. First, we collect and annotate images from several

existing affordance datasets and internet sources, creating a challenging evaluation dataset of great diversity to assess the model's performance. Second, we design a real-world robotic manipulation evaluation with 7 tasks and 34 diverse objects. It is worth nothing that both evaluations include out-of-domain data and novel objects, making this a challenging cross-domain and zero-shot setup.

Overall, the contributions of this work can be summarized as follows: (1) *Automated Affordance Data Collection:* We propose an automated pipeline for collecting and annotating affordance data from egocentric human-object interaction videos. Different from previous work, the data are collected with precise segmentation maps for both graspable and functional affordances. (2) *Advanced Affordance Learning Model:* We introduce Geometry-guided Affordance Transformer (GAT) that incorporates shape and geometric priors in a flexible and innovative way to tackle the challenging affordance segmentation task. (3) *Affordance-Oriented Manipulation:* We present Aff-Grasp, a framework designed for affordance-oriented grasping. Aff-Grasp can identify the most suitable object based on task instructions, grasp the appropriate part, and utilize its functional component to complete the task (without specifying explicit object or part names). (4) *Extensive Vision and Robot Evaluations:* We conduct experiments on both visual data and a real robot. A challenging affordance evaluation dataset is created for vision evaluation and diverse tasks are designed for robot experiments across a wide range of objects.

5.2 Method

In this chapter, our goal is to develop a holistic system that covers data collection, model learning, and robot deployment. To this end, we first develop an automated pipeline to collect images and related affordance annotations from human videos. Next, we propose an effective affordance learning model termed Geometry-guided Affordance Transformer (GAT). Finally, we introduce Aff-Grasp that couples the trained model with an off-the-shelf grasp estimation model to achieve affordance-oriented manipulation. In Sec. 5.2.1, we describe how affordance data are collected from large-scale egocentric videos of human interactions. We then elaborate on the design of GAT that enables effective affordance learning from collected data in Sec. 5.2.2. Lastly, in Sec. 5.2.3, we explain Aff-Grasp, detailing how it yields affordance-oriented grasp poses.

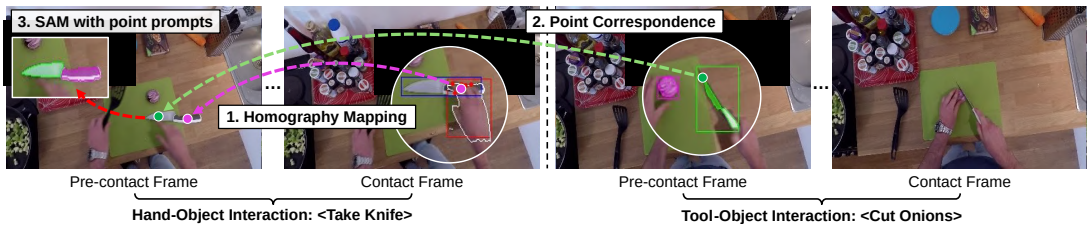


Figure 5.2: Illustration of the data collection process from egocentric videos. First, graspable points (depicted in purple) are localized from clips of hand-object interaction and then projected to pre-contact frame by homography. Next, functional points (depicted in green) are identified from tool-object interactions and mapped to the pre-contact frame of hand-object interaction through point correspondence. Lastly, these points are used as prompts for the SAM to obtain affordance masks.

5.2.1 Data Collection from Egocentric Videos

Given an egocentric video of a human interacting with an object, our aim is to first locate contact points. Human-object interaction videos can generally be categorized into two types: hand-object interaction and tool-object interaction. In hand-object interaction, contact points indicate where the human grasps the object. In tool-object interaction, contact points reveal which part of the tool is used to interact with the target object. These points represent sparse *graspable* and *functional* areas of an object, carrying rich affordance information. As shown in Fig. 5.2, we propose a pipeline to automatically collect these points without manual annotation. The collected points then serve as prompts to produce precise segmentation masks using SAM [62].

Graspable Point Localization. Egocentric videos, such as Epic-Kitchens [29] and Ego4D [44], include timestamped narrations that describe actions and their respective start and end times. Based on these narrations, we first retrieve hand-object interaction clips (associated with actions such as "take" or "hold") and employ a hand-object detector [118] to generate contact states and hand-object bounding boxes for all frames. Next, we use labeled timestamps to extract the contact frame, which is typically annotated as the start of an action. We conduct an additional hand segmentation in this frame using the hand box as a prompt via EfficientSAM [138]. We then locate the intersection region of the hand mask and object bounding box to sample n contact points $P = \{p_1, p_2, ..., p_n\}$. However, the sampled points are often occluded by hands, and therefore do not accurately represent the graspable affordance area of the object. To collect clean object images free of occlusion, it is necessary to identify the pre-contact frame, *i.e.*, the last frame where the object is fully visible before contact occurs. We

utilize the contact states to detect the frame that is closest to the contact frame but without hand-object contact, designating this as the pre-contact frame. Since human motion between adjacent frames is minimal, we follow a similar pipeline to previous studies [57, 78] to project the average position of sampled graspable points to the pre-contact frame by computing a homography transformation [124], as illustrated by the purple dashed line in Fig. 5.2.

Functional Point Localization. To localize functional points, we first retrieve the relationship between objects and affordances from existing affordance datasets. We then extract related tool-object interaction clips from video narrations based on the object-affordance relationship. For example, most datasets associate affordances "cut" and "grasp" with a knife. After localizing the graspable points from a clip showing a person grasping a knife, we then retrieve the following nearest clip depicting a cutting action with the knife.

However, the tool is often heavily occluded or invisible in the contact frame. Similar to graspable point localization, we need to find the pre-contact frame that shows minimal or no intersection between the tool and the target object. To achieve this, we first employ the same combination of the hand-object detector and EfficientSAM to obtain the bounding box and mask of the hand-held tool. Next, we use an open-vocabulary object segmentation model, GroundedSAM [111], to segment the target object. We then measure the Intersection over Union (IoU) between tool and object bounding boxes in frames prior to the contact frame until the IoU is below a preset threshold. Lastly, we calculate the point distances between masks in this pre-contact frame and extract the point within the tool mask that has the shortest distance to all points in the object mask. An erosion operation is applied to the tool mask to ensure that the functional point is inside the tool.

Nonetheless, not all object categories have related action clips in the narrations. In such cases, we use the farthest sampling to determine the functional points based on the distance to the grasp points. This simple method produces accurate functional points, as most tools are designed with graspable and functional parts distributed at opposite ends.

Data Generation. After extracting functional points, we first project these points to the pre-contact frame of the hand-object interaction clip where we infer the graspable points. Since the object category remains the same, we compute the point correspondence within object bounding boxes using foundation model features [1], which map the functional point from the tool-object pre-contact frame to the hand-object pre-

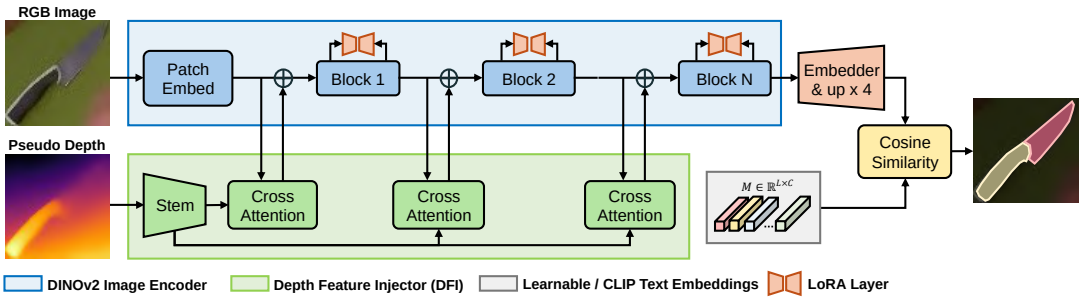


Figure 5.3: The architecture of GAT. It consists of a DINOv2 image encoder, a depth feature injector, an embedder, and LoRA layers. The model performs segmentation by computing cosine similarity between upsampled features and learnable / CLIP text embeddings.

contact frame (illustrated by the green dashed line in Fig. 5.2). We then label the graspable points as positive and the functional points as negative to obtain the graspable affordance mask. Conversely, the functional affordance mask is generated by reversing the positive and negative labels. Finally, we crop the object images and store them along with the generated segmentation masks as annotations.

5.2.2 Geometry-guided Affordance Transformer

While the affordance data are collected from egocentric videos without manual labor, there are two issues that hinder effective model training. The first issue is the low resolution of the collected images. The object of interest occupies very small areas of the video frames, often resulting in cropped images of less than 100 pixels in either height or width. The second is the limited diversity of the training data, characterized by monotonous backgrounds and mostly restricted to indoor scenes. To cope with these issues, we propose an affordance learning architecture called GAT, as illustrated in Fig. 5.3. It includes a novel Depth Feature Injector (DFI) that integrates geometric priors into image features using pseudo depth maps, a DINOv2 image encoder as a feature extractor, and additional LoRA layers for effective fine-tuning.

Depth Feature Injector. We argue that depth maps introduce rich geometric information that can help with foreground-background and part separation. Additionally, they exclude color information, allowing the model to fully focus on the shape information, which is highly relevant to affordances. For instance, graspable parts often consist of shapes like cylinders or spheres, while parts designed for cutting typically feature sharp edges and flat surfaces.

Specifically, we first obtain pseudo depth maps for each training image with a state-of-the-art depth estimation model Depth-Anything [144]. During model training, the pseudo depth map is first encoded into feature maps using a stem block, which contains three standard 3×3 convolution layers. These feature maps are then processed by a 1×1 convolution that transforms the channel dimension to match that of the RGB image features.

We divide the whole model into four blocks. At the beginning of each block, the DFI takes the image features $F_i \in \mathbb{R}^{N \times C}$ and depth features $F_d \in \mathbb{R}^{N \times C}$ as input, and outputs updated image features $\hat{F}_i \in \mathbb{R}^{N \times C}$, where N denotes the number of patches. Concretely, DFI contains several cross-attention layers followed by residual connections. In the cross-attention layer, F_i is used as the query, and F_d is adopted as the key and value:

$$Q = \phi_q(F_i),\ K = \phi_k(F_d),\ V = \phi_v(F_d), \tag{5.1}$$

$$\hat{F}_i = \beta \cdot \text{softmax}(QK^T / \sqrt{d_k}) \cdot V + F_i, \tag{5.2}$$

where ϕ is a linear transformation, and d_k is the dimension of the key acting as a scaling factor. Following [19], we set a learnable vector $\beta \in \mathbb{R}^C$, initialized to 0, to balance the output from the cross-attention layer and the image feature. This strategy prevents the image feature from being excessively affected by the depth feature, making the training process more stable. We observe that DFI constantly brings improvement, even when integrated solely during training (see Sec. 5.3.3). This indicates that it can act as a regularization mechanism during training, and can be discarded during inference to speed up the process.

DINOv2 with Low-Rank Adaptation. We notice that directly training from the typical ImageNet pre-trained representation often leads to inferior results. This can be attributed to two primary reasons: First, affordance segmentation focuses on fine-grained object parts, whereas representations trained for image classification emphasize global object features [1]. Second, ImageNet pre-trained models exhibit limited diversity, making it challenging to handle data from diverse domains. To address this issue, we employ the self-supervised visual foundation model DINOv2, which has been demonstrated to be highly effective for data-limited affordance learning [68]. Furthermore, we introduce LoRA [50] to fine-tune the model without modifying the parameters of the original DINOv2. This strategy helps adaptation across different domains and prevents overfitting. LoRA was originally developed to fine-tune large language models for different downstream tasks. Specifically, it injects trainable rank decomposition

matrices to a pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$ by $W_0 + \Delta W = W_0 + BA$, where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, and the rank $r \ll \min(d,k)$. During training, only A and B are trainable, while W_0 remains frozen. This incurs minimal computational cost and memory usage.

Classifier and Loss Functions. To speed up inference time for real-world applications, we avoid adding any complex decoder structures. Instead, we process the output feature with an embedder (an MLP), reshape it, and upsample it by a factor of four to increase the resolution $F_{out} \in \mathbb{R}^{C\times\frac{4H}{p}\times\frac{4W}{p}}$, where p is the patch size. Next, we intialize $M \in \mathbb{R}^{L\times C}$ learnable embeddings, where L is the number of affordance categories. We compute the cosine similarity between M and F_{out} to yield the segmentation output, which is then restored to the same size as the input image via bilinear interpolation. Due to the domain gap, we do not add a learnable embedding for the background classification to prevent overfitting. Alternatively, we determine a pixel as background if all its affordance predictions are below a preset threshold τ. Compared to a linear layer and explicit background classifier, the cosine similarity-based segmentation and implicit background prediction are more robust and can effectively improve the performance, as detailed in Sec. 5.3.3. In addition, to achieve open vocabulary affordance segmentation, M can also be replaced with corresponding CLIP text embeddings, as verified in [68].

Since the collected data are highly unbalanced, we utilize a combination of focal loss [75] and dice loss [91] as training objectives:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{focal} + \mathcal{L}_{dice}, \tag{5.3}$$

where α is a weighting factor to balance loss values.

5.2.3 Affordance-Oriented Robotic Manipulation

Our ultimate goal is affordance-oriented robotic manipulation, where given a task and a cluttered scene, the robot can select the object that possesses the related affordance, grasp the correct part, and apply the functional part to the target object to perform desired actions. To achieve this, we propose Aff-Grasp, which integrates GAT to achieve affordance segmentation and transforms the visual affordance to available grasp poses. The framework of Aff-Grasp is shown in Fig. 5.4. Given a task consisting of a verb and a target, such as “cut cake”, it first uses an open-vocabulary object detection model [79] to detect the target (cake) and other visible objects. For objects other than the target,

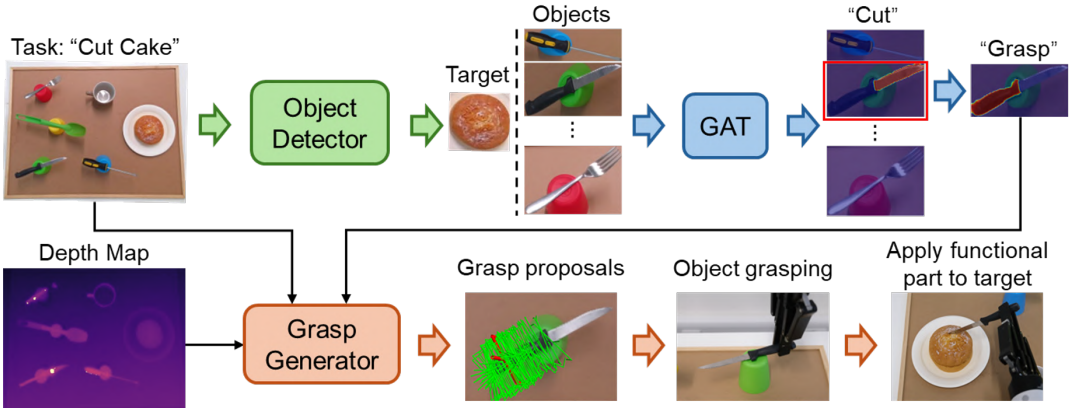


Figure 5.4: The framework of Aff-Grasp. It first employs an open-vocabulary detector [79] to locate all objects within the scene, which are then sent to GAT to determine if they possess the corresponding affordance required for the task. Afterwards, a 6 DoF grasp generation model, Contact-GraspNet, leverages the object's graspable affordance and the depth map to generate dense grasp proposals. Finally, the robot executes affordance-specific sequential motion primitives to apply the functional part to the target.

the input vocabulary is simply set to "objects" for class-agnostic detection. These detected objects are then cropped and sent to GAT to predict affordance areas. The object with the most certain affordance area for the required action (cut) is identified. After that, we extract the graspable affordance area of this object, *i.e.*, the knife handle, and generate potential grasp poses within it. For grasp pose estimation, we select Contact-GraspNet [123] that can produce dense grasp proposals within the specified mask area. The one with the highest score is then chosen as the execution pose. Once the object is grasped and lifted, we execute affordance-specific sequential motion primitives to apply the functional affordance area to the target object to complete the task. For the handover task, Aff-Grasp is instructed to find available grasp proposals within the functional affordance area and then pass the graspable part to the human hand. When CLIP text embeddings are used as classifiers, the action required for a task can be transformed into text embeddings for open-vocabulary affordance segmentation [68, 102]. Therefore, unseen affordance vocabularies can also be used at inference, enhancing the model's adaptability and versatility.

