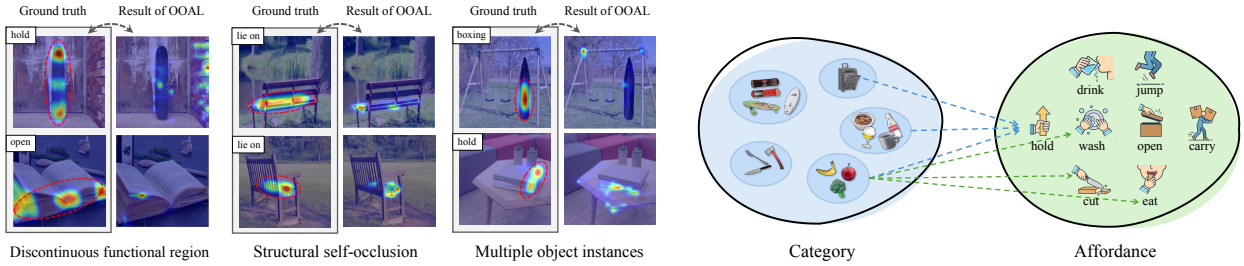
2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Visual category-guided one-shot open affordance grounding

Yangfan Wang[1], Hongyang Yu[2*], and Xiying Li[1*]



(a) Typical hard cases on the AGD20K dataset

(b) Function-appearance heterogeneity

Fig. 1: **Challenges in one-shot open affordance grounding.** (a) Difficult scenarios arising from discontinuous functional regions, structural self-occlusion, and multiple object instances. (b) Function-appearance heterogeneity: visually diverse objects (e.g., skateboards and suitcases) may share the same affordance "hold". Conversely, a single object (e.g., a piece of fruit) can possess multiple functional properties such as "eat", "cut", and "wash".

***Abstract*— Affordance grounding is a challenging task that aims to locate functional regions in object images enabling potential human-object interactions. One-shot open affordance grounding leverages the generalization capability of visual foundation models to overcome limitations of training data scale. However, existing methods often fail to locate functional regions in complex scenarios due to the lack of fine-grained perception, function-appearance heterogeneity, and the overfitting of affordance prompts to known categories. To improve generalization to unseen categories, we introduce a category-conditioned affordance prompt learning, which constructs a complete semantic category-affordance prompt from instance-level visual features. To further improve the accuracy of affordance localization for objects with complex structures, we propose a coarse-to-fine semantic-guided Transformer decoder. This design enhances the decoder's ability to understand the semantic mapping between the affordance words and corresponding object part-level regions. On multiple standard benchmarks, our method achieves competitive performance compared to related methods with less than 1% of the training cost. Notably, our approach shows more robust generalization to unseen objects and novel affordances than the recent SOTA baseline methods.**

I. Introduction

Affordance grounding aims to localize the fine-grained object parts that afford potential functional interactions. It can be an important technology for robotics [1], [2], [3], scene understanding [4], [5], human-object interaction [6], [7], [8], visual navigation [9], [10], and other related fields. To address the scarcity of annotated data and weak generalization inherent in conventional affordance grounding, **the one-shot open affordance grounding (OOAG)** task has attracted researchers' attention. This paradigm finetunes a visual foundation model ([11], [12]) using only one annotated sample per category, thereby minimizing training costs while substantially enhancing localization accuracy.

Benefiting from the excellent generalization capability of visual foundation models, OOAG methods can localize novel objects and associated affordance terms. For example, once the model learns that "the handle of an axe can be grasped", this knowledge can be transferred to semantically similar items, even for unseen categories during the training phase, such as knives and tennis rackets. Furthermore, OOAG frameworks can process queries semantically related to affordance concepts (e.g., 'hit' vs. 'strike'), overcoming the limitations of predefined query sets.

Despite significant progress, current OOAG methods still grapple with several critical challenges. First, existing models exhibit a notable deficiency in fine-grained perceptual precision. We evaluate the state-of-the-art OOAL on the AGD20K dataset [13], and the visual results of hard cases reveal a substantial degradation in localization accuracy within complex scenarios, such as those involving discontinuous functional regions, structural self-occlusion, or multiple object instances, as illustrated in Fig. 1(a). Second, OOAG

[1]School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen, China.
[2]Peng Cheng Laboratory, Shenzhen, China.
*Corresponding authors: Xiying Li and Hongyang Yu.
This work was supported by the National Natural Science Foundation of China (U21B2090, 62402251, 62472238) and the Guangdong Basic and Applied Basic Research Foundation (2022A1515010361).

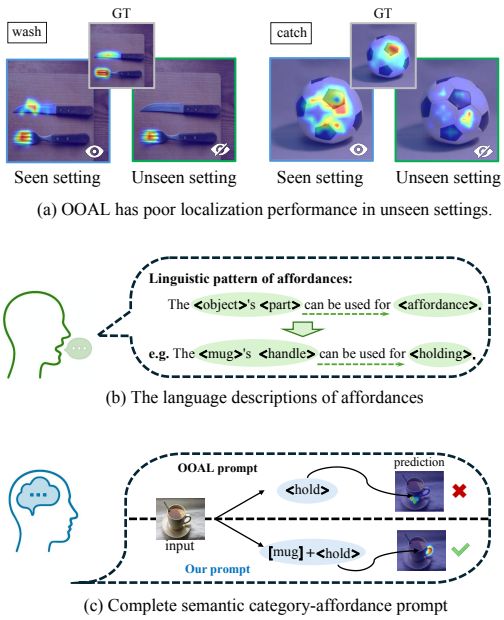


Fig. 2: The limitations of OOAL in unseen settings and our research motivation.

faces the challenge of "function-appearance heterogeneity". As illustrated in Fig. 1(b), knives, skateboards, and suitcases all support the same "hold" action although they possess significantly different appearances. Therefore, affordance localization methods require a robust semantic mapping between affordance concepts and object regions. Third, the current approach generally inserts randomly initialized context vectors that can be learned as part of the affordance prompt. However, we find that such affordance prompts can't generalize well to the unseen classes, even for the exact same test image, as shown by the example in Fig. 2(a).

Generally, the linguistic structure of an affordance can be conceptualized as: <object>'s <part> can be used for <affordance>. Following this pattern, we observe that existing methods neglect the utilization of visual category information. In fact, directly localizing target parts using only affordance terms remains challenging for vision-language models. Motivated by this, we propose category-conditioned affordance prompt learning. Instead of relying on static and handcrafted prompts, our approach encodes instance-level visual features as category tokens during the inference phase to construct a complete semantic category-affordance prompt, as illustrated in Fig. 2(b) and 2(c). By incorporating dynamic category tokens into prompt, the context vectors can better capture transferable affordance knowledge and effectively narrow the search space, thereby enhancing part-level localization accuracy and generalization to unseen categories.

Furthermore, to address the challenges of inaccurate localization in complex scenarios, we propose a coarse-to-fine semantic-guided Transformer decoder. This decoder uses multi-level visual features to progressively guide the textual features, thereby enabling the decoder to learn robust semantic mappings between functional words and object areas.

Overall, our contributions can be summarized as follows:

1) Firstly, we propose a category-conditioned affordance prompt learning approach. By encoding category information from instance-level visual features, our method constructs a complete semantic category-affordance prompt, which effectively narrows the search space for visual localization.
2) Secondly, we propose a coarse-to-fine semantic-guided Transformer decoder. This design enhances the decoder's ability to understand the semantic mapping between the affordance words and corresponding object part-level regions, which improve the accuracy of affordance localization for objects with complex structures.
3) Finally, we conduct extensive experiments on two standard affordance grounding datasets. Experimental results demonstrate that our method achieves competitive performance against related approaches, with less than 1% of the training data.

II. Related Work

Affordance grounding. Affordance grounding aims to combine visual perception and decision-making tasks by retrieving "potentially manipulable" regions in the provided image. Previous research [14], [15] has focused on using convolutional neural networks. Subsequently, many studies [16], [17] have explored weakly supervised methods to mitigate the high cost associated with manually obtaining affordance annotations. Recently, researchers have further extended affordance grounding research to include egocentric videos [18], 3D models [19], hand pose generation ([20], [21]), and associating it with human parts [22].

With the rapid development of large-scale pre-training, vision-language models have been widely applied to open-world detection ([23], [24]) and segmentation ([25], [26]). Many studies have explored the applications of large language models and visual foundation models in affordance learning and reasoning. Locate [27] utilizes DINO [12] features to transfer affordance knowledge from human-object interaction images in a weakly supervised manner. Voxposer [28] introduces an LLM for affordance reasoning to interact with a VLM and generate 3D affordance maps for robot manipulation. AffCorrs [29] leverages the vision foundation model DINO to simplify the task by explicitly selecting relevant objects as support images. AffordanceLLM [30] uses a pre-trained large-scale vision-language model for human-robot interaction knowledge, significantly improving zero-shot generalization to unseen objects and actions.

Unlike most existing approaches that depend on large-scale datasets or domain-specific transfer learning, our method finetunes the vision foundation model to achieve a balance between training efficiency and generalization performance.

Prompt learning. To overcome the limited generalization and lack of task-specific knowledge in handcrafted prompts, prompt learning has been designed as a paradigm for adapting large-scale vision-language models.