5.3 Experiments

In this section, we present experiments from both vision and robot perspectives, along with ablation studies that examine the design choices of GAT. Note that all evaluations are conducted in a zero-shot or cross-domain setting, since the test data and robot setup contain novel objects or affordances, and differ greatly from the training data collected from egocentric videos. As a result, only models with strong generalization ability can achieve high performance. Implementation details are provided in the appendix.

5.3.1 Vision Experiments

Evaluation Dataset. To evaluate the effectiveness of GAT, a diverse and challenging affordance dataset that has consistent object and affordance categories with the collected training data is needed. After carefully inspecting existing datasets, we found that most of them are not compatible with our evaluation requirements. Many datasets either have a small number of categories [23, 101] or are collected in the lab environment with limited diversity [64, 96]. Some datasets contain a large number of images but have coarse keypoint-based annotations [39, 85] or small resolutions [58]. Therefore, we create an Affordance Evaluation Dataset (AED) by manually annotating 721 images collected from several existing affordance datasets and internet resources. It contains 13 object categories and 8 affordance classes. See appendix for more details on AED.

Pre-train	Method	mIoU	F1	Acc
ImageNet	DeepLabV3+ [17]	13.46	22.27	23.05
	PSPNet [155]	16.90	27.32	26.46
	SegFormer [136]	23.72	36.86	37.19
Foundation Models	ZegCLIP [160]	18.33	26.41	25.55
	DINOv2 [106]	46.16	62.49	63.61
	ViT-Adapter [19]	50.86	66.88	65.21
	OOAL [68]	54.82	70.58	68.00
	GAT (Ours)	**68.62**	**81.09**	**83.51**

Table 5.1: Quantitative comparison on the AED.

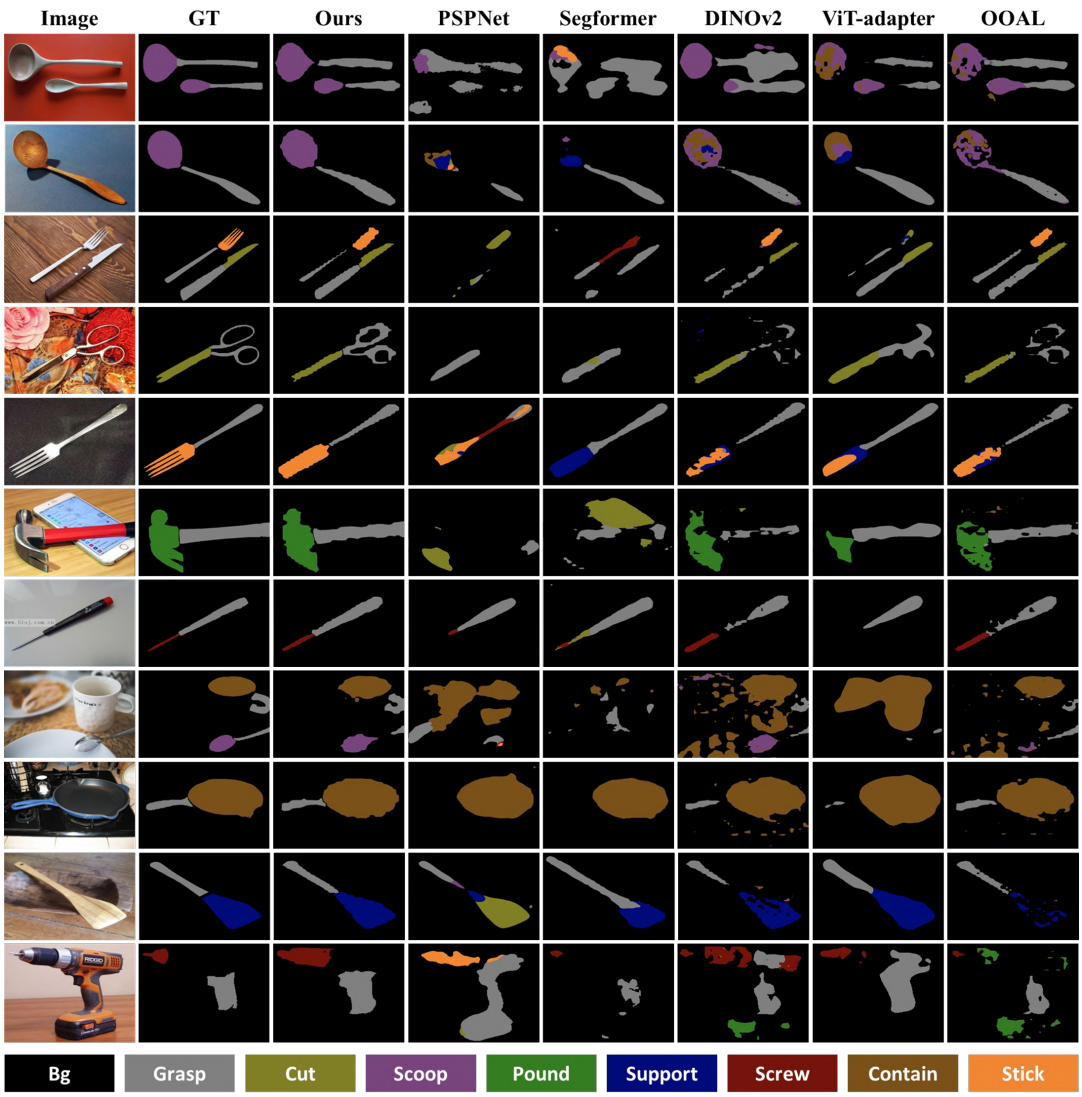


Figure 5.5: Qualitative comparison between our approach and other segmentation models on the AED.

Quantitative and Qualitative Comparisons. Table 5.1 shows the results of different state-of-the-art segmentation approaches on the proposed AED. They use either pre-trained ImageNet backbones to extract feature maps, or obtain representations from visual foundation models like CLIP [109] and DINOv2 [106]. Thus, we divide these models into two sections based on the pre-training strategies. For ImageNet pre-trained models, we employ classical CNN segmentation models such as DeepLabV3+ [17] and PSPNet [155], as well as a transformer-based segmentation model SegFormer [136]. For visual foundation-based models, we choose ZegCLIP [160], DINOv2 [106], ViT-Adapter [19], and OOAL [68] to compare with GAT, as they represent the state-of-the-art in leveraging visual foundation models for semantic or affordance segmen-

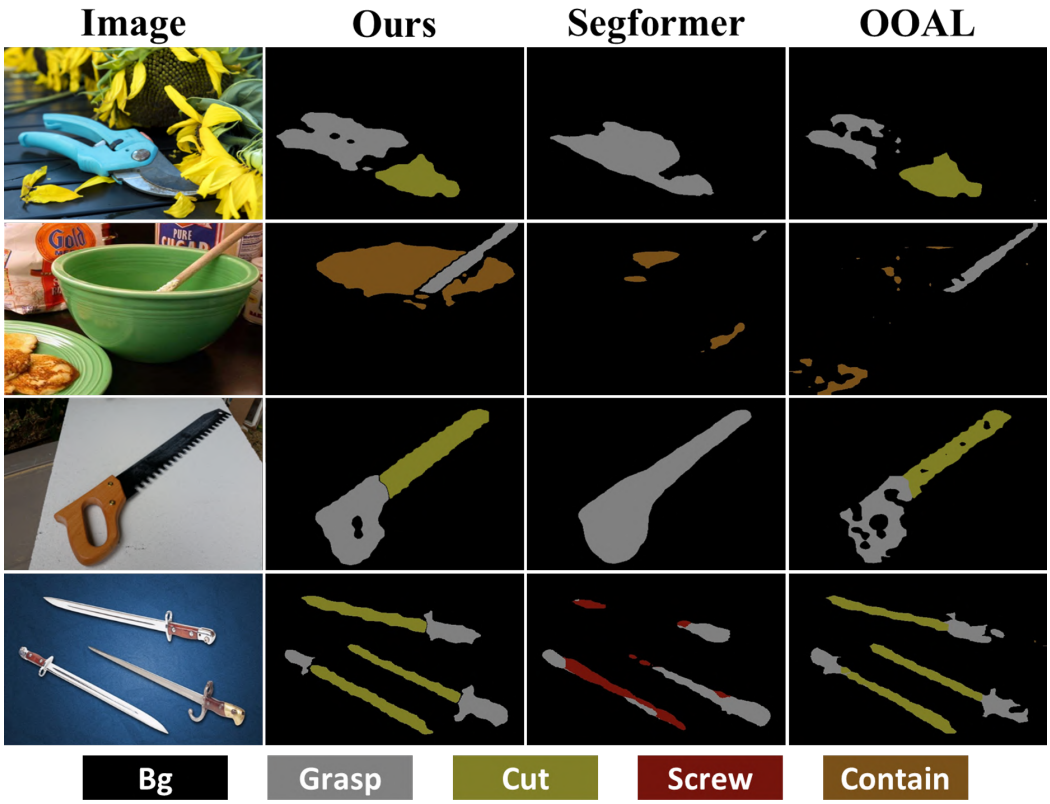


Figure 5.6: Qualitative comparison on unseen objects.

tation tasks. We notice that methods using pre-trained ImageNet backbones generally produce much inferior results compared to those based on foundation models. This confirms the substantial domain gap between training and evaluation sources, and demonstrates that foundation models are more resistant to noisy training samples and have better cross-domain capabilities. Among the foundation model based approaches, our model GAT outperforms other methods by a large margin across all metrics. While GAT leverages additional depth information during training, ablation studies in Sec. 5.3.3 show that depth can be disabled during inference without substantial performance degradation.

Figure 5.5 depicts the qualitative comparison between our methods and other models. We find that models like PSPNet and Segformer often yield incomplete or incorrect affordance predictions, which may result from the low diversity in the pre-trained ImageNet representation. On the other hand, most models based on DINOv2 can coarsely generate correct affordance prediction map, but often suffer from incomplete part activation and noisy segmentation around object boundaries. In contrast, the results from GAT are part-focused, exhibit well-preserved boundary segmentation, and are capa-

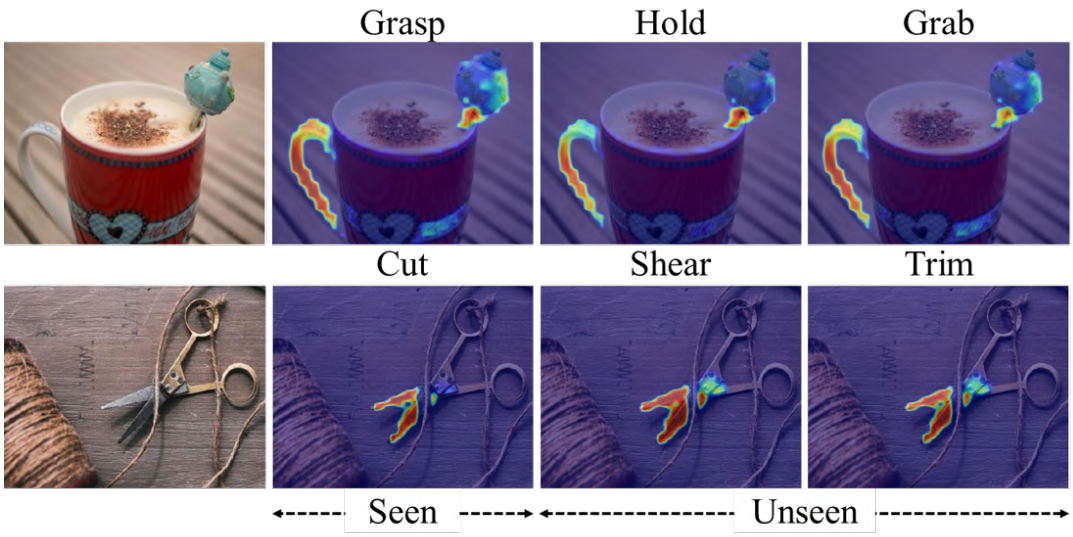


Figure 5.7: Qualitative examples on novel affordances.

ble of handling complex objects like drills. Notably, our results stand out from other counterparts when dealing with images containing multiple objects, as shown in the 1st, 3rd, 6th, and 8th examples in Fig. 5.5. Generalization tests on unseen objects and affordances are provided in the appendix.

In addition, we perform a further qualitative comparison on unseen objects to explore the models' generalization ability. As displayed in Fig. 5.6, novel objects such as shears, saw, bowl, and sword are used. It is apparent that Segformer cannot make accurate affordance predictions for these objects. OOAL demonstrates acceptable potential on unseen objects but often produce less confident and inconsistent results. In comparison, GAT shows excellent performance on these out-of-distribution objects with much more complete segmentation maps. Furthermore, in Fig. 5.7, we present examples to showcase that our model can generalize to novel affordances that are synonymous with the trained actions when using CLIP text embeddings as the classifier.

5.3.2 Robot Experiments

Experiment Setup and Comparison Methods. Real-world experiments are conducted to evaluate three essential properties: accuracy, robustness, and generalization. Illustrations for these experiments are shown in Fig. 5.8. For the comparison methods, since this work focuses on visual affordance learning (i.e., identifying appropriate graspable and functional regions) rather than grasp success, manipulation models are not directly comparable. Instead, we select two relevant methods, LOCATE [67] and Robo-

Figure 5.8: Illustration of accuracy, robustness, and generalization evaluations. The accuracy evaluation requires the model to recognize the affordance of a single object and execute related task. The robustness evaluation involves accurately selecting a object in a cluttered scene to perform a specified affordance task. The generalization evaluation accesses if the model can reason about the graspable area of unseen objects.

Models	Correct Affordance	Successful Grasp	Successful Interaction
LOCATE [67]	42/72 (58.3%)	33/72 (45.8%)	n/a
Robo-ABC [57]	62/72 (86.1%)	44/72 (61.1%)	n/a
OOAL (Ours)	70/72 (97.2%)	57/72 (80.6%)	47/72 (65.3%)

Table 5.2: Success rates for accuracy evaluation.

ABC [57], that learn affordances in a similar manner—from egocentric images and videos, respectively.

Quantitative and Qualitative Comparisons. The results for the accuracy evaluation are shown in Tab. 5.2. It is clear that Aff-Grasp outperforms its competitors significantly, achieving an 11.1% higher affordance prediction rate and a 19.5% increase in successful grasping compared to Robo-ABC. Also, it is worth mentioning that our success rate for affordance prediction is measured on both graspable and functional affordances, whereas other two methods are measured solely on the graspable affordance. We observe that LOCATE struggles to make accurate predictions in real-world scenarios. Robo-ABC has relatively accurate affordance predictions, but it generates grasp proposals based on point correspondences, which do not represent the most confident grasp. Consequently, even though Robo-ABC frequently makes correct affordance predictions, 29% of its proposed grasping points do not lead to successful grasps.

In addition, we note that Aff-Grasp is capable of recognizing the correct affordance in cluttered scenes. As presented in Tab. 5.3 for the robustness evaluation, Aff-Grasp achieves a high success rate in affordance prediction, accurately predicting affordances 95% of the time, even in the presence of multiple seen and unseen objects acting as dis-

	Cut	Stir	Scoop	Screw	Pour	Stick	Handover	Total
Correct Affordance	8/9	9/9	9/9	8/9	9/9	9/9	17/18	69/72 (95.8%)
Successful Grasp	7/9	6/9	7/9	6/9	7/9	6/9	13/18	53/72 (73.6%)
Successful Interaction	7/9	6/9	5/9	5/9	6/9	5/9	11/18	46/72 (63.9%)

Table 5.3: Success rates for robustness evaluation.

Models	Correct Affordance	Successful Grasp	Inference Time (s)
LOCATE [67]	20/35 (57.1%)	15/35 (42.9%)	0.0047
Robo-ABC [57]	24/35 (68.6%)	21/35 (60.0%)	12.92
OOAL (Ours)	32/35 (91.4%)	28/35 (80.0%)	0.0063

Table 5.4: Success rates for generalization evaluation and inference time for affordance prediction components.

tractors. Table 5.4 reports results from the generalization evaluation and the inference time for the affordance prediction component of the models. It can be observed that our method is efficient and significantly more accurate in predicting the correct graspable areas for unseen objects, leading to a much higher success rate in grasping. While LOCATE also has a fast inference speed, it fails to accurately infer graspable affordances. In contrast, Robo-ABC's performance on unseen objects is considerably reduced and suffers from a much longer inference time. Although it does not require additional training, its retrieval and correspondence mapping processes are quite time-consuming and computationally expensive, making it less suitable for real-world applications.