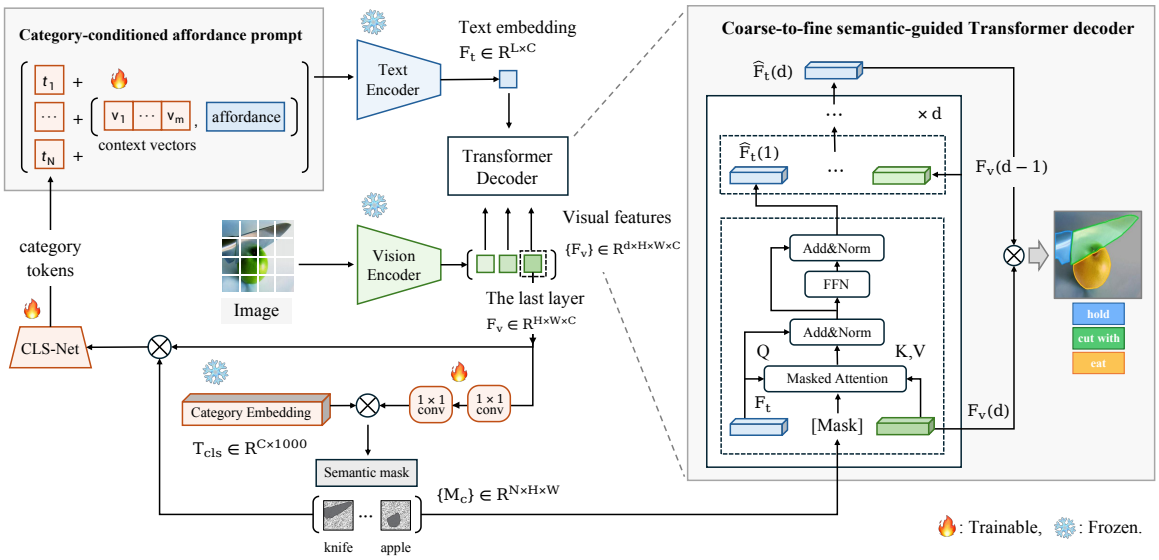


Fig. 3: **Architecture of the proposed Visual Category-Guided framework.** Our method consists of the category-conditioned affordance prompt and a coarse-to-fine semantic-guided transformer decoder that integrates multi-scale intermediate features to enhance localization robustness in complex scenarios.

CoOp [31] introduces a foundational scheme that replaces hand-crafted templates with learnable context vectors, enabling efficient adaptation of vision-language models while keeping the backbone frozen. Building upon this, MaPLe [32] implements multi-modal prompt learning to ensure deep alignment between vision and language branches across multiple transformer layers. More recently, OVAL-Prompt [33] extends affordance grounding to open-vocabulary settings by leveraging LLMs to reason over part-level maps, achieving zero-shot localization on RGB-D images without additional fine-tuning.

However, these methods are less effective for complex, part-level localization tasks such as affordance grounding, particularly when generalizing to unseen categories.

III. Method

A. Overview

Let I denote the input image, T denote the affordance text, and S denote the dense annotations for the corresponding image-affordance pair. The goal of One-Shot Open Affordance Grounding (OOAG) is to align the visual embedding space of object part-level areas with corresponding affordance text. Given only one annotated sample per category, this alignment is achieved by minimizing the localization loss L_{loc} between the predicted results and dense annotations. This process can be expressed as:

$$\theta^* = \arg\min_{\theta} L_{loc}(f_\theta(I,T),S) \tag{1}$$

where f_θ represents the OOAG model. The predicted result $f_\theta(I,T)$ is a dense heatmap where the score at each spatial location is computed via the cosine similarity between the corresponding visual feature and the affordance text embedding. The localization loss L_{loc} is then instantiated based on the specific annotations of each dataset: KL Divergence is employed for datasets with normalized heatmap annotations, while the hybrid BCE-Dice loss is utilized for binary mask annotations.

The architecture of our proposed framework, which primarily consists of a category-conditioned affordance prompt and a coarse-to-fine semantic-guided transformer decoder, is illustrated in Fig. 3. Specifically, our approach encodes instance-level visual features through semantic masks and a proposed lightweight network. To strengthen the alignment between affordance words and fine-grained visual areas, the decoder progressively integrates multi-scale intermediate feature to update the textual features. Furthermore, our method introduces a small number of trainable parameters, thereby ensuring flexible training even under the one-shot learning setting.

B. Category-conditioned affordance prompt

Following the linguistic patterns of affordance descriptors, we propose category-conditioned affordance prompt learning. Our approach employs semantic masks to separate instance-level visual features, which are then fed into a lightweight network to be ultimately encoded as category tokens.

For the visual backbone, we adopt DINOv2 to extract spatial feature maps, denoted as F_v. To effectively process these features, we adapt the AttentionPool module from CLIP-ViT, which was originally designed to generate image embeddings via a multi-head attention mechanism, to encode our visual feature maps. Specifically, we extract the weights of its value projector W_v and final linear output projection W_F, reshaping them into two 1×1 convolutional layers, which allows them to be applied to every pixel of feature map F_v while preserving its dimensions. Considering the distribution

gap between DINOv2 and CLIP's representations, We set these convolutional layers as trainable.

To efficiently derive spatial masks from common semantic concepts, we leverage the 1,000 ImageNet categories, providing a comprehensive vocabulary to compute a static category embedding $T_{cls} \in \mathbb{R}^{c\times 1000}$ via the CLIP text encoder. Then a per-category score tensor $S_c \in \mathbb{R}^{H\times W\times 1000}$ is obtained by computing the dot product between these embeddings and the processed feature map:

$$S_c = W_F(W_v(F_v)) \cdot T_{cls}. \tag{2}$$

From S_c, we subsequently extract N semantic masks. Specifically, a category index map I is generated by taking the channel-wise argmax at each spatial location: $I(u,v) = \text{argmax}_c S_c(u,v,c)$. The resulting binary masks $\{M_c\}_{c=1}^N$ are then formed for all appearing categories, where $M_c = \mathbb{I}[I(u,v) = c]$ and $\mathbb{I}$ denotes the indicator function. Then we discard any semantic mask whose relative area is below the threshold τ. The number of masks N is the count of categories that satisfy this constraint:

$$\frac{1}{H\times W}\sum_{u=1}^{H}\sum_{v=1}^{W} M_c \geq \tau. \tag{3}$$

To extract instance-level visual features, we perform weighted average pooling over these semantic masks. The pooled vector is then fed into a lightweight CLS-Net (h_θ) to further extract category semantic features, which are subsequently encoded as the category token t_c. To maintain efficiency, the CLS-Net adopts a two-layer bottleneck architecture to avoid excessive parameters, which is crucial for preventing overfitting in the one-shot training setting. The computation process for generating category tokens $\{t_c\}_{c=1}^N$ can be expressed as:

$$t_c = \text{h}_\theta\left(\frac{\sum_{u,v} M_c \cdot F_v}{\sum_{u,v} M_c}\right). \tag{4}$$

Finally, we construct the category-affordance prompt using a set of multi-instance category token t_c. First, the token t_c is added element-wise to a set of M learnable context vectors $V = \{v_1, \cdots, v_M\}$, and then concatenate the result with an affordance word T to construct the category-affordance prompt, which is subsequently used as the text input, expressed as $\{\hat{T}\}_{c=1}^N = \{t_c + v_1, \cdots, t_c + v_M, T\}_{c=1}^N$.

During the phase of finetune training, we jointly optimize the context vectors v_c, the parameter h_θ of CLS-Net, and the two 1×1 convolutional layers (W_F,W_v). By incorporating dynamic category tokens during inference, our prompts facilitate the learning of category-agnostic affordance knowledge, thereby improving generalization performance on unseen categories.

C. Coarse-to-fine semantic-guided Transformer decoder

To address the challenge of inaccurate localization in complex scenarios, we propose a coarse-to-fine semantic-guided Transformer decoder. This module feeds intermediate visual features into cascaded decoder blocks to progressively refine the image perception of textual features. Furthermore, the attention mechanism from coarse to fine guides the decoder to understand the semantic mapping between affordance words and object parts.

Specifically, our decoder receives three inputs: the affordance text embedding F_t, a set of multi-level visual feature maps $\{F_v(i)\}_{i=1}^d$, and a set of semantic masks $\{M_c\}_{c=1}^N$. Our decoder receives d visual features from different layers and the Transformer decoder consists of d Transformer blocks, which references visual features to guide the update of the affordance text embedding. In each Transformer block, masked cross-attention is performed by treating text embeddings as queries and visual tokens as keys and values, while utilizing semantic masks to constrain attention to relevant areas. First, we apply linear projections to produce the query (Q), the key (K), and the value (V):

$$Q = \phi_q(F_t), K = \phi_k(F_v), V = \phi_v(F_v). \tag{5}$$

It's worth noting that the masked attention mechanism is applied individually to each category mask M_c. The masked cross-attention allows the model to focus on the update of text embeddings by retrieving relevant visual information corresponding to the affordance text, which is computed as:

$$\hat{F_t} = \text{softmax}\left(\frac{QK^T + log(M_c)}{\sqrt{d_k}}\right)\cdot V + F_t. \tag{6}$$

F_t is passed through a feed-forward network with residual connections to update the text embedding. Then the updated embeddings $\hat{F_t}$ are fed into the next Transformer block. The visual feature from the shallow level containing local detail information continues to guide the text features, and this process progressively enhances part-level perception capabilities for objects with complex structures. For each of the N category masks, we compute a corresponding heatmap via temperature-scaled cosine similarity between its text embedding $\hat{F_t}$ and the visual features F_v. The final output heatmap $H(u,v)$ is then generated by combining these N individual heatmaps through a pixel-wise maximum operation.