In Fig. 5.9, we present raw predictions of graspable affordance from each model for seen and unseen objects. It can be observed that LOCATE often produces incomplete and wrong predictions. Robo-ABC occasionally makes predictions within the right object part area, but also produces high activation for the background or the entire object. In comparison, Aff-Grasp consistently makes precise segmentation predictions for both seen and unseen objects and is not affected by the background.

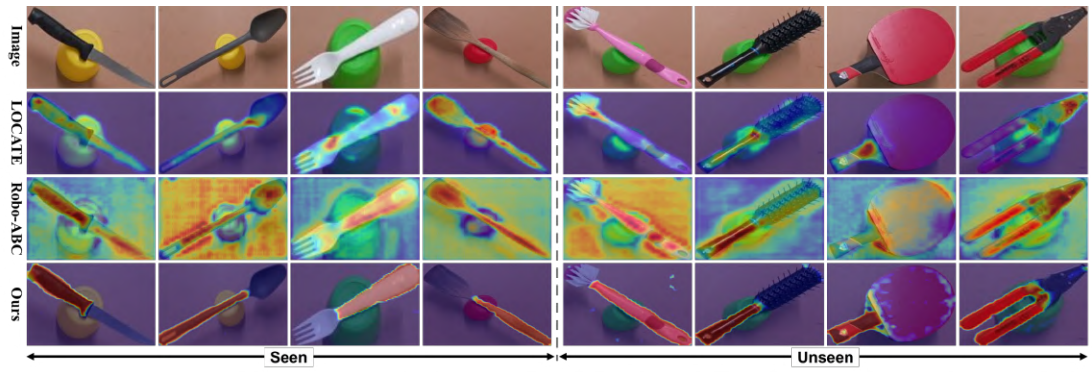


Figure 5.9: Qualitative comparison of graspable affordance predictions on seen and unseen object categories.

Methods	mIoU	F1	Accuracy
Baseline - DeiT III	31.02	44.55	35.85
w/ DINOv2	45.45	61.78	70.86
w/ embedder	48.83	65.10	71.07
w/ embedder & up×4	51.41	64.26	67.27
w/ focal loss	50.70	66.97	70.12
w/ focal & dice loss	53.12	69.13	74.55
linear layer w/o bg	54.96	70.50	71.97
cosine similarity	55.52	71.01	71.54
cosine similarity w/o bg	56.70	72.00	71.22
+ DFI-training only	60.15	74.92	79.87
+ DFI	64.66	78.35	79.74
+ LoRA	68.62	81.09	83.51

Table 5.5: Ablation results of embedder, loss functions, classifiers, and proposed modules. The baseline model is a DeiT III model with a linear layer and binary cross entropy loss. “w/o bg” means that there is no background classifier. “DFI-training only” denotes that the DFI is used during training, and discarded at inference.

5.3.3 Ablation Study

To explore the impact of each component in our model, we perform ablation experiments on the embedder, loss function, classifiers, designed modules, and hyperparameters. The ablation results are summarized in Tab. 5.5. We first set up a baseline

Inference	#Params (M)	GFLOPs	Time (ms)
w/ DFI	10.9	204.9	10.1
w/o DFI	$5.7_{\downarrow 47.7\%}$	$185.5_{\downarrow 9.5\%}$	$6.3_{\downarrow 37.6\%}$

Table 5.6: Ablation study on DFI on inference efficiency.

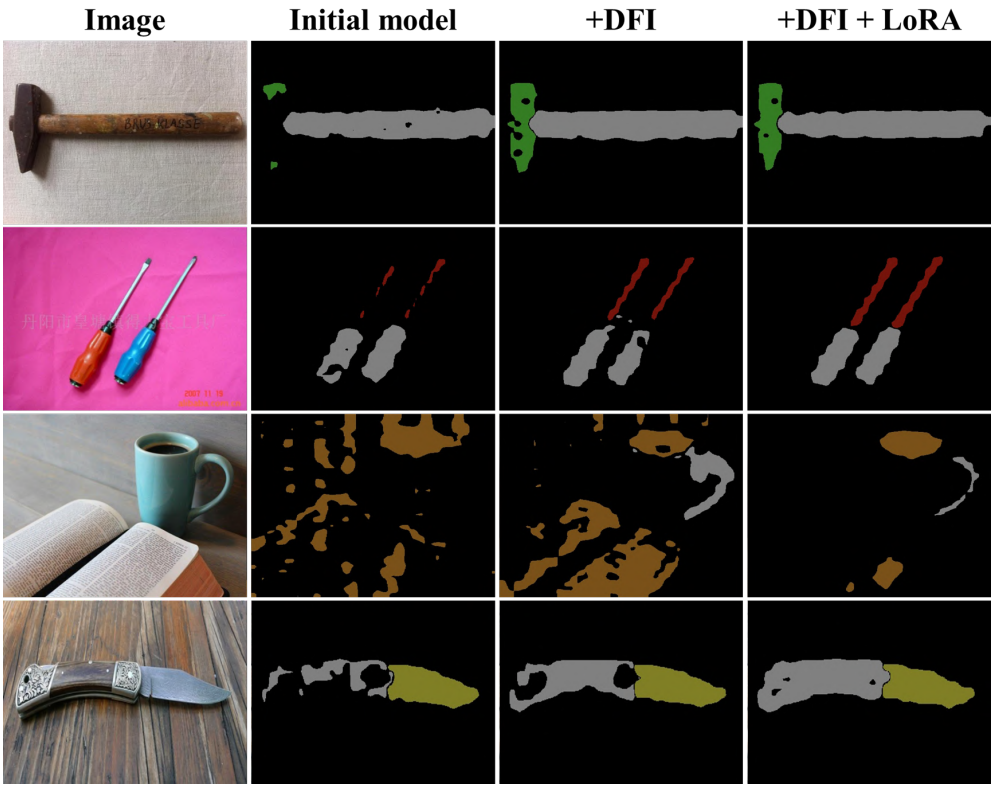


Figure 5.10: Qualitative improvements with DFI and LoRA.

model, which employs a frozen DeiT III [128] backbone that is fully supervised on ImageNet. We then add a simple linear layer for patch-wise classification and utilize binary cross-entropy as the loss function. Based on this baseline, we first explore the impact of the embedder and loss functions. We find that a larger feature map followed by an embedder is beneficial, and the combination of focal loss and dice loss also brings improvements. Then, we analyze the results under different classification schemes, including the linear layer, cosine similarity, and whether to learn a background classifier. It is clear that implicit background prediction leads to better performance. Given the large gap between training and evaluation data, learning a background classifier can easily result in overfitting. Also, employing cosine similarity as the classifier can better utilize the inherent features of DINOv2, producing better results than a linear classi-

fier. Lastly, we investigate the influence of DFI and LoRA. Notably, DFI improves performance significantly by a large margin, with 7.96% and 6.35% increases in mIoU and F1 scores. In particular, DFI can also be used solely in training and discarded at inference, improving results without extra computational cost. The impact of this operation is analyzed in Tab. 5.6, showing that model efficiency greatly improves when deactivating DFI during inference. Additionally, integrating LoRA layers to fine-tune the foundation features is also helpful, leading to a 3.96% improvement in mIoU with marginal additional parameters.

In Fig. 5.10, we show the qualitative ablation results to visually examine the effects of DFI and LoRA. The segmentation results indicate that DFI is particularly effective at locating tiny and slender parts, while LoRA further enhances performance with refined boundaries and more complete segmentation maps. More ablation studies are provided in the appendix.

5.4 Conclusion and Limitations

In this chapter, we present a streamlined affordance learning system that integrates automatic data collection from egocentric videos, effective model learning, and deployment on a real robot. Specifically, we first collect training samples with segmentation masks as annotations from videos of humans interacting with common objects. To effectively train on the collected data, we introduce an affordance learning model named Geometry-guided Affordance Transformer (GAT). GAT features a depth feature injector that incorporates geometric and shape information, which is relevant and beneficial for affordance understanding. Building on GAT, we develop a framework, Aff-Grasp, that facilitates affordance-oriented manipulation. Aff-Grasp enables robots to select the desired object and grasp the correct part without explicitly specifying the object category. To demonstrate the effectiveness of our data collection process and the proposed model, we perform evaluations from both vision and robot perspectives. Extensive experiments show consistent and robust performance, demonstrating the effectiveness of the entire system from data collection to model training and robot deployment.

It is worth noting that there are several limitations in the proposed affordance learning system: (1) *Data Collection:* In this work, we focus primarily on tools with distinct graspable and functional parts, which is a key stepping-stone for more general tools. However, current data collection pipeline exhibits certain limitations in handling thin

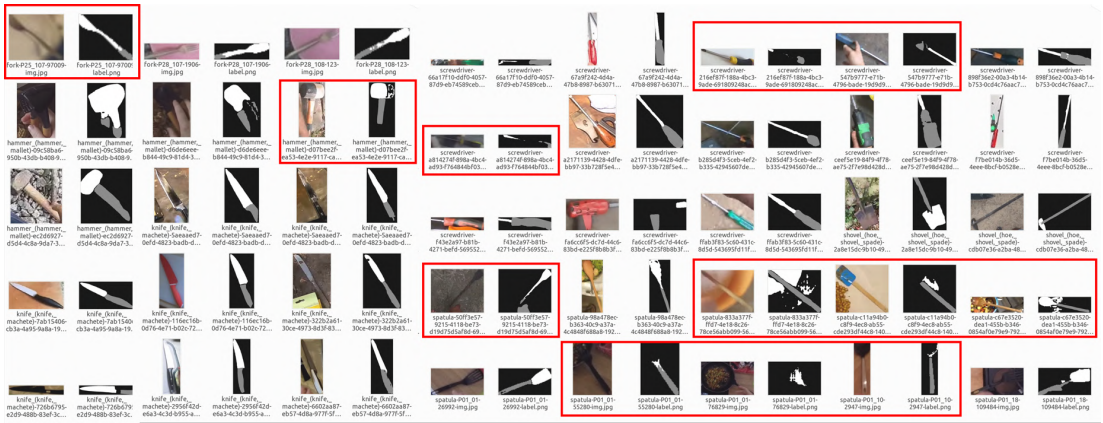


Figure 5.11: Screenshot of the collected data. Noisy annotations are highlighted with red bounding boxes.

and deformable objects that have small or indistinct graspable parts, such as chopsticks or wiping cloths. To address these limitations in future work, we consider integrating LLMs to acquire task priors, which will enable better distinction between graspable and functional parts. Additionally, the quality of the collected data is affected by a variety of factors. On the one hand, occlusion, motion blur, poor lighting conditions, inaccurate narrations, and unpredictable subject behavior from the video data can lead to noisy results. On the other hand, the hand-object detector and the open-vocabulary object detection model can produce incorrect predictions, further affecting the usability of the data. To mitigate these issues, we first add some constraints to reduce the error rate, such as setting high thresholds to filter out uncertain predictions. We then visualize all data samples and manually remove those with completely wrong annotations. Figure 5.11 displays a screenshot of the collected data. Notably, the proposed data collection pipeline is not perfect and the annotations of many samples are noisy and incomplete. Nevertheless, we retain these noisy data to assess the model's performance in this challenging situation. (2) *Model Weakness:* The model prediction can be susceptible to complex texture. As shown in Fig. 5.12, the model fails to make correct predictions when the target objects have complex textures or packaging. Also, the model sometimes confuses object parts with similar materials and shapes. For example, the head of a trowel is incorrectly recognized as having a "cutting" affordance. (3) *Robot Experiments:* Our work improves robotic grasping and interaction performance in real-world scenarios by advancing affordance prediction, as more complete and accurate object part segmentation allows the grasp estimation model to identify grasp poses with greater confidence. However, due to issues such as depth measurement er-

Figure 5.12: Failure cases. The model fails to recognize objects with complex texture and confuses parts with similar shapes and appearances.

rors, partial point clouds, unreliable grasp poses, and robot self-collision, correct affordance prediction does not guarantee a successful grasp, and a successful grasp does not always result in effective tool-object interaction. Since the focus of this chapter is primarily on visual affordance learning, we did not fine-tune the grasp generation model, nor did we perform policy learning to improve the grasp and interaction success rate. Moreover, in our robot experiments, we assume the operating directions of tools are known to simplify the evaluation and design of motion primitives. To enhance practicality and scalability, a 6D object pose estimation model, such as Omni6DPose [153], can be utilized to infer the operating direction.

Chapter 6

Conclusion

In this thesis, we present methods for efficient affordance learning across vision, language, and robotic manipulation, demonstrating that the proposed approaches advance the state of the art and offer promising insights into how affordances can be learned from human-object interactions, integrated with vision and language foundation models, and leveraged for robotic manipulation. Our methods make accurate affordance predictions for diverse base and novel objects, requiring minimal annotation effort. This stands in contrast to traditional methods that often demand extensive training data or numerous simulated interaction trials to gain affordance knowledge. Despite the encouraging performance of our methods [67–69] in both vision benchmarks and real-world robot experiments, several limitations persist. We highlight these limitations and propose future research directions that can further advance the field of affordance learning.

6.1 Limitations

While affordance learning has received increasing attention and shown remarkable progress, current research in this field, including the methods proposed in this thesis, still faces several key limitations:

Limited Diversity. Compared to other mainstream computer vision tasks, the datasets used and proposed for affordance learning cover a relatively narrow range of object and affordance categories. This lack of diversity restricts the system's ability to capture the breadth of interactions found in real-world environments. As a result, current methods may struggle to recognize subtle affordance cues that fall outside the training distribution.

Limited Generalization Ability. While our proposed approaches demonstrate reasonable generalization ability on unseen objects, they still struggle with significantly different objects or environmental conditions. Notably, current vision-language models have a stronger understanding of nouns than verbs. While they can recognize synonyms or synonymous expressions within seen affordance vocabularies, they often fail to generalize to entirely new affordance types. This gap underscores the need for developing vision-language models that can better capture the underlying invariance and subtle differences in affordance vocabularies across diverse scenarios.

Low Reliability in Unstructured Environments. Although the proposed methods in this thesis perform well under lab conditions and on affordance benchmarks, their reliability declines in real-world scenarios. For instance, experimental setups often feature clean tables with minimal distractors, and objects are often positioned or supported in ways that facilitate easier grasping. In contrast, real-world environments introduce challenges such as occlusions, lighting variations, and complex backgrounds—factors that current perception modules are not yet capable of handling robustly. Therefore, existing methods are not reliable enough for many real-world applications. We hope that future advancements in vision and robotic foundation models will help alleviate these limitations.

6.2 Future Work

The work presented in this thesis lays a foundational step in affordance learning, but much diverse and important topics remain unexplored. To advance the development of more robust, versatile, and scalable affordance learning systems, we outline several promising directions for future research:

Affordances in 3D Physical Space. One particularly compelling direction is the integration of 3D affordance learning with spatial understanding. While substantial research has explored 3D semantic and instance segmentation, work specifically related to 3D affordance learning remains scarce. In real-world settings, agents must understand the function of different interactive components in an environment, such as handles, knobs, and buttons, and determine how to manipulate them to achieve desired goals. For example, a recent work named SceneFun3D [31] introduced a large-scale dataset with precise interaction annotations for real-world 3D indoor scenes, offering a valuable resource for learning 3D affordances. Moreover, spatial understanding is essential for accurate interaction and manipulation, as it directly relates to affor-

dance learning. Recognizing affordances requires an understanding of spatial properties, such as shape, orientation, size, and relative positioning of objects. Future work could explore integrating 3D affordance learning with spatial awareness to enhance the robustness and generalization of affordance prediction in complex 3D environments.