$$H(u,v) = \max_{c\in\{1,\ldots,N\}} \frac{\hat{F_t}\cdot F_v}{||\hat{F_t}||\cdot||F_v||}. \tag{7}$$

IV. Experiment

A. Experimental setup

Dataset. We selected two benchmark datasets, AGD20K [16] and the UMD Part Affordance dataset [34]. **AGD20K** is a large-scale affordance grounding dataset with 36 affordance types and 50 object categories, containing 23,816 images. As a weakly supervised dataset, the training images are provided with only image-level labels. In addition, AGD20K uses sparse keypoint annotations and applies Gaussian kernels over each point to generate dense annotations of the affordance areas. The **UMD** Part Affordance dataset consists of 28,843 RGB-D images covering 7 affordance types across 105 kitchen, workshop, and gardening tools. For each image, the dataset provides pixel-level annotations of the affordance regions.

Method	Venue	Training Data	Seen			Unseen (Easy Split)			Unseen (Hard Split)		
			KLD ↓	SIM ↑	NSS ↑	KLD ↓	SIM ↑	NSS ↑	KLD ↓	SIM ↑	NSS ↑
Cross-View-AG	CVPR'22	seen/unseen split: 26,990/	1.538	0.334	0.927	1.787	0.285	0.829	2.092	0.209	0.138
LOCATE	CVPR'23	18,138 image-level labels	1.226	0.401	1.177	1.405	0.372	1.157	1.829	0.282	0.276
AffordanceLLM	CVPR'24	unseen split:	-	-	-	1.463	0.377	1.070	1.661	0.361	0.947
AffordanceSAM	arXiv'25	1,675 dense labels	-	-	-	1.083	**0.543**	**1.800**	1.128	0.514	**1.761**
OOAL	CVPR'24	seen/unseen split: 50/33	0.740	0.577	1.745	1.070	0.461	1.503	1.302	0.410	1.119
Ours	-	keypoint labels	**0.636**	**0.672**	**1.823**	**0.924**	0.525	1.697	**0.972**	**0.533**	1.415

TABLE I: Quantitative comparison on AGD20K on seen and unseen splits.

Method	Setting	Seen (%)	Unseen (%)	mIoU (%)
SegFormer	Fully supervised	74.6	57.7	65.0
PSPNet		72.0	60.8	66.0
OOAL	One-shot learning	74.6	59.7	66.4
Ours		**78.2**	**62.5**	**69.5**

TABLE II: Quantitative comparison on UMD dataset. We evaluate the performance on both seen and unseen splits.

Metrics. We evaluate AGD20K results using the Kullback-Leibler Divergence (KLD), Similarity (SIM), and Normalized Scanpath Saliency (NSS) metrics. For KLD, a lower value is better; for SIM and NSS, a higher value is better. For the UMD dataset, we employ the mean Intersection over Union (mIoU) metric under both seen and unseen settings.

Baselines. Our one-shot open affordance grounding method is directly comparable to the OOAL [13] method. The weakly supervised affordance grounding (WSAG) methods we compared include Cross-View-AG [16], LOCATE [27], AffordanceLLM [30], and AffordanceSAM [35]. Furthermore, we compared our approach on the UMD dataset against representative fully-supervised semantic segmentation methods: PSPNet [36] and SegForm [37].

B. Implementation details

Our experiments are implemented on two NVIDIA GeForce RTX 3090 GPUs, using the base-sized Vision Transformer (ViT-B) for all visual foundation models. Considering the effective proportion of semantic regions, we set the area threshold hyperparameter τ to 0.25. Additionally, the number of decoder blocks d is set to 3, the selection of which will be further discussed in the ablation study. We train the model for 25,000 iterations using the SGD optimizer with a learning rate of 0.015. Input images are initially resized to 256 $\times$ 256 pixels and subsequently randomly cropped to 224 $\times$ 224 with horizontal flipping.

Following a one-shot training protocol, we took one example from each object category to form the training set. Both datasets provide two train-test splits for seen and unseen settings, and we used these splits to evaluate performance. To ensure the reliability of our one-shot evaluation, we report the mean and standard deviation over five independent runs with different random seeds.

C. Comparison to state-of-the-art methods

For the AGD20K dataset, we conduct training under the "seen" setting (trained on all categories) and the "unseen"

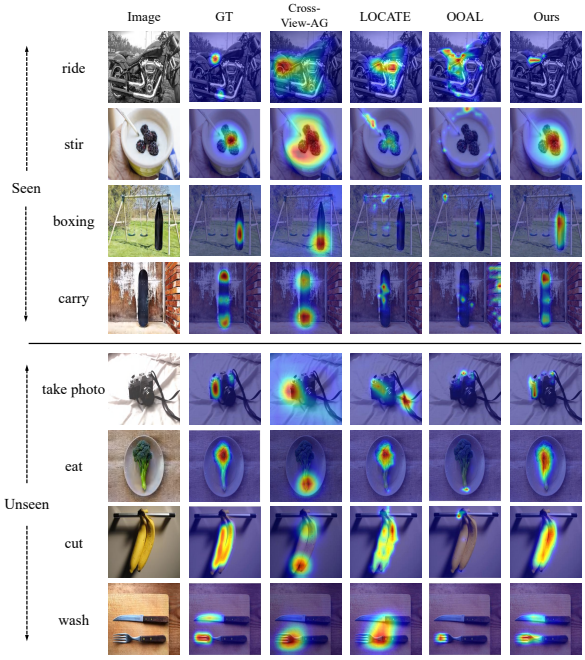


Fig. 4: **Qualitative comparison on AGD20K.**

setting (where certain categories are excluded from training). To more comprehensively evaluate the generalization capability, we benchmark our method not only on the original unseen split (easy split) but also on the more challenging hard split proposed by AffordanceLLM [30], where the test set is markedly less similar to the training data. While weakly supervised methods like Cross-View-AG [16] and LOCATE [27] utilize image-level labels, they still require a massive 26,990 and 18,138 training samples, an amount over 540$\times$ greater than our one-shot learning method. Furthermore, both AffordanceLLM [30] and AffordanceSAM [35] rely on expensive dense annotations and an extra pseudo-labeling phase, implying that their actual training costs significantly exceed the 1,675 images mentioned in their studies. The most similar method to our training setting is the OOAL [13] baseline. We randomly selected and manually annotated one image per category from the 50 categories in the AGD20K dataset, totaling 50 training images. Notably, to utilize the dense annotations from AGD20K, AffordanceLLM and AffordanceSAM only retained on the unseen splits.

As shown in Table 1, our method achieves competitive performance across multiple metrics on the AGD20K

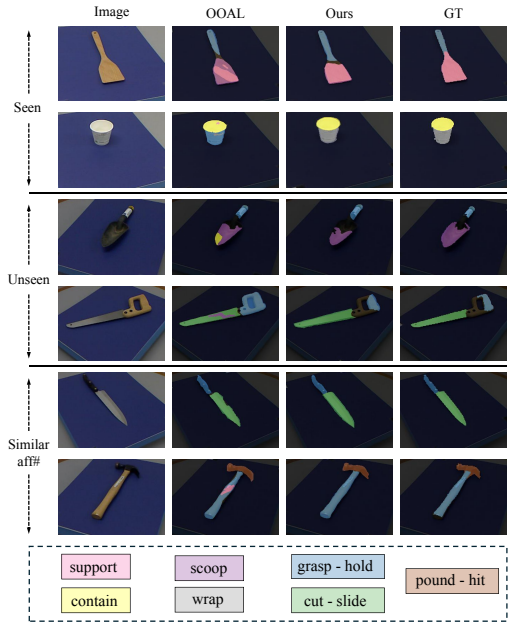


Fig. 5: **Qualitative comparison on UMD.**

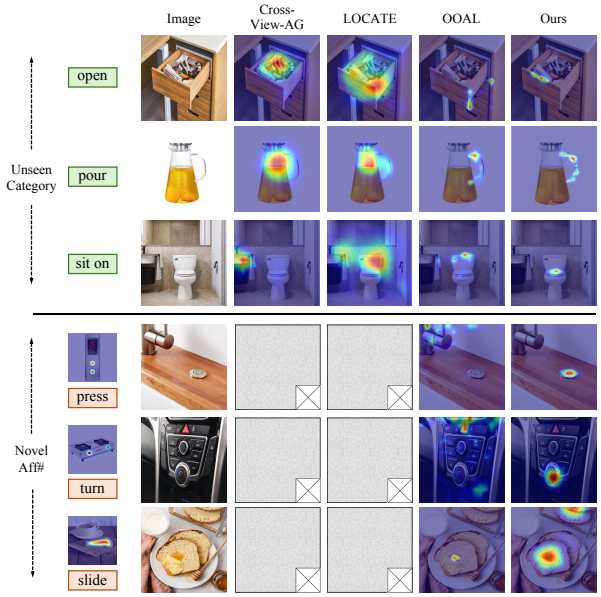


Fig. 6: **Generalization on Internet images.**

dataset. While AffordanceSAM achieves higher performance on some metrics, this can be attributed to its pre-training on a large dataset and pseudo-label generation, which are undeniably effective. However, our method achieves equally competitive performance with only 1% of the training cost.