Affordance Learning with Generative AI. Another interesting topic involves leveraging generative AI to produce rich and diverse training data. To tackle the challenge of limited training data diversity, synthetic data generation using advanced generative AI techniques provides a promising solution. By simulating realistic objects, interactions, and environmental variations, synthetic datasets can augment real-world samples, enhancing both diversity and scalability while reducing the reliance on laborious manual annotations. In this thesis, we have demonstrated that the human-object interaction data, either images or videos, can be effectively utilized for affordance learning. However, current generation models often fail to capture the subtle details and dynamics of human-object interactions. Future research could explore the generation of controlled and realistic interactions in images, videos, or 3D spaces. In addition to human data, recent studies in robotics [59, 131] have amassed large-scale real-world robot data, which could be leveraged to generate more realistic and diverse data from a robotic perspective.

Integration with Vision-Language-Action Models. In parallel, the rise of Vision-Language-Action (VLA) models [7, 61, 134, 162] offers a fertile avenue for rethinking how affordances are represented and utilized. Unlike previous models that often rely on unimodal inputs for specific tasks, VLAs integrate visual perception, language understanding, and action generation into a unified framework. The goal is to develop a versatile model capable of handling diverse tasks and environments. However, existing VLAs often acquire affordances implicitly by learning from vast datasets that link visual inputs with task instructions and corresponding actions. For instance, by training on numerous instances of a robot grasping a handle to open doors or drawers, VLAs learn that the handle affords the action of opening. While effective, this learning process is inefficient and requires extensive data collection and computational resources. Future work could explore integrating affordance knowledge into VLAs in an explicit manner, which could reduce the need for extensive retraining and improve adaptability.

Fine-Grained and Holistic Affordance Representation. Finally, a more fine-grained and holistic affordance representation will be crucial as tasks grow in complexity. Current representations like masks or heatmaps, while effective, may fall short when it comes to capturing the intricacies of real-world manipulation. Moreover, affordances

encompass more than just the graspable or functional parts of objects, and can be interpreted in a variety of ways. For example, a recent work [99] defines affordances as the pose of the robot end-effector at key stages of a task, using this information as intermediate representations to aid robot manipulation. Another work [71] proposes a chain of affordances that better reflects the sequential interactions between objects in a scene. Specifically, it prompts the model to consider four types of affordances when taking action: object, grasp, spatial, and movement affordances. Future work could explore more comprehensive affordance representations that encapsulate not only graspable or functional aspects but also the subtle contextual cues necessary for effective manipulation. Expanding the representation to include multi-modal signals, such as tactile and force feedback, and temporal dynamics, could lead to a more holistic model that better mirrors how humans interact with objects.

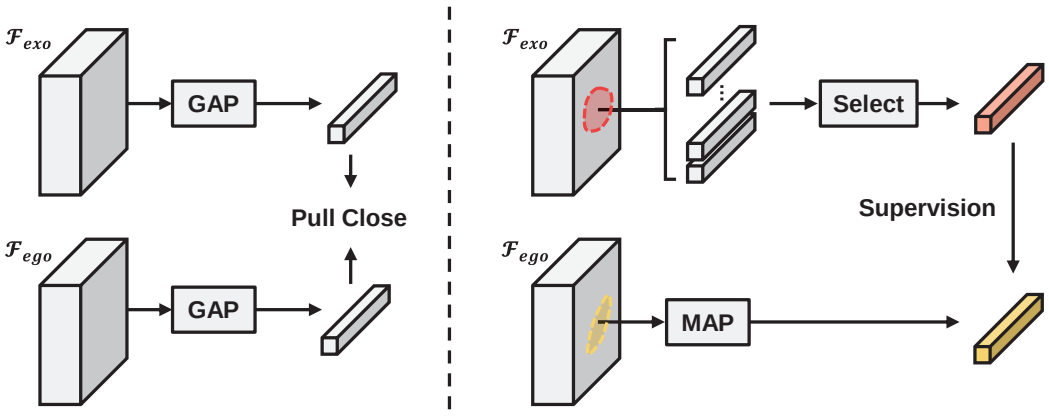
In summary, the path forward is both challenging and exciting. By deepening the affordance understanding in 3D spatial contexts, harnessing generative AI for scalable data synthesis, embedding affordances into VLA models, and enriching the way affordances are represented, we can advance not only the field of affordance learning itself but also broader domains of embodied AI and robotics. We invite the research community to engage with these challenges and contribute to the development of more capable, generalizable, and human-aligned affordance learning paradigms.

Appendix A

Affordance Grounding with Weak Supervision

A.1 Framework Comparison

In Fig. A.1, we show the framework comparison between state-of-the-art affordance grounding work [85, 97] and LOCATE. Previous work performs knowledge transfer by pulling close two global embeddings, and the prediction is only generated at the inference stage. Instead, we conduct part-level knowledge transfer by selecting the object part prototype from exocentric features, and utilizing it as a high-level pseudo-supervision to guide egocentric localization in an explicit manner.



(a) Previous Work (b) LOCATE

Figure A.1: Comparison of LOCATE and previous work. Previous affordance grounding work [85, 97] performs knowledge transfer in a global and implicit manner. In contrast, LOCATE conducts knowledge transfer in a more localized and explicit manner. (GAP denotes global average pooling, and MAP is masked average pooling).

A.2 Additional Experimental Details

Evaluation Metrics. Different from the semantic segmentation task that uses the binary mask as ground truth, the GT for affordance grounding is a probability distribution that indicates the affordance area, i.e., "action possibilities". Following previous work [39, 78, 85, 87, 97], we use KLD, SIM, and NSS as the metrics to evaluate the prediction performance. KLD, SIM, and NSS are used to measure the difference, similarity, and correspondence between two probability distributions, respectively. Here, we detail the calculation of each metric. Specifically, we first feed the prediction $\mathcal{P} \in \mathbb{R}^{H \times W}$ and ground truth $\mathcal{M} \in \mathbb{R}^{H \times W}$ to a min-max normalization. To compute KLD and SIM, input maps are divided by the sum of all elements:

$$\hat{\mathcal{P}}_i = \mathcal{P}_i / \sum \mathcal{P}, \quad \hat{\mathcal{M}}_i = \mathcal{M}_i / \sum \mathcal{M}. \tag{A.1}$$

Then, KLD and SIM are calculated as

$$\mathrm{KLD}(\hat{\mathcal{M}} \parallel \hat{\mathcal{P}}) = \sum_i \hat{\mathcal{M}}_i \cdot log(\frac{\hat{\mathcal{M}}_i}{\hat{\mathcal{P}}_i}), \tag{A.2}$$

$$\mathrm{SIM}(\hat{\mathcal{P}}, \hat{\mathcal{M}}) = \sum_i min(\hat{\mathcal{P}}_i, \hat{\mathcal{M}}_i). \tag{A.3}$$

For NSS, the input maps are first processed as follows:

$$\bar{\mathcal{M}} = \mathbb{1}(\mathcal{M} > 0.1), \quad \bar{\mathcal{P}} = \frac{\mathcal{P} - \mu(\mathcal{P})}{\sigma(\mathcal{P})}, \tag{A.4}$$

where $\mathbb{1}(\cdot)$ is an indicator function, μ and σ denote the arithmetic mean and standard deviation of $\mathcal{P}$, respectively. NSS is then computed as the average normalized prediction at binary GT locations:

$$\mathrm{NSS}(\bar{\mathcal{P}}, \bar{\mathcal{M}}) = \frac{1}{\sum \bar{\mathcal{M}}} \sum_i \bar{\mathcal{P}} \cdot \bar{\mathcal{M}}_i. \tag{A.5}$$

Details of the Unseen Setting. For the unseen setting in the AGD20K dataset [85], there are 35 object classes in the training set and 12 classes in the test set. It is worth noting that there is no object category intersection between training and test sets, so the model can learn how humans interact with objects from the training set and generalize the ability to novel objects in the test set. In Tab. A.1, we show the object categories in training and test set, respectively.

Train	apple, badminton racket, baseball, baseball bat, basketball, bench, book, bottle, bowl, carrot, cell phone, chair, couch, discus, fork, frisbee, hammer, hot dog, javelin, keyboard, knife, microwave, motorcycle, orange, oven, punching bag, rugby ball, scissors, skateboard, snowboard, suitcase, surfboard, tennis racket, toothbrush, wine glass
Test	axe, banana, bed, bicycle, broccoli, camera, cup, golf clubs, laptop, refrigerator, skis, soccer ball

Table A.1: Training and test object categories under the unseen setting.

	Method	Big			Middle			Small		
		KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
Seen	EIL [89]	1.047	0.461	0.389	1.794	0.284	0.710	3.057	0.123	0.231
	SPA [107]	5.745	0.317	0.222	4.990	0.228	0.440	6.076	0.118	0.297
	TS-CAM [41]	1.039	0.424	0.166	1.814	0.248	0.401	2.652	0.132	0.352
	Hotspots [97]	0.986	0.448	0.408	1.738	0.265	0.672	2.587	0.149	0.683
	Cross-view-AG [85]	0.766	0.533	0.652	1.485	0.322	1.040	2.373	0.175	0.927
	Cross-view-AG+* [87]	0.787	0.521	0.660	1.481	0.314	1.089	2.381	0.167	0.959
	LOCATE (Ours)	**0.676**	**0.580**	**0.706**	**1.178**	**0.390**	**1.316**	**2.029**	**0.216**	**1.349**
Unseen	EIL [89]	1.199	0.393	0.271	1.906	0.246	0.482	3.082	0.113	0.116
	SPA [107]	8.299	0.259	0.254	6.938	0.186	0.333	7.784	0.095	0.144
	TS-CAM [41]	1.238	0.351	0.072	1.970	0.208	0.236	2.766	0.113	0.124
	Hotspots [97]	1.015	0.425	0.548	1.872	0.242	0.605	2.693	0.134	0.544
	Cross-view-AG [85]	0.884	0.500	0.728	1.595	0.303	0.945	2.558	0.147	0.692
	Cross-view-AG+* [87]	0.867	0.485	0.776	1.658	0.279	0.988	2.630	0.133	0.754
	LOCATE (Ours)	**0.571**	**0.629**	**0.956**	**1.302**	**0.373**	**1.257**	**2.223**	**0.189**	**1.071**

Table A.2: Comparison to state-of-the-art methods on different object scales. The test set is divided into three subsets (Big, Middle and Small) based on the ratio of the mask to the image. The **best** and second-best results are highlighted in bold and underlined, respectively (↑/↓ means higher/lower is better). The symbol * indicates that we reproduce the results using the official code.

A.3 Additional Experimental Results

Comparison on Different Scales. To investigate the effect of different affordance region scales on the model, we follow [85] to split the test set into three subsets: Big, Middle and Small. The egocentric images in the "Big" subset have large affordance regions (the proportion of the mask to the image content is larger than 0.1), while the "Small" subset contains samples with fairly small affordance region (mask ratio is below 0.03), which is challenging to make accurate prediction. The remaining samples will be classified to the "Middle" subset. The results are shown in Tab. A.2, LOCATE

Loss	Seen			Unseen		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
MSE	1.395	0.357	1.112	1.648	0.301	1.033
Triplet	1.281	0.399	1.114	1.426	0.382	1.121
Cos†	1.240	0.400	1.156	1.418	0.370	1.148
Cos	1.226	0.401	1.177	1.405	0.372	1.157

Table A.3: Ablation study on the choice of loss functions. Cos† denotes the cosine embedding loss without margin.

τ	Seen			Unseen		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
0.4	1.232	0.397	1.178	1.409	0.368	1.179
0.5	1.229	0.400	1.177	1.405	0.370	1.165
0.6	1.226	0.401	1.177	1.405	0.372	1.157
0.7	1.226	0.400	1.176	1.414	0.372	1.140
0.8	1.239	0.400	1.159	1.423	0.373	1.125

Table A.4: Ablation study on the localization map threshold τ.

achieves the best performance among all the other methods on all scales and metrics.

Ablation Study on Loss Function. In LOCATE, we use cosine embedding loss to perform the supervision, which pulls the egocentric embedding towards the direction of the selected prototype. To explore the impact of different objective functions, we show the performance with different loss functions in Tab. A.3. For triplet loss, we select the prototype with the lowest PartIoU as the negative example. Results show that cosine embedding loss achieves the best results, and the margin can compensate for domain gaps to further improve performance.

Ablation Study on Hyper-parameters. There are two main hyper-parameters in LOCATE. The first one is τ which controls the portion of extracted exocentric feature embeddings. A larger τ leads to more localized extraction and fewer embeddings. The other one is μ, which indicates the confidence that the selected prototype represents the object part. Ablation results of these two hyper-parameters are displayed in Tab. A.4 and Tab. A.5, respectively. Our model is not sensitive to the choice of hyper-parameters, as results only vary within a small range. We set (τ, μ) to 0.6 as the

μ	Seen			Unseen		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
0.4	1.228	0.399	1.180	1.406	0.369	1.161
0.5	1.232	0.399	1.177	1.404	0.371	1.161
0.6	1.226	0.401	1.177	1.405	0.372	1.157
0.7	1.235	0.400	1.163	1.405	0.372	1.167
0.8	1.230	0.400	1.163	1.411	0.372	1.149

Table A.5: Ablation study on the PartIoU threshold μ.

final setting.

A.4 Additional Visualizations

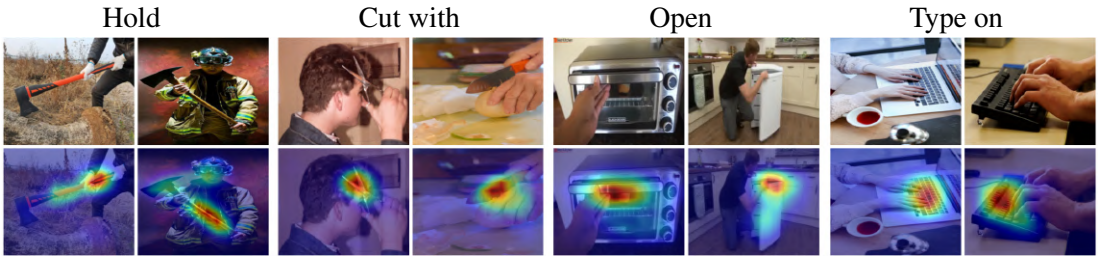


Figure A.2: Localization maps for exocentric images.

Exocentric Localizations. In LOCATE, we first extract feature embeddings from highly activated positions in the exocentric localization maps, allowing the model to focus more on object parts. In Fig. A.2, we show some examples of exocentric localization maps. We find these maps mainly focus on the interaction regions, which contain strong affordance-specific knowledge of how humans interact with objects.

Part-aware Features in DINO-ViT. Our framework is built on the deep feature of a self-supervised vision transformer (DINO-ViT [11]), whose part-aware features can provide good semantic correspondences across images [1]. We provide some illustrations in Fig. A.3. It can be seen that the features of DINO-ViT can help find the object parts involved in the exocentric interaction.

Additional Qualitative Results. In Fig. A.4, we show more qualitative comparison with state-of-the-art methods. In the seen setting, state-of-the-art methods can locate the general affordance area, but the predicted heatmaps are very coarse with blurred

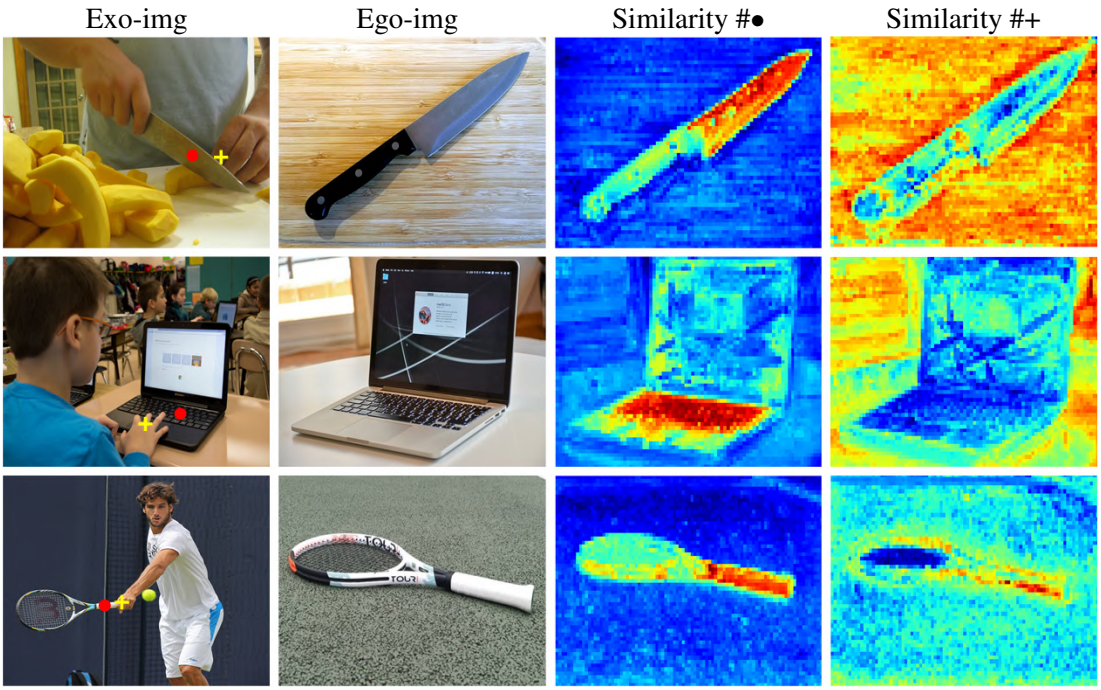


Figure A.3: Similarity maps computed between exocentric embeddings corresponding to the dot/cross and all egocentric features. Here dots and crosses are placed on positions of object parts and humans, respectively.

boundaries. In comparison, LOCATE performs much better with more part-focused results. As for the unseen setting, most state-of-the-art approaches often locate the wrong affordance region, while our results consistently show better performance.

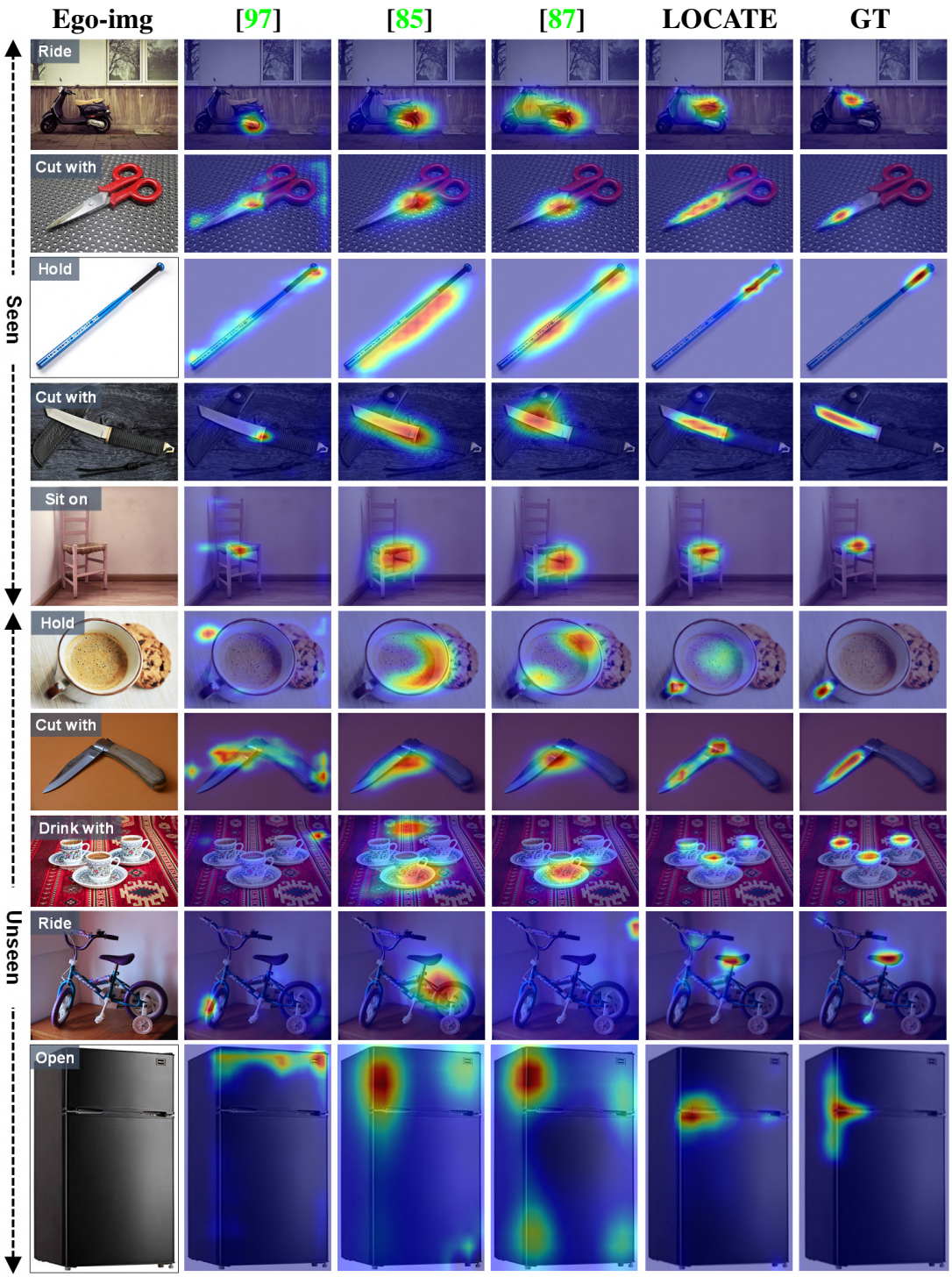


Figure A.4: Qualitative comparison between LOCATE and state-of-the-art affordance grounding methods (Hotspots [97], Cross-view-AG [85], and Cross-view-AG+ [87]) in both seen and unseen settings.

Appendix B

Data-Limited Vision-Language Affordance Learning

B.1 Dataset Details

Dataset	Affordance	Object
UMD	(7) grasp, cut, scoop, contain, pound, support, wrap-grasp	(17) bowl, cup, hammer, knife, ladle, mallet, mug, pot, saw, scissors, scoop, shears, shovel, spoon, tenderizer, trowel, turner
AGD20K	(37) beat, boxing, brush with, carry, catch, cut, cut with, drag, drink with, eat, hit, hold, jump, kick, lie on, lift, look out, open, pack, peel, pick up, pour, push, ride, sip, sit on, stick, stir, swing, take photo, talk on, text on, throw, type on, wash, write	(50) apple, axe, badminton racket, banana, baseball, baseball bat, basketball, bed, bench, bicycle, binoculars, book, bottle, bowl, broccoli, camera, carrot, cell phone, chair, couch, cup, discus, drum, fork, frisbee, golf clubs, hammer, hot dog, javelin, keyboard, knife, laptop, microwave, motorcycle, orange, oven, pen, punching bag, refrigerator, rugby ball, scissors, skateboard, skis, snowboard, soccer ball, suitcase, surfboard, tennis racket, toothbrush, wine glass

Table B.1: Affordance and object classes in the UMD and AGD20K dataset. The number of classes is shown in parentheses.

To evaluate the model's generalization ability in the challenging One-shot Open Affordance Learning (OOAL) setting, datasets with a large number of object categories are required. In addition, at least two object categories are needed for each affordance so that the model can be trained on one object and tested on the other. After an investigation of existing affordance datasets, we find only two datasets, AGD20K [85] and UMD [96], that fulfill the prerequisites and can be used to evaluate the affordance segmentation task. Specific affordance and object categories of these two datasets are shown in Tab. B.1. For the unseen split, we display the object category division in

Dataset	Base Objects (Train)	Novel Objects (Test)
UMD	(8) bowl, hammer, knife, mallet, mug, scissors, spoon, turner	(9) cup, ladle, pot, saw, scoop, shears, shovel, tenderizer, trowel
AGD20K	(33) apple, badminton racket, baseball, baseball bat, bench, book, bottle, bowl, carrot, cell phone, chair, couch, discus, fork, frisbee, hammer, hot dog, javelin, keyboard, microwave, motorcycle, orange, oven, punching bag, rugby ball, scissors, skateboard, snowboard, suitcase, surfboard, tennis racket, toothbrush, wine glass	(14) axe, banana, basketball, bed, bicycle, broccoli, camera, cup, golf clubs, knife, laptop, refrigerator, skis, soccer ball

Table B.2: Object category division in the unseen split of UMD and AGD20K dataset. The number of categories is shown in parentheses.

Tab. B.2. The model is trained on base object classes, and evaluated on novel objects categories.

Moreover, it is worth noting that annotations in AGD20K and UMD are of different types. UMD uses pixel-level dense binary maps, while the ground truth of AGD20K consist of sparse keypoints within the affordance areas, and a gaussian distribution is then applied on each point to generate dense annotation. The difference of dense and sparse affordance annotation is highlighted in Fig. B.1.

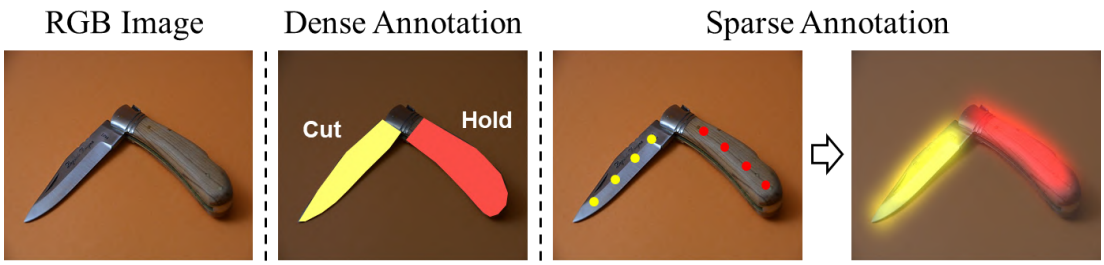


Figure B.1: Different affordance annotation schemes. Dense affordance annotation is labeled as binary masks. Sparse affordance annotation is first labeled as keypoints, and then a gaussian kernel is performed over each point to produce pixel-wise ground truth.

B.2 Ablation Study on Hyperparameters

The proposed framework involves three primary hyperparameters, *i.e.*, the number of learnable text tokens p, vision encoder fusion layers j, decoder transformer layers t.

p	Seen			Unseen		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
2	0.774	0.568	1.710	1.119	0.457	1.434
4	0.765	0.573	1.714	1.102	**0.469**	1.449
6	0.760	0.572	1.726	1.162	0.440	1.383
8	**0.740**	0.577	**1.745**	**1.070**	0.461	**1.503**
10	0.768	**0.581**	1.726	1.111	0.460	1.463

Table B.3: Ablation study on the number of learnable token p in text prompt learning.

j	Seen			Unseen		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
1	1.060	0.455	1.422	1.338	0.390	1.302
2	0.748	0.576	**1.756**	1.105	0.456	1.452
3	**0.740**	0.577	1.745	**1.070**	**0.461**	**1.503**
4	0.762	**0.579**	1.713	1.129	0.453	1.401

Table B.4: Ablation study on the number of fusion layers j in multi-layer feature fusion.

t	Seen			Unseen		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
0	0.846	0.537	1.622	1.115	0.447	1.440
1	0.753	0.574	1.737	1.094	0.449	**1.492**
2	**0.740**	**0.577**	**1.745**	**1.067**	**0.465**	**1.492**
3	0.746	0.575	1.738	1.110	0.458	1.456

Table B.5: Ablation study on the number of transformer decoder layers t.

We conduct ablation studies individually to explore the impact of these hyperparameters, as detailed in Tab. B.3, Tab. B.4, and Tab. B.5. Notably, increasing the number of learnable text tokens up to 8 showcases a gradual improvement in performance within the seen setting, but leads to fluctuating results in the unseen setting, indicating its susceptibility to generalization when confronted with unseen objects. In terms of the fusion layers, the fusion of the last two layers demonstrates an obvious performance gain compared to the single-layer counterpart, and integrating the last three layers

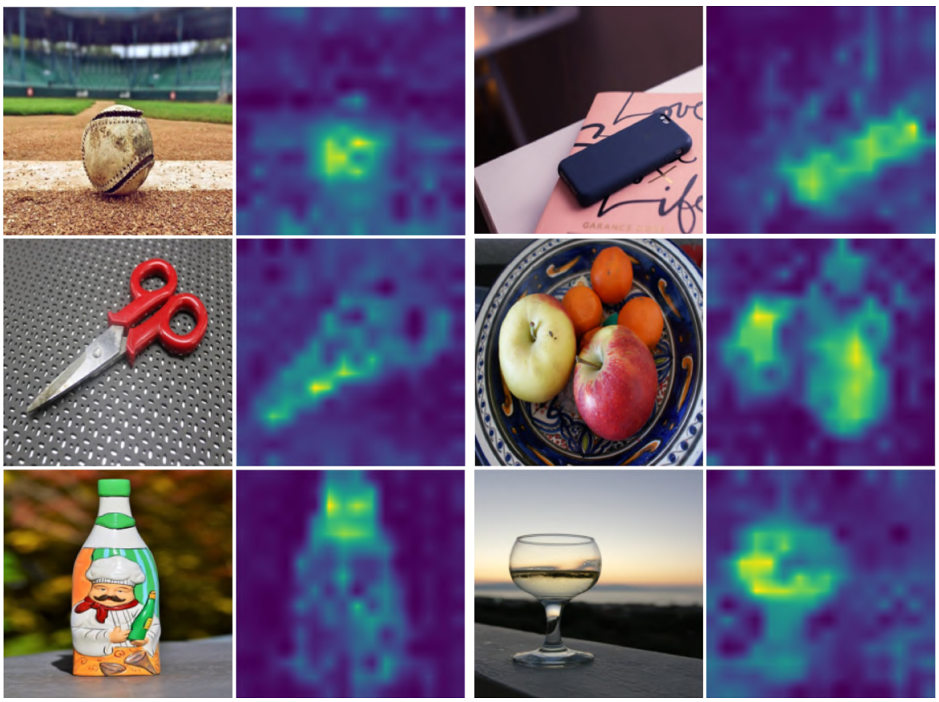


Figure B.2: Visualization of CLS-guided mask.

yields the best results. Lastly, we note that the transformer decoder can effectively improve performance in both seen and unseen setting, and a two-layer transformer decoder produces the most optimal results.

B.3 Additional Visualizations

Visualization of CLS-guided Mask. In Fig. B.2, we display the visualization of the CLS-guided mask from the proposed CLS-guided transformer decoder. It can be seen that the mask primarily concentrates on foreground objects, thus facilitating the cross-attention within salient regions.

Visualization of Unseen Affordances. In Fig. B.3, we further display examples on AGD20K dataset to showcase that our model has the ability to recognize unseen affordances. It is evident that the model can consistently activate relevant affordance areas when receiving text that are previously unseen during training.

Additional Qualitative Results. In Fig. B.4, we present more qualitative results on AGD20K dataset. The comparison demonstrates that predictions from our methods exhibit clear separation among object parts, while predictions from other approaches often bias towards one part or the whole object. In particular, our methods can locate very fine-grained affordance areas even for unseen objects, such as the saddle of a

bicycle for “sit on”, and the handle of a golf club for “hold”.

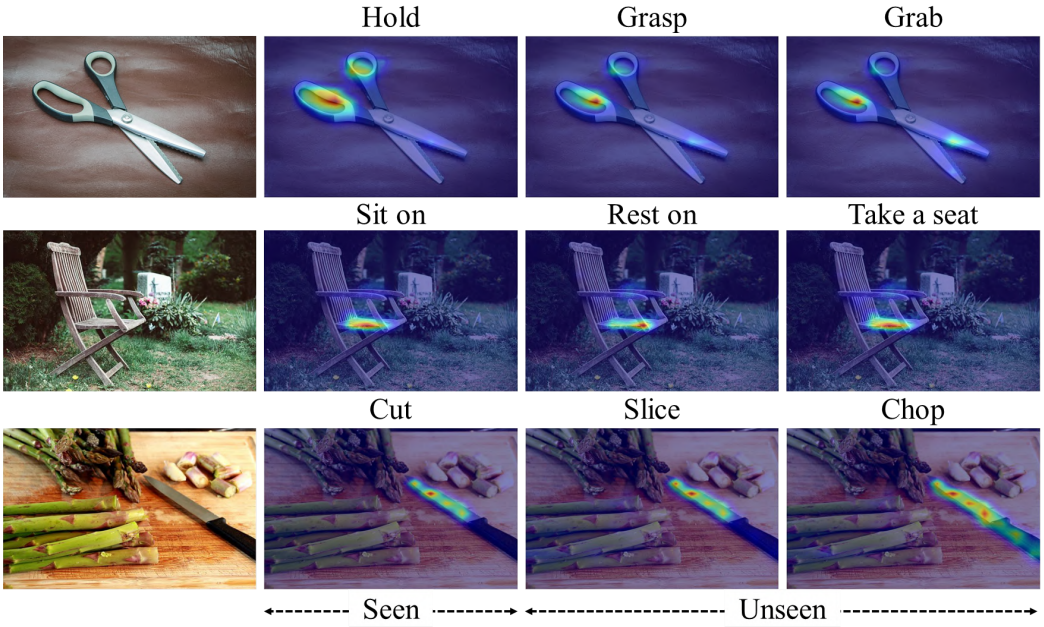


Figure B.3: Qualitative examples of unseen affordance prediction on AGD20K dataset. The 2nd column shows the results on seen affordances, and the 3rd and 4th columns show results with unseen affordances.