The comparison on the UMD dataset is shown in Table 2, where we benchmark our approach against several representative fully supervised methods and the baseline method. Experimental results demonstrate that our approach significantly outperforms the compared methods across multiple metrics and achieves state-of-the-art performance.

D. Qualitative results

The qualitative comparison on the AGD20K dataset is illustrated in Fig. 4. We observe that the OOAL method exhibits unsatisfactory accuracy in the unseen setting. This is primarily because its context vectors overfit to seen categories while neglecting the learning of generalizable affordance knowledge. Specifically, the method fails to accurately localize functional regions even for basic examples like broccoli and bananas.

Although LOCATE and Cross-View-AG can make reasonable predictions to some extent, they are biased toward salient regions (e.g., localizing to the entire body of a motorcycle). Furthermore, in multi-instance scenarios, these methods often produce inaccurate spatial localizations. In contrast, our method can clearly localize complex affordance regions. Even under unseen settings, our model effectively handles complex affordances involving multiple affordance regions (e.g., the left and right edges of a camera for taking photos) and localizes the washable regions, including multiple instances (e.g., the blade of a knife and the tines of a fork can be washed).

Fig. 5 presents the visualization results for the UMD dataset. Beyond the seen and unseen settings, we also replaced the affordance labels with semantically similar words. Across all these settings, our method consistently demonstrates robust affordance semantic understanding.

E. Generalization on Internet Images

To further demonstrate the generalization capability in real-world scenarios, we conducted a qualitative comparison on internet images. As shown in Fig. 6, the top three rows illustrate our method's zero-shot generalization to unseen categories. The bottom three rows, labeled as "novel aff#", illustrate generalize to unlearned affordances finetuned in a one-shot manner.

For the "unseen categories", the compared methods show poor affordance localization ability. For the more challenging "novel aff#", the weakly supervised methods Cross-View-AG and LOCATE fail to provide localization capabilities for undefined affordance terms, OOAL also shows limited generalization. In contrast, our method maintains robust localization even when encountering novel affordance terms. These results demonstrate that our method achieves superior robustness in real-world scenarios by rapidly finetuning the vision foundation model.

F. Ablation study

We conduct ablation studies on the category-conditioned prompt (Prompt) and the coarse-to-fine semantic-guided decoder (Decoder) using the unseen setting (hard split) of the AGD20K dataset. The results are summarized in Table 3.

- **Category-conditioned prompts.** As shown in Table 3, the integration of category information substantially improves part-level localization accuracy. Specifically, semantic masks separate multiple instance-level visual features (Line 1 vs. 2), and the lightweight CLS-Net further encodes visual features (Line 1 vs. 3, Line 2 vs. 4), both of which consistently enhance performance. These results indicate that our category-conditioned

#	Prompt		Decoder		Metrics		
	Mask	CLS-Net	Mask	MLF	KLD ↓	SIM ↑	NSS ↑
1					1.302	0.410	1.119
2	✓				1.235	0.470	1.178
3		✓			1.220	0.442	1.140
4	✓	✓			1.187	0.473	1.244
5			✓	✓	1.218	0.459	1.291
6	✓	✓	✓		1.060	0.506	1.338
7	✓	✓		✓	1.025	0.520	1.390
8	✓	✓	✓	✓	**0.972**	**0.533**	**1.415**

TABLE III: Ablation study of the Prompt and Decoder components.

Block num.	KLD ↓	SIM ↑	NSS ↑
2	1.02	0.515	1.372
3	**0.972**	**0.533**	**1.415**
4	1.274	0.303	1.195
6	4.450	0.146	0.197

TABLE IV: Effect of decoder block number on localization performance.

prompt effectively constrains the localization search space, enabling context vectors to capture more generalizable affordance knowledge.

- **Coarse-to-fine decoder.** We incorporate multi-level features (MLF) from the visual encoder into cascaded Transformer decoder blocks to update text embeddings (Line 4 vs. 6) and employ spatial semantic masks (Mask) to guide the cross-attention mechanism (Line 4 vs. 7). The performance gains demonstrate that this novel block facilitates a deeper understanding of the mapping between affordance words and visual regions.
- **Number of decoder blocks.** We also conducted an ablation study on the number of decoder blocks. Table 4 presents the ablation results for the number of decoder blocks. The performance is highest at $d = 3$, but drops sharply at $d = 4$ or 6. We attribute this to the fact that stacking too many blocks introduces excessive parameters, causing the model to overfit during one-shot training.

V. CONCLUSIONS

In this work, we propose a visual category-guided one-shot open affordance grounding method. To improve generalization to unseen categories, we introduce category-conditioned affordance prompt learning, which constructs a complete semantic category-affordance prompt from instance-level visual features. Additionally, we designed a coarse-to-fine semantic-guided Transformer decoder to further improve accurate affordance localization for objects with complex structures, by strengthening the decoder's understanding of the semantic mapping between affordance and object part- level regions. Experiment results on two affordance grounding datasets demonstrate that we achieve competitive performance compared to related methods with less than 1% of the full training data. Furthermore, our approach shows more robust generalization to unseen objects and novel affordances.

REFERENCES

[1] H. Chen, B. Sun, A. Zhang, M. Pollefeys, and S. Leutenegger, "Vidbot: Learning generalizable 3d actions from in-the-wild 2d human videos for zero-shot robotic manipulation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 27 661–27 672.

[2] R. Xu, F.-J. Chu, C. Tang, W. Liu, and P. A. Vela, "An affordance keypoint detection network for robot manipulation," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2870–2877, 2021.

[3] S. Nasiriany, S. Kirmani, T. Ding, L. Smith, Y. Zhu, D. Driess, D. Sadigh, and T. Xiao, "Rt-affordance: Affordances are versatile intermediate representations for robot manipulation," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 8249–8257.

[4] S. Qian and D. F. Fouhey, "Understanding 3d object interaction from a single image," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 753–21 763.

[5] W. Xu, V. Ila, L. Zhou, and C. T. Jin, "Tb-hsu: Hierarchical 3d scene understanding with contextual affordances," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 2025, pp. 8960–8968.

[6] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, "Hico: A benchmark for recognizing human-object interactions in images," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1017–1025.

[7] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8359–8367.

[8] D. Shan, J. Geng, M. Shu, and D. F. Fouhey, "Understanding human hands in contact at internet scale," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9869–9878.

[9] A. Kumar, S. Gupta, D. Fouhey, S. Levine, and J. Malik, "Visual memory for robust path following," *Advances in neural information processing systems*, vol. 31, 2018.

[10] A. Halilovic and S. Krivic, "Affordance-based explanations of robot navigation," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 13 523–13 529.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[12] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

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

[14] A. Myers, A. Kanazawa, C. Fermuller, and Y. Aloimonos, "Affordance of object parts from geometric features," in *International Conference on Robotics and Automation2*, 2015, pp. 5–6.

[15] T.-T. Do, A. Nguyen, and I. Reid, "Affordancenet: An end-to-end deep learning approach for object affordance detection," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 5882–5889.

[16] H. Luo, W. Zhai, J. Zhang, Y. Cao, and D. Tao, "Learning affordance grounding from exocentric images," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 2252–2261.

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

[18] L. Mur-Labadia, J. J. Guerrero, and R. Martinez-Cantin, "Multi-label affordance mapping from egocentric vision," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 5238–5249.

[19] Y. Yang, W. Zhai, H. Luo, Y. Cao, J. Luo, and Z.-J. Zha, "Grounding 3d object affordance from 2d interactions in images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 10 905–10 915.

[20] J. Jian, X. Liu, M. Li, R. Hu, and J. Liu, "Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand

pose," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 14 713–14 724.

[21] Y. Ye, X. Li, A. Gupta, S. De Mello, S. Birchfield, J. Song, S. Tulsiani, and S. Liu, "Affordance diffusion: Synthesizing hand-object interactions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 479–22 489.

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

[23] T. Cheng, L. Song, Y. Ge, W. Liu, X. Wang, and Y. Shan, "Yolo-world: Real-time open-vocabulary object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 16 901–16 911.

[24] S. Fu, Q. Yang, Q. Mo, J. Yan, X. Wei, J. Meng, X. Xie, and W.-S. Zheng, "Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 14 987–14 997.

[25] H. Zhang, F. Li, X. Zou, S. Liu, C. Li, J. Yang, and L. Zhang, "A simple framework for open-vocabulary segmentation and detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1020–1031.

[26] Y. Liu, S. Bai, G. Li, Y. Wang, and Y. Tang, "Open-vocabulary segmentation with semantic-assisted calibration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3491–3500.

[27] G. Li, V. Jampani, D. Sun, and L. Sevilla-Lara, "Locate: Localize and transfer object parts for weakly supervised affordance grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 922–10 931.

[28] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei, "Voxposer: Composable 3d value maps for robotic manipulation with language models," *arXiv preprint arXiv:2307.05973*, 2023.

[29] D. Hadjivelichkov, S. Zwane, L. Agapito, M. P. Deisenroth, and D. Kanoulas, "One-shot transfer of affordance regions? affcorrs!" in *Conference on Robot Learning*. PMLR, 2023, pp. 550–560.

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

[31] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[32] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "Maple: Multi-modal prompt learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 19 113–19 122.

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

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

[35] D. Jiang, M. Wang, T. Ma, H. Li, G. Dai, L. Zhang, *et al.*, "Affordancesam: Segment anything once more in affordance grounding," *arXiv preprint arXiv:2504.15650*, 2025.

[36] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[37] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.