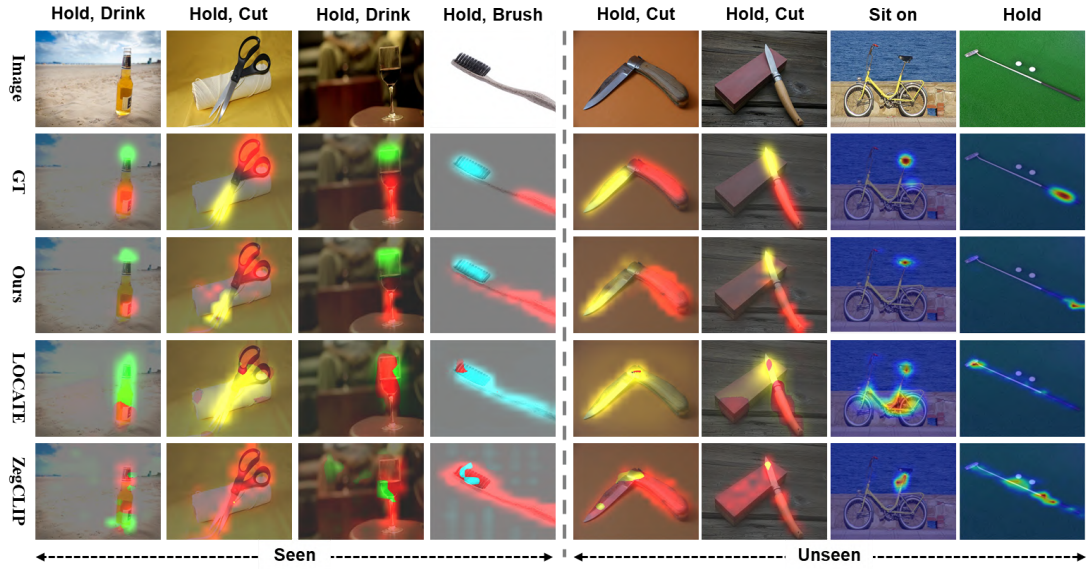


Figure B.4: Additional qualitative comparison on AGD20K dataset.

Appendix C

An Affordance Learning System for Robotic Manipulation

C.1 Dataset Details

In Fig. C.1, we present examples from existing affordance datasets alongside our proposed Affordance Evaluation Dataset (AED), highlighting the necessity for a new evaluation dataset. UMD [96] is collected in a fixed lab environment with coarse annotations. RGBD-AFF [58] has very low resolution and clean background. IIT-AFF [101] includes humans and also annotates occluded object parts. AGD20K [85] is annotated with keypoints and transformed to coarse heatmaps with a Gaussian kernel. In contrast, AED contains natural images with pixel-wise annotations. The statistics regarding the number of images per object category are listed in Tab. C.1.

Figure C.1: Examples from existing affordance datasets.

Total	knife	cup	scissors	hammer	fork	screwdriver	spatula	ladle	pan	shovel	spoon	drill	trowel
721	156	95	78	78	72	59	46	46	31	22	18	10	10

Table C.1: Statistics on the number of images for each object on the AED.

(a) Robot experiment setup.

(b) Experimental objects.

Figure C.2: (a) Experimental setup. (b) Seen (left) and unseen (right) objects used in the experiments.

C.2 Implementation Details

Affordance Data Collection. We gather training data from two large-scale egocentric video datasets: Epic-kitchens [29] and Ego4d [44]. We utilize narratives to collect data of 9 object categories from Epic-kitchens, and 12 classes from Ego4d, resulting in a total of 13 object classes. For graspable point localization, correspondences between the pre-contact and contact frames are detected using the SURF descriptor [5], and the homography is then estimated by sampling at least four pairs of points with the RANSAC [40] algorithm to maximize the number of inliers. For functional point localization, the IoU threshold is set to 0.3 to detect the pre-contact frame. We set

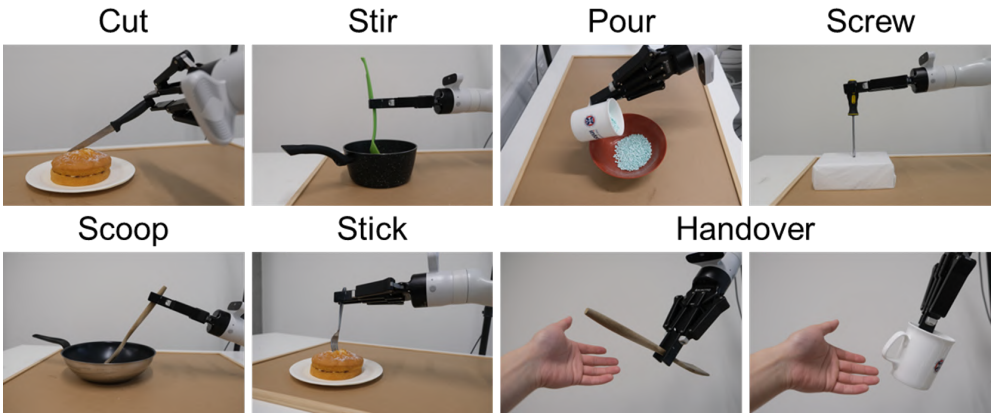


Figure C.3: Illustrations of 7 tasks in the robot experiments.

the detection thresholds to 0.1 for the hand-object detector and 0.35 for the open-vocabulary detector.

Vision Experiments. All experiments are conducted on two GeForce RTX 3090 GPUs using the Adamw [81] optimizer, with a learning rate of 1e−3 and batch size 8 for 15 epochs. DINOv2-base is used as the feature extractor. Collected images are first resized to 476×476 and then randomly cropped to 448×448. Both horizontal and vertical flipping are used for data augmentation. During training, LoRA is applied to all query, key, and value projection layers in the transformer block. Focal and Dice losses are used as training objectives:

$$\mathcal{L}_{focal} = -\frac{1}{n}\sum_{i=1}^{n}\left[(1-\hat{y}_i)^{\gamma}\cdot\hat{y}_i\log(y_i)+\hat{y}_i^{\gamma}\cdot(1-\hat{y}_i)\log(1-y_i)\right], \tag{C.1}$$

$$\mathcal{L}_{dice} = 1-\frac{2\sum_i^n y_i\hat{y}_i+\varepsilon}{\sum_i^n y_i+\sum_i^n \hat{y}_i+\varepsilon}, \tag{C.2}$$

where n is the number of valid pixels in output, $\gamma = 2$ is a focusing parameter to balance easy and hard samples, and $\varepsilon = 1$ is a smoothing factor that prevents division by zero and stabilizes the training. y and $\hat{y}$ represent the predicted probabilities and ground truth, respectively. Three metrics, mean intersection-over-union (mIoU), F1-score (F1), and accuracy (Acc), are adopted for evaluation.

Robot Experiments. To evaluate the effectiveness of learned visual affordances, we deploy Aff-Grasp in a 7 DoF Kinova Gen3 robot arm. The arm is equipped with a Robotiq 2F-85 parallel jaw gripper, and a calibrated Azure Kinect RGB-D camera is mounted next to the robot to capture the scene of the workspace (see Fig. C.2(a)). To enable open-vocabulary affordance recognition, we utilize CLIP text embeddings as the classifier, and discard the DFI to speed up inference time. Real-world experiments

are conducted with 34 diverse objects (shown in Fig. C.2(b)) and 7 tasks (see Fig. C.3) to evaluate three essential properties: accuracy, robustness, and generalization. We introduce them in detail as follows:

1. *Accuracy evaluation:* Given a single seen object on the workspace, we evaluate whether the model can recognize correct affordances of the object and perform the related affordance task. This evaluation is performed with 24 objects, each of which is repositioned 3 times during the experiment.

2. *Robustness evaluation:* Given multiple seen and unseen objects in a cluttered scene and an affordance task, we evaluate the model's ability to identify which object should be selected to perform the specific task. This requires the model to make robust predictions in the presence of distractors. The evaluation is conducted across 7 affordance tasks. Each task is tested with 3 diverse objects, except for the handover task, which is tested with 6 objects, each possessing different functional affordances. Every object is repositioned 3 times during the experiments.

3. *Generalization evaluation:* Given novel object categories not encountered during training, we evaluate if the model can still recognize the correct graspable areas. This evaluation assesses if the model can generalize the graspable affordance prediction to novel objects, which is a crucial factor in robotic manipulation. It is conducted with 7 novel objects, each repositioned 5 times.

We compare GAT with two relevant affordance grounding methods: LOCATE and Robo-ABC. LOCATE is an affordance grounding model that learns affordances from human-object interaction images using action labels as weak supervision. The method builds on DINO-ViT to identify object parts by clustering visual features from interaction regions of exocentric images, and then transfers the discovered parts to egocentric images for affordance grounding. Robo-ABC extracts object images and contact points from egocentric videos and stores these as an affordance memory. During inference, it first retrieves the most similar objects to the target and then utilizes semantic correspondence from the diffusion model to map the contact point to the current object.

Success rate is adopted as metric and reported from three aspects: correct affordance prediction, successful grasp, and successful interaction. For experiments in cluttered scenes, we assume that only one object is available to complete the target task. We do not perform manipulation policy learning, as it is beyond the focus of this

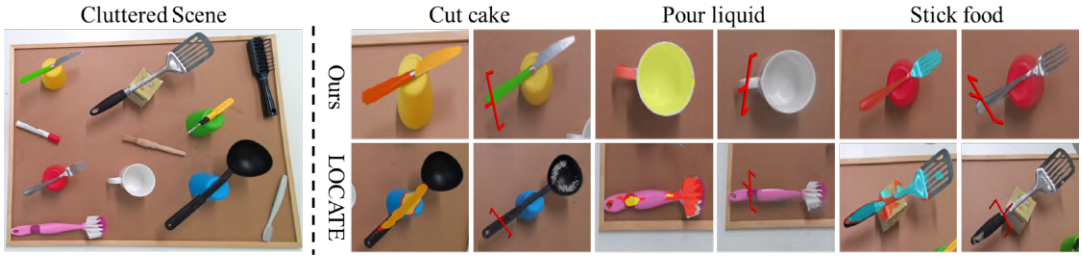
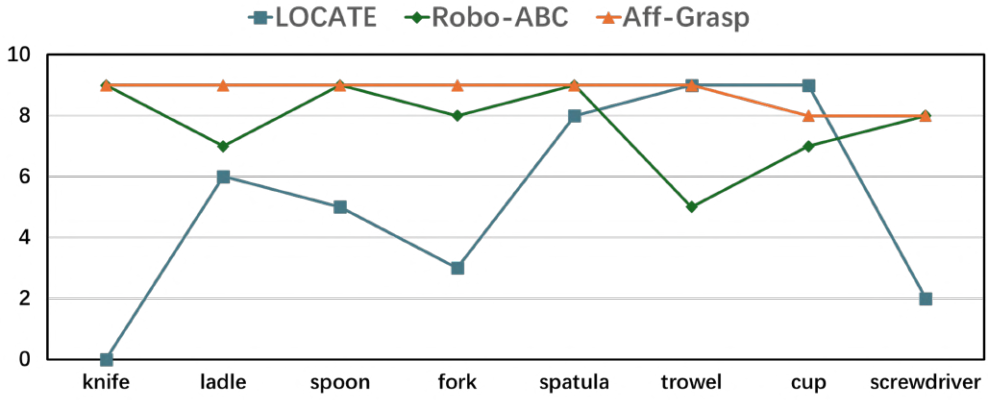
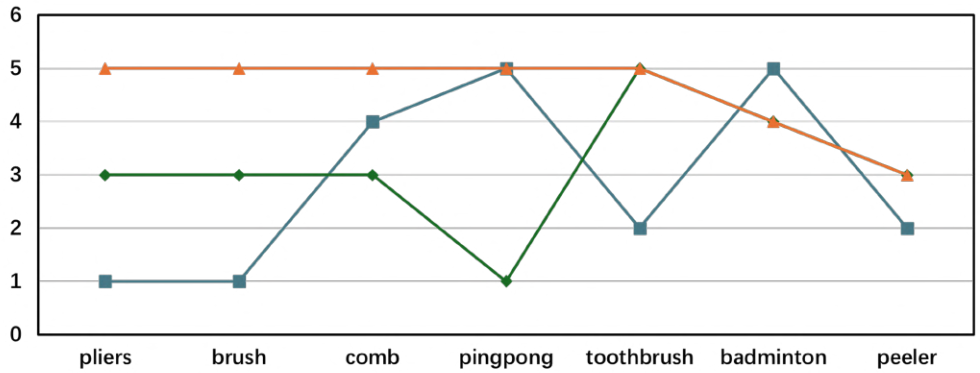


Figure C.4: Qualitative comparison of affordance prediction and final grasp pose for 3D point clouds in the cluttered scene. LOCATE fails to identify related objects for desired tasks, whereas Aff-Grasp can select the correct object with accurate affordance segmentation and is not affected by cluttered scenes.



(a) Seen classes in accuracy evaluation



(b) Unseen classes in generalization evaluation

Figure C.5: Success rates of correct affordance predictions for each individual object from the accuracy and generalization evaluations. The total numbers of trials are 9 and 5, respectively.

work. Instead, we design motion primitives for each affordance and assume that the operating direction of the tool is known. For instance, in the task of "stir in the pot", the ladle is first grasped and lifted to a height of 20 cm. Next, the gripper is rotated 90

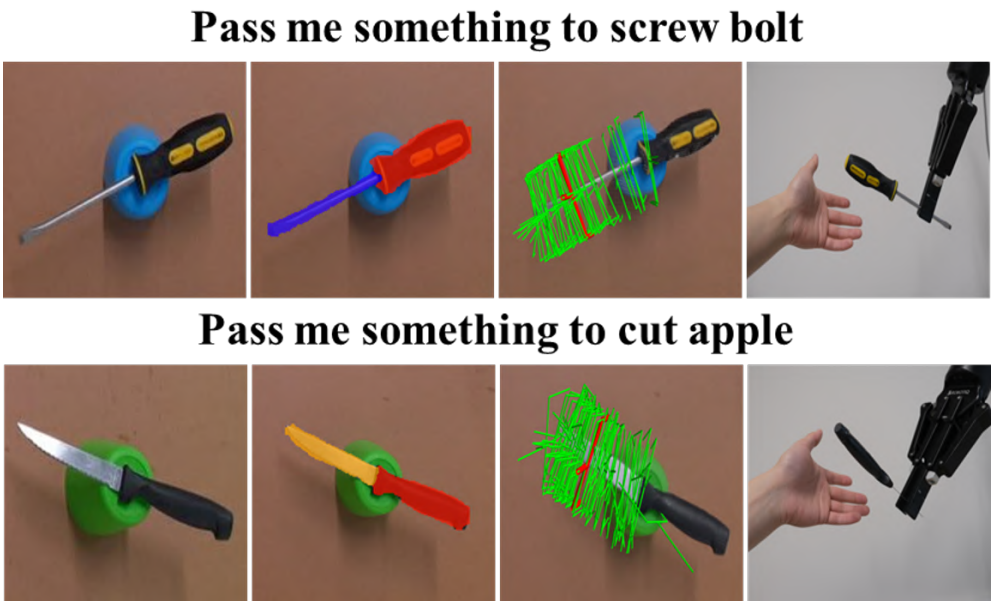


Figure C.6: The Aff-Grasp framework can perform the handover task by generating grasp poses within the functional parts of objects, and orienting the graspable parts towards the human hand. Green indicates all potential grasps, while red marks the final selected grasp.

degrees along its x-axis while simultaneously moving the ladle above the pot. Finally, it lowers the ladle to a certain distance and moves in a circular trajectory around the center of the pot.

C.3 Additional Experimental Results

Additional Results for Robot Experiments. In Fig. C.4, we show a visual example from the robustness evaluation, where both seen and unseen objects serve as interferences. The predicted affordance segmentation maps and corresponding grasp poses on point clouds are displayed. It is noted that LOCATE is unable to localize the correct object to execute the specified affordance task, while our model successfully identifies the matching object and predicts accurate segmentation maps. In Fig. C.5, we show success rates of individual classes for accuracy and generalization evaluations. It is evident that our results are accurate and stable over all categories, while the results of LOCATE and Robo-ABC show frequent fluctuations. Furthermore, we display affordance and grasp pose predictions for the handover task in Fig. C.6. When the robot is asked to pass something to the subject for a task, Aff-Grasp generates grasp propos-

Depth map	mIoU	F1	Accuracy
Color depth (jet)	61.82	76.29	78.34
Color depth (inferno)	62.38	76.57	77.31
Color depth (viridis)	63.92	77.81	78.95
Grayscale depth	64.66	78.35	79.74

Table C.2: Ablation study on different depth representations in DFI.

Embeddings	mIoU	F1	Accuracy
CLIP-B/32	66.47	79.70	79.31
CLIP-B/16	66.04	79.37	79.15
CLIP-L/14	66.91	80.02	81.09
Learnable embeds	68.62	81.09	83.51

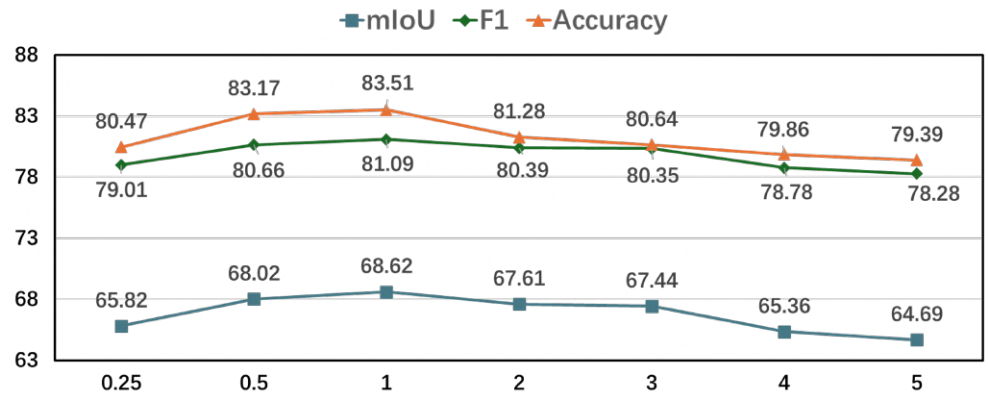
Table C.3: Ablation study on different classification embeddings: learnable or CLIP text embeddings.

als based on the functional affordance mask and directs graspable parts towards the subject's hand.

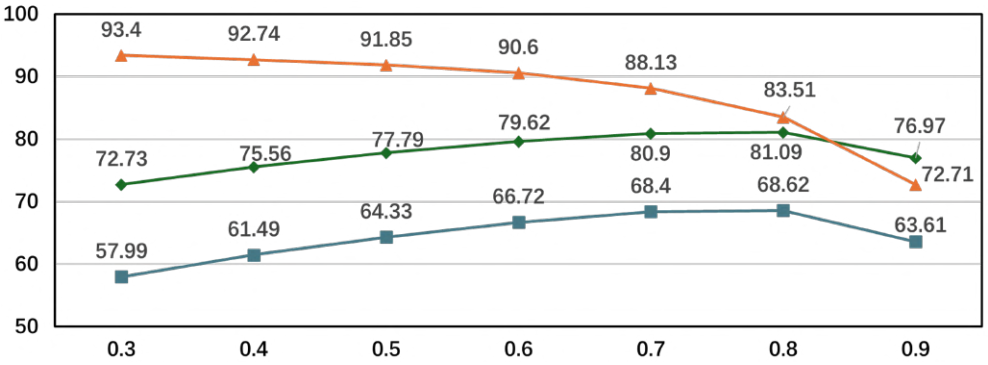
Additional Ablation Studies. To further understand the effectiveness of DFI module, we perform experiments using different depth maps as input. We observe that DFI is more effective with depth representations that have low contrast. As listed in Tab. C.2, the jet colormap, known for high-contrast visual effect, yields the worst results in DFI. In comparison, the less expressive grayscale depth achieves the best performance among other colored counterparts. We speculate that grayscale input focuses more on the geometric information, whereas color depth may introduce noise to some extent.

In Tab. C.3, we show the impact of different classification embeddings. The learnable embeddings yield the best results, but lose the ability to reason about unseen affordances. While performance degrades slightly when using CLIP text embeddings as classifiers, this approach retains the ability for open-vocabulary affordance segmentation. Therefore, we use learnable embeddings for vision evaluation and CLIP-L/14 text embeddings for robot experiments.

Finally, we conduct experiments with different hyer-parameter settings, focusing on the threshold τ for background classification and the weighting factor α in the



(a) Weighting factor α



(b) Threshold τ for background classification

Figure C.7: Ablation study on hyper-parameters.

loss function. As presented in Fig. C.7, the model obtains the highest performance in mIoU and F1-score with a weighting factor α of 1. For the background classification threshold τ, a smaller value leads to higher accuracy, as only confident predictions are counted as foreground. In this case, only mIoU and F1 score can truly reflect the performance. We thus choose 0.8 as the default threshold.

C.4 Discussion

Affordance vs. Part. One may argue that parts are more direct and explicit instructions than affordances, as actions or verbs are often more abstract than semantics or nouns. A spectrum of recent work [51, 110, 127, 130] also utilize open-vocabulary part segmentation models [122, 133] and large language models [104, 105] to specify the desired grasping parts for robots. However, understanding object affordances holds great significance for embodied intelligence. Firstly, human's instructions are typically high-level and abstract. For example, we would instruct a robot to "cut an apple for me", rather than specifying "grasp the knife handle and cut the apple with the knife

blade". Therefore, affordance understanding helps in the interpretation of natural instructions from humans. Second, reasoning about object parts from task instructions using large language models is time-consuming. A direct understanding of affordances can streamline the process by allowing robots to infer actionable areas from high-level instructions without extensive part-based prompting and reasoning. Thus, affordance-based approaches contribute to more intuitive and efficient interactions between robots and their environments, aligning more closely with how humans naturally communicate and perform tasks.

Points vs. Masks. In this work, we represent affordances as segmentation masks, whereas some related previous work [4, 57] represents them as points. While one may argue that using points instead of masks to acquire grasping poses is a more straightforward choice, we deem that masks are more robust and informative for the following reasons: (1) Predicting points is challenging due to their sparsity. Also, computing point correspondences is time-consuming and susceptible to variations in background and orientation. (2) A segmentation mask provides a broad region that, when combined with a grasp pose estimation model, can lead to the most confident grasp proposal. Point-based methods, on the other hand, heavily rely on the accuracy of point predictions and may fail if the predicted point is far from the object's center of mass.

Video Datasets. Although this work collects affordance data from egocentric videos, we observe that the same pipeline can also be applied to exocentric human-object interaction videos. This flexibility highlights the robustness and adaptability of our approach in different visualization perspectives. Egocentric videos provide a first-person viewpoint, which is highly beneficial for capturing the user's direct interactions with objects, allowing for a more intimate and precise understanding of affordances. On the other hand, exocentric videos, which capture interactions from a third-person perspective, can offer a comprehensive view of the context in which interactions occur. Additionally, video datasets collected in simple or laboratory environments [80, 145, 152] are preferable for ensuring high accuracy and usability of the training data. These controlled settings typically offer good lighting, background uniformity, and clear object boundaries, providing consistent and reliable data.

Bibliography

[1] Amir, S., Gandelsman, Y., Bagon, S., and Dekel, T. (2022). Deep vit features as dense visual descriptors. *ECCVW What is Motion For*.

[2] Ardón, P., Pairet, È., Lohan, K. S., Ramamoorthy, S., and Petrick, R. (2020). Affordances in robotic tasks–a survey. *arXiv preprint arXiv:2004.07400*.

[3] Ardón, P., Pairet, E., Petrick, R. P., Ramamoorthy, S., and Lohan, K. S. (2019). Learning grasp affordance reasoning through semantic relations. *IEEE Robotics and Automation Letters*, 4(4):4571–4578.

[4] Bahl, S., Mendonca, R., Chen, L., Jain, U., and Pathak, D. (2023). Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13778–13790.

[5] Bay, H., Ess, A., Tuytelaars, T., and Van Gool, L. (2008). Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359.

[6] Bharadhwaj, H., Gupta, A., and Tulsiani, S. (2023). Visual affordance prediction for guiding robot exploration. In *IEEE International Conference on Robotics and Automation*, pages 3029–3036.

[7] Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al. (2024). π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*.

[8] Borja-Diaz, J., Mees, O., Kalweit, G., Hermann, L., Boedecker, J., and Burgard, W. (2022). Affordance learning from play for sample-efficient policy learning. In *IEEE International Conference on Robotics and Automation*, pages 6372–6378.

[9] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are

few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

[10] Burke, C. J., Tobler, P. N., Baddeley, M., and Schultz, W. (2010). Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences*, 107(32):14431–14436.

[11] Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660.

[12] Chao, Y.-W., Liu, Y., Liu, X., Zeng, H., and Deng, J. (2018). Learning to detect human-object interactions. In *Winter conference on Applications of Computer Vision*, pages 381–389. IEEE.

[13] Chao, Y.-W., Wang, Z., He, Y., Wang, J., and Deng, J. (2015). Hico: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1017–1025.

[14] Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. (2018). Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *Winter conference on Applications of Computer Vision*, pages 839–847. IEEE.

[15] Chen, C., Cong, Y., and Kan, Z. (2024). Worldafford: Affordance grounding based on natural language instructions. In *36th International Conference on Tools with Artificial Intelligence*, pages 822–828. IEEE.

[16] Chen, J., Gao, D., Lin, K. Q., and Shou, M. Z. (2023a). Affordance grounding from demonstration video to target image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6799–6808.

[17] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision*, pages 801–818.

[18] Chen, Q., Yang, L., Lai, J.-H., and Xie, X. (2022a). Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4288–4298.

[19] Chen, Z., Duan, Y., Wang, W., He, J., Lu, T., Dai, J., and Qiao, Y. (2023b). Vision transformer adapter for dense predictions. In *The Eleventh International Conference on Learning Representations*.

[20] Chen, Z., Wang, T., Wu, X., Hua, X.-S., Zhang, H., and Sun, Q. (2022b). Class re-activation maps for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 969–978.

[21] Cho, S., Shin, H., Hong, S., Arnab, A., Seo, P. H., and Kim, S. (2024). Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123.

[22] Chu, H., Deng, X., Lv, Q., Chen, X., Li, Y., Hao, J., and Nie, L. (2025). 3d-affordancellm: Harnessing large language models for open-vocabulary affordance detection in 3d worlds. *The Thirteenth International Conference on Learning Representations*.

[23] Chuang, C.-Y., Li, J., Torralba, A., and Fidler, S. (2018). Learning to act properly: Predicting and explaining affordances from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 975–983.

[24] Contributors, M. (2020). MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation.

[25] Corona, E., Pumarola, A., Alenya, G., Moreno-Noguer, F., and Rogez, G. (2020). Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5031–5041.

[26] Cruz, F., Magg, S., Weber, C., and Wermter, S. (2016). Training agents with interactive reinforcement learning and contextual affordances. *IEEE Transactions on Cognitive and Developmental Systems*, 8(4):271–284.

[27] Cui, L., Chen, X., Zhao, H., Zhou, G., and Zhu, Y. (2023). Strap: Structured object affordance segmentation with point supervision. *arXiv preprint arXiv:2304.08492*.

[28] Damen, D., Doughty, H., Farinella, G. M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al. (2018). Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision*, pages 720–736.

[29] Damen, D., Doughty, H., Farinella, G. M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al. (2020). The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141.

[30] Damen, D., Doughty, H., Farinella, G. M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al. (2022). Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23.

[31] Delitzas, A., Takmaz, A., Tombari, F., Sumner, R., Pollefeys, M., and Engelmann, F. (2024). 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.

[32] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255.

[33] Deng, S., Xu, X., Wu, C., Chen, K., and Jia, K. (2021). 3d affordancenet: A benchmark for visual object affordance understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1778–1787.

[34] Di Palo, N. and Johns, E. (2024a). Dinobot: Robot manipulation via retrieval and alignment with vision foundation models. In *IEEE International Conference on Robotics and Automation*, pages 2798–2805.

[35] Di Palo, N. and Johns, E. (2024b). On the effectiveness of retrieval, alignment, and replay in manipulation. *IEEE Robotics and Automation Letters*, 9(3):2032–2039.

[36] Do, T.-T., Nguyen, A., and Reid, I. (2018). Affordancenet: An end-to-end deep learning approach for object affordance detection. In *IEEE International Conference on Robotics and Automation*, pages 5882–5889.

[37] Fang, H.-S., Wang, C., Fang, H., Gou, M., Liu, J., Yan, H., Liu, W., Xie, Y., and Lu, C. (2023). Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, 39(5):3929–3945.

[38] Fang, H.-S., Wang, C., Gou, M., and Lu, C. (2020). Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11444–11453.

[39] Fang, K., Wu, T.-L., Yang, D., Savarese, S., and Lim, J. J. (2018). Demo2vec: Reasoning object affordances from online videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2139–2147.

[40] Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395.

[41] Gao, W., Wan, F., Pan, X., Peng, Z., Tian, Q., Han, Z., Zhou, B., and Ye, Q. (2021). Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2886–2895.

[42] Geng, Y., An, B., Geng, H., Chen, Y., Yang, Y., and Dong, H. (2023). Rlafford: End-to-end affordance learning for robotic manipulation. In *IEEE International Conference on Robotics and Automation*, pages 5880–5886.

[43] Gibson, J. J. (1979). The ecological approach to visual perception.

[44] Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al. (2022). Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 18995–19012.

[45] Hadjivelichkov, D., Zwane, S., Agapito, L., Deisenroth, M. P., and Kanoulas, D. (2022). One-shot transfer of affordance regions? affcorrs! In *6th Annual Conference on Robot Learning*.

[46] Hassanin, M., Khan, S., and Tahtali, M. (2021). Visual affordance and function understanding: A survey. *ACM Computing Surveys (CSUR)*, 54(3):1–35.

[47] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778.

[48] Hinton, G., Vinyals, O., Dean, J., et al. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

[49] Hou, Z., Yu, B., Qiao, Y., Peng, X., and Tao, D. (2021). Affordance transfer learning for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 495–504.

[50] Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

[51] Huang, H., Lin, F., Hu, Y., Wang, S., and Gao, Y. (2024). Copa: General robotic manipulation through spatial constraints of parts with foundation models. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 9488–9495.

[52] Huang, W., Wang, C., Zhang, R., Li, Y., Wu, J., and Fei-Fei, L. (2023). Voxposer: Composable 3d value maps for robotic manipulation with language models. In *7th Annual Conference on Robot Learning*.

[53] Hung, W.-C., Jampani, V., Liu, S., Molchanov, P., Yang, M.-H., and Kautz, J. (2019). Scops: Self-supervised co-part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 869–878.

[54] Hurst, A. et al. (2024). Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

[55] Jang, J. H., Seo, H., and Chun, S. Y. (2024). Intra: Interaction relationship-aware weakly supervised affordance grounding. In *Proceedings of the European Conference on Computer Vision*, pages 18–34. Springer.

[56] Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

[57] Ju, Y., Hu, K., Zhang, G., Zhang, G., Jiang, M., and Xu, H. (2024). Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation. In *Proceedings of the European Conference on Computer Vision*, pages 222–239. Springer.

[58] Khalifa, Z. and Shah, S. A. A. (2023). A large scale multi-view rgbd visual affordance learning dataset. In *IEEE International Conference on Image Processing*, pages 1325–1329.

[59] Khazatsky, A. et al. (2024). DROID: A large-scale in-the-wild robot manipulation dataset. In *RSS 2024 Workshop: Data Generation for Robotics*.

[60] Kim, E., Kim, S., Lee, J., Kim, H., and Yoon, S. (2022). Bridging the gap between classification and localization for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14258–14267.

[61] Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al. (2024). Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.

[62] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026.

[63] Kokic, M., Stork, J. A., Haustein, J. A., and Kragic, D. (2017). Affordance detection for task-specific grasping using deep learning. In *IEEE-RAS 17th International Conference on Humanoid Robotics*, pages 91–98.

[64] Koppula, H. S., Gupta, R., and Saxena, A. (2013). Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970.

[65] Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., and Jia, J. (2024). Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589.

[66] Lee, S., Lee, M., Lee, J., and Shim, H. (2021). Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5495–5505.

[67] Li, G., Jampani, V., Sun, D., and Sevilla-Lara, L. (2023). Locate: Localize and transfer object parts for weakly supervised affordance grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10922–10931.

[68] Li, G., Sun, D., Sevilla-Lara, L., and Jampani, V. (2024a). One-shot open affordance learning with foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3086–3096.

[69] Li, G., Tsagkas, N., Song, J., Mon-Williams, R., Vijayakumar, S., Shao, K., and Sevilla-Lara, L. (2024b). Learning precise affordances from egocentric videos for robotic manipulation. *arXiv preprint arXiv:2408.10123*.

[70] Li, J., Li, D., Xiong, C., and Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

[71] Li, J., Zhu, Y., Tang, Z., Wen, J., Zhu, M., Liu, X., Li, C., Cheng, R., Peng, Y., and Feng, F. (2024c). Improving vision-language-action models via chain-of-affordance. *arXiv preprint arXiv:2412.20451*.

[72] Li, Y., Liu, M., and Rehg, J. M. (2018). In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European Conference on Computer Vision*, pages 619–635.

[73] Li, Y., Wang, H., Duan, Y., Zhang, J., and Li, X. (2025). A closer look at the explainability of contrastive language-image pre-training. *Pattern Recognition*, 162:111409.

[74] Li, Y., Zhao, N., Xiao, J., Feng, C., Wang, X., and Chua, T.-s. (2024d). Laso: Language-guided affordance segmentation on 3d object. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14251–14260.

[75] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2980–2988.

[76] Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023). Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916.

[77] Liu, M., Tang, S., Li, Y., and Rehg, J. M. (2020). Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *Proceedings of the European Conference on Computer Vision*, pages 704–721. Springer.

[78] Liu, S., Tripathi, S., Majumdar, S., and Wang, X. (2022a). Joint hand motion and interaction hotspots prediction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3282–3292.

[79] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al. (2024). Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision*, pages 38–55. Springer.

[80] Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., and Yi, L. (2022b). Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21013–21022.

[81] Loshchilov, I. and Hutter, F. (2019). Decoupled weight decay regularization. In *International Conference on Learning Representations*.

[82] Lu, L., Zhai, W., Luo, H., Kang, Y., and Cao, Y. (2023). Phrase-based affordance detection via cyclic bilateral interaction. *IEEE Transactions on Artificial Intelligence*, 4(5):1186–1198.

[83] Luddecke, T. and Worgotter, F. (2017). Learning to segment affordances. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 769–776.

[84] Luo, H., Zhai, W., Zhang, J., Cao, Y., and Tao, D. (2021). One-shot affordance detection. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 895–901.

[85] Luo, H., Zhai, W., Zhang, J., Cao, Y., and Tao, D. (2022). Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2252–2261.

[86] Luo, H., Zhai, W., Zhang, J., Cao, Y., and Tao, D. (2023). Leverage interactive affinity for affordance learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6809–6819.

[87] Luo, H., Zhai, W., Zhang, J., Cao, Y., and Tao, D. (2024). Grounded affordance from exocentric view. *International Journal of Computer Vision*, 132(6):1945–1969.

[88] Ma, T., Wang, Z., Zhou, J., Wang, M., and Liang, J. (2024). Glover: Generalizable open-vocabulary affordance reasoning for task-oriented grasping. *arXiv preprint arXiv:2411.12286*.

[89] Mai, J., Yang, M., and Luo, W. (2020). Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8766–8775.

[90] Mi, J., Tang, S., Deng, Z., Goerner, M., and Zhang, J. (2019). Object affordance based multimodal fusion for natural human-robot interaction. *Cognitive Systems Research*, 54:128–137.

[91] Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D vision*, pages 565–571. IEEE.

[92] Mon-Williams, R., Li, G., Long, R., Du, W., and Lucas, C. G. (2025). Embodied large language models enable robots to complete complex tasks in unpredictable environments. *Nature Machine Intelligence*, pages 1–10.

[93] Montesano, L., Lopes, M., Bernardino, A., and Santos-Victor, J. (2007). Affordances, development and imitation. In *IEEE 6th International Conference on Development and Learning*, pages 270–275.

[94] Mur-Labadia, L., Guerrero, J. J., and Martinez-Cantin, R. (2023a). Multi-label affordance mapping from egocentric vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5238–5249.

[95] Mur-Labadia, L., Martinez-Cantin, R., and Guerrero, J. J. (2023b). Bayesian deep learning for affordance segmentation in images. In *IEEE International Conference on Robotics and Automation*, pages 6981–6987.

[96] Myers, A., Teo, C. L., Fermüller, C., and Aloimonos, Y. (2015). Affordance detection of tool parts from geometric features. In *IEEE International Conference on Robotics and Automation*, pages 1374–1381.

[97] Nagarajan, T., Feichtenhofer, C., and Grauman, K. (2019). Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8688–8697.

[98] Nagarajan, T. and Grauman, K. (2020). Learning affordance landscapes for interaction exploration in 3d environments. *Advances in Neural Information Processing Systems*, 33:2005–2015.

[99] Nasiriany, S., Kirmani, S., Ding, T., Smith, L., Zhu, Y., Driess, D., Sadigh, D., and Xiao, T. (2024). Rt-affordance: Affordances are versatile intermediate representations for robot manipulation. *arXiv preprint arXiv:2411.02704*.

[100] Nguyen, A., Kanoulas, D., Caldwell, D. G., and Tsagarakis, N. G. (2016). Detecting object affordances with convolutional neural networks. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2765–2770.

[101] Nguyen, A., Kanoulas, D., Caldwell, D. G., and Tsagarakis, N. G. (2017). Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5908–5915.

[102] Nguyen, T., Vu, M. N., Vuong, A., Nguyen, D., Vo, T., Le, N., and Nguyen, A. (2023). Open-vocabulary affordance detection in 3d point clouds. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5692–5698.

[103] Oh, S. W., Lee, J.-Y., Xu, N., and Kim, S. J. (2019). Video object segmentation using space-time memory networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9226–9235.

[104] OpenAI (2023a). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

[105] OpenAI (2023b). Gpt-4v(ision) system card.

[106] Oquab, M. et al. (2024). DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*.

[107] Pan, X., Gao, Y., Lin, Z., Tang, F., Dong, W., Yuan, H., Huang, F., and Xu, C. (2021). Unveiling the potential of structure preserving for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11642–11651.

[108] Qian, S., Chen, W., Bai, M., Zhou, X., Tu, Z., and Li, L. E. (2024). Affordancellm: Grounding affordance from vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 7587–7597.

[109] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

[110] Rashid, A., Sharma, S., Kim, C. M., Kerr, J., Chen, L. Y., Kanazawa, A., and Goldberg, K. (2023). Language embedded radiance fields for zero-shot task-oriented grasping. In *7th Annual Conference on Robot Learning*.

[111] Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al. (2024). Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*.

[112] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

[113] Roy, A. and Todorovic, S. (2016). A multi-scale cnn for affordance segmentation in rgb images. In *Proceedings of the European Conference on Computer Vision*, pages 186–201. Springer.

[114] Sawatzky, J. and Gall, J. (2017). Adaptive binarization for weakly supervised affordance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision workshops*, pages 1383–1391.

[115] Sawatzky, J., Srikantha, A., and Gall, J. (2017). Weakly supervised affordance detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2795–2804.

[116] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 618–626.

[117] Shaban, A., Bansal, S., Liu, Z., Essa, I., and Boots, B. (2017). One-shot learning for semantic segmentation. In *British Machine Vision Conference*.

[118] Shan, D., Geng, J., Shu, M., and Fouhey, D. F. (2020). Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9869–9878.

[119] Shao, Y., Zhai, W., Yang, Y., Luo, H., Cao, Y., and Zha, Z.-J. (2025). Great: Geometry-intention collaborative inference for open-vocabulary 3d object affordance grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[120] Shen, W., Yang, G., Yu, A., Wong, J., Kaelbling, L. P., and Isola, P. (2023). Distilled feature fields enable few-shot language-guided manipulation. In *7th Annual Conference on Robot Learning*.

[121] Stark, M., Lies, P., Zillich, M., Wyatt, J., and Schiele, B. (2008). Functional object class detection based on learned affordance cues. In *Computer Vision Systems*, pages 435–444, Berlin, Heidelberg. Springer Berlin Heidelberg.

[122] Sun, P., Chen, S., Zhu, C., Xiao, F., Luo, P., Xie, S., and Yan, Z. (2023). Going denser with open-vocabulary part segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15453–15465.

[123] Sundermeyer, M., Mousavian, A., Triebel, R., and Fox, D. (2021). Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes. In *IEEE International Conference on Robotics and Automation*, pages 13438–13444.

[124] Szeliski, R. et al. (2007). Image alignment and stitching: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 2(1):1–104.

[125] Tang, C., Yu, J., Chen, W., and Zhang, H. (2021). Relationship oriented affordance learning through manipulation graph construction. *arXiv preprint arXiv:2110.14137*.

[126] Tang, Y., Huang, W., Wang, Y., Li, C., Yuan, R., Zhang, R., Wu, J., and Fei-Fei, L. (2025). Uad: Unsupervised affordance distillation for generalization in robotic manipulation. In *IEEE International Conference on Robotics and Automation*.

[127] Tong, E., Opipari, A., Lewis, S., Zeng, Z., and Jenkins, O. C. (2024). Oval-prompt: Open-vocabulary affordance localization for robot manipulation through llm affordance-grounding. *arXiv preprint arXiv:2404.11000*.

[128] Touvron, H., Cord, M., and Jégou, H. (2022). Deit iii: Revenge of the vit. In *Proceedings of the European Conference on Computer Vision*, pages 516–533. Springer.

[129] Tsagkas, N., Rome, J., Ramamoorthy, S., Mac Aodha, O., and Lu, C. X. (2024). Click to grasp: Zero-shot precise manipulation via visual diffusion descriptors. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 11610–11617.

[130] van Oort, T., Miller, D., Browne, W. N., Marticorena, N., Haviland, J., and Suenderhauf, N. (2024). Open-vocabulary part-based grasping. *arXiv preprint arXiv:2406.05951*.

[131] Vuong, Q., Levine, S., Walke, H. R., Pertsch, K., Singh, A., Doshi, R., Xu, C., Luo, J., Tan, L., Shah, D., et al. (2023). Open x-embodiment: Robotic learning datasets and rt-x models. In *Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@ CoRL2023*.

[132] Wang, Y., Zhang, M., Li, Z., Driggs-Campbell, K. R., Wu, J., Fei-Fei, L., and Li, Y. (2024). D3fields: Dynamic 3d descriptor fields for zero-shot generalizable robotic manipulation. In *ICRA 2024 Workshop on 3D Visual Representations for Robot Manipulation*.

[133] Wei, M., Yue, X., Zhang, W., Kong, S., Liu, X., and Pang, J. (2023). Ov-parts: Towards open-vocabulary part segmentation. *Advances in Neural Information Processing Systems*, 36:70094–70114.

[134] Wen, J., Zhu, Y., Li, J., Zhu, M., Tang, Z., Wu, K., Xu, Z., Liu, N., Cheng, R., Shen, C., et al. (2025). Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. *IEEE Robotics and Automation Letters*.

[135] Wu, H., Zhang, Z., Cheng, H., Yang, K., Liu, J., and Guo, Z. (2020). Learning affordance space in physical world for vision-based robotic object manipulation. In *IEEE International Conference on Robotics and Automation*, pages 4652–4658.

[136] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. (2021). Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090.

[137] Xie, J., Hou, X., Ye, K., and Shen, L. (2022). Clims: Cross language image matching for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4483–4492.

[138] Xiong, Y., Varadarajan, B., Wu, L., Xiang, X., Xiao, F., Zhu, C., Dai, X., Wang, D., Sun, F., Iandola, F., et al. (2024). Efficientsam: Leveraged masked image pretraining for efficient segment anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16111–16121.

[139] Xu, J., Hou, J., Zhang, Y., Feng, R., Zhao, R.-W., Zhang, T., Lu, X., and Gao, S. (2022). Cream: Weakly supervised object localization via class re-activation mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9437–9446.

[140] Xu, L., Gao, Y., Song, W., and Hao, A. (2024). Weakly supervised multimodal affordance grounding for egocentric images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6324–6332.

[141] Xu, M., Zhang, Z., Wei, F., Hu, H., and Bai, X. (2023). Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954.

[142] Xu, P. and Yadong, M. (2025). Weakly-supervised affordance grounding guided by part-level semantic priors. In *The Thirteenth International Conference on Learning Representations*.

[143] Xu, R., Zhang, J., Guo, M., Wen, Y., Yang, H., Lin, M., Huang, J., Li, Z., Zhang, K., Wang, L., et al. (2025). A0: An affordance-aware hierarchical model for general robotic manipulation. *arXiv preprint arXiv:2504.12636*.

[144] Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., and Zhao, H. (2024). Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381.

[145] Yang, L., Li, K., Zhan, X., Wu, F., Xu, A., Liu, L., and Lu, C. (2022). OakInk: A large-scale knowledge repository for understanding hand-object interaction. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[146] Yang, X., Ji, Z., Wu, J., and Lai, Y.-K. (2023a). Recent advances of deep robotic affordance learning: a reinforcement learning perspective. *IEEE Transactions on Cognitive and Developmental Systems*.

[147] Yang, Y., Zhai, W., Luo, H., Cao, Y., Luo, J., and Zha, Z.-J. (2023b). Grounding 3d object affordance from 2d interactions in images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10905–10915.

[148] Yoshida, T., Kurita, S., Nishimura, T., and Mori, S. (2024). Text-driven affordance learning from egocentric vision. *arXiv preprint arXiv:2404.02523*.

[149] Yu, Z., Huang, Y., Furuta, R., Yagi, T., Goutsu, Y., and Sato, Y. (2023). Fine-grained affordance annotation for egocentric hand-object interaction videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2155–2163.

[150] Yuan, W., Duan, J., Blukis, V., Pumacay, W., Krishna, R., Murali, A., Mousavian, A., and Fox, D. (2024). Robopoint: A vision-language model for spatial affordance prediction in robotics. In *8th Annual Conference on Robot Learning*.

[151] Zhai, W., Luo, H., Zhang, J., Cao, Y., and Tao, D. (2022). One-shot object affordance detection in the wild. *International Journal of Computer Vision*, 130(10):2472–2500.

[152] Zhan, X., Yang, L., Zhao, Y., Mao, K., Xu, H., Lin, Z., Li, K., and Lu, C. (2024). Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 445–456.

[153] Zhang, J., Huang, W., Peng, B., Wu, M., Hu, F., Chen, Z., Zhao, B., and Dong, H. (2025). Omni6dpose: a benchmark and model for universal 6d object pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision*, pages 199–216. Springer.

[154] Zhang, X., Wei, Y., Feng, J., Yang, Y., and Huang, T. S. (2018). Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1325–1334.

[155] Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). Pyramid scene parsing network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2881–2890.

[156] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2921–2929.

[157] Zhou, C., Loy, C. C., and Dai, B. (2022a). Extract free dense labels from clip. In *Proceedings of the European Conference on Computer Vision*, pages 696–712. Springer.

[158] Zhou, K., Yang, J., Loy, C. C., and Liu, Z. (2022b). Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

[159] Zhou, T., Zhang, M., Zhao, F., and Li, J. (2022c). Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4299–4309.

[160] Zhou, Z., Lei, Y., Zhang, B., Liu, L., and Liu, Y. (2023). Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11175–11185.

[161] Zhu, H., Kong, Q., Xu, K., Xia, X., Deng, B., Ye, J., Xiong, R., and Wang, Y. (2025). Grounding 3d object affordance with language instructions, visual observations and interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[162] Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al. (2023). Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR.