INTRA: Interaction Relationship-aware Weakly Supervised Affordance Grounding

Ji Ha Jang[1*], Hoigi Seo[1*], and Se Young Chun[1,2†]

[1]Dept. of Electrical and Computer Engineering, [2]INMC & IPAI
Seoul National University, Republic of Korea
{jeeit17, seohoiki3215, sychun}@snu.ac.kr

Abstract. Affordance denotes the potential interactions inherent in objects. The perception of affordance can enable intelligent agents to navigate and interact with new environments efficiently. Weakly supervised affordance grounding teaches agents the concept of affordance without costly pixel-level annotations, but with exocentric images. Although recent advances in weakly supervised affordance grounding yielded promising results, there remain challenges including the requirement for paired exocentric and egocentric image dataset, and the complexity in grounding diverse affordances for a single object. To address them, we propose INTeraction Relationship-aware weakly supervised Affordance grounding (INTRA). Unlike prior arts, INTRA recasts this problem as representation learning to identify unique features of interactions through contrastive learning with exocentric images only, eliminating the need for paired datasets. Moreover, we leverage vision-language model embeddings for performing affordance grounding flexibly with any text, designing text-conditioned affordance map generation to reflect interaction relationship for contrastive learning and enhancing robustness with our text synonym augmentation. Our method outperformed prior arts on diverse datasets such as AGD20K, IIT-AFF, CAD and UMD. Additionally, experimental results demonstrate that our method has remarkable domain scalability for synthesized images / illustrations and is capable of performing affordance grounding for novel interactions and objects. Project page: https://jeeit17.github.io/INTRA

Keywords: Affordance grounding · Weak supervision · Exocentric image · Contrastive learning · Interaction relation

1 Introduction

Affordance [17] refers to the perceived possible interactions based on an object's inherent or recognized properties (*e.g.*, the rim of a wine glass affords sipping while stem of it affords holding). Humans can identify affordances of objects and interact with proper parts despite the diversity in their physical attributes.

* Authors contributed equally. † Corresponding author.

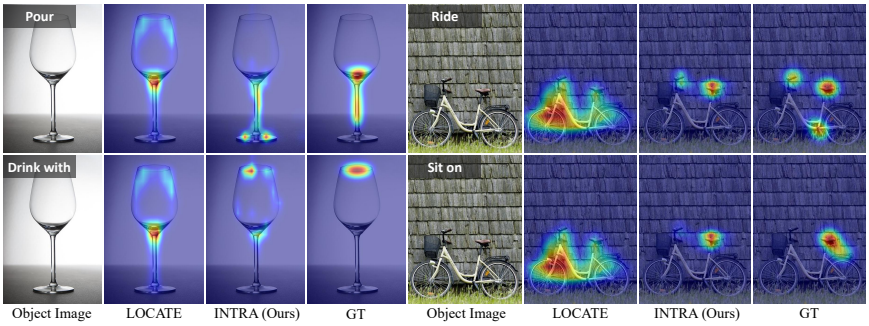


Fig. 1: Prior works on weakly-supervised affordance grounding like LOCATE [25] often failed to ground different affordances for the same object. However, our proposed INTRA yielded finer and more accurate grounding results for them that are closer to the ground truth (GT) by considering interaction relationship among them.

This ability can be acquired through individual learning, by directly interacting with objects, and observational learning [8], by observing others' interactions. The sense of affordance enables effective interaction in new environments or with novel objects, without step-by-step instructions [57]. Affordance plays an essential role across numerous applications involving intelligent agents, enabling them to provide flexible and timely responses in complex, dynamic environments [5]. These applications include task planning, robot grasping, manipulation, scene understanding and action prediction [2, 6, 7, 16, 51, 62].

Affordance grounding is the task to teach intelligent systems how to locate possible action regions in objects for a certain interaction. While fully supervised learning [4, 18, 42, 58] is the most straightforward approach, its reliance on costly annotations may limit its applicability across diverse contexts. Another approach is weakly supervised learning, similar to human's observational learning [8], that does not require GT, but *weak* labels. In this setting, *exocentric* images illustrating human-object interactions, along with corresponding *egocentric* images depicting the objects, are provided during training. During inference, intelligent systems perform affordance grounding on the egocentric images, identifying object parts relevant to the given interactions. Recent advances in weakly supervised affordance grounding [25, 31, 32, 41] proposed to use *pairs* of exocentric and egocentric images, yielding great performance. The deep neural networks learn affordances by pulling features from exocentric and egocentric images closer, aiming to focus on object parts related to interactions.

However, weakly supervised affordance grounding remains challenging. Firstly, the requirement for current weak labels with *pairs* of exocentric and egocentric images is still strong. Note that human observational learning does not usually require egocentric images. Secondly, a complex relationship between interactions exists, which has not been adequately addressed in prior works. Many instances in object-interaction relationships exhibit intricate many-to-many associations, occasionally with one entailing another. For example, some distinct

interactions represent the same affordance regions (*e.g.*, 'wash' and 'brush with' a tooth brush), and there are closely related interactions that always come together (*e.g.*, 'sip' usually includes 'hold'. 'ride' usually includes 'sit on'). This complexity poses challenges in extracting interaction-relevant features based on image-level affordance labels, introducing biases towards objects in affordance grounding as illustrated in Fig. 1 (LOCATE [25] often yielded similar affordance grounding with different interactions for the same object).

Here, we propose a novel weakly supervised affordance grounding method, INTRA (INTeraction Relationship-aware weakly supervised Affordance grounding) to address these unexplored challenges. While previous studies [31,41] solved the weak supervision problem as supervised learning by pulling object features of exocentric and egocentric images closer and LOCATE [25] enhanced this approach by generating more localized pseudo labels based on prior information for exocentric images for supervised learning (*i.e.*, containing human, object part, and background), our INTRA framework recasts the weak supervision problem as representation learning. This novel reformulation allows us to use *weaker* labels (*i.e.*, exocentric images only) for training so that the requirement to use *pairs* of exocentric / egocentric images is now alleviated. Moreover, unlike prior works, our INTRA method actively exploits large language model (LLM) as well as the text encoder of the vision-language model (VLM) to leverage linguistic information and existing textual knowledge on affordances, which further enhances our interaction relationship-guided contrastive learning. This novel scheme also allows excellent scalability for unseen objects across diverse domains as well as zero-shot inference for novel interactions, which was not possible in prior arts. In summary, our main contributions are three-fold as follows:

- We propose a novel approach for weakly supervised affordance grounding by recasting the problem as representation learning and by leveraging VLM, leading to relaxing the need for paired training datasets for *more* weak supervision and enhancing scalability across domains for unseen objects.
- We proposed INTRA, a novel method that consists of text synonym augmentation and text-conditioned affordance map generation module along with interaction relationship-guided contrastive learning, so that inference on unseen interactions is possible.
- Our INTRA outperforms the prior arts in weakly supervised affordance grounding on diverse datasets such as AGD20K, IIT-AFF, CAD and UMD, demonstrating both qualitative and quantitative excellence (see Fig. 1).

2 Related Works

2.1 Affordance Grounding

Supervised affordance grounding. Supervised affordance grounding methods [10, 14] analyze interaction videos / images to predict affordance regions on an object, trained with pixel-level GT masks / heat maps. Though successful in localizing fine-grained affordance regions through supervised learning, they are

limited by the costly GT mask annotation process and their limited generalizability to unseen objects. Furthermore, they require paired demonstration videos and target object images, making real-world application challenging.

Weakly supervised affordance grounding. Weakly supervised affordance grounding methods [13,20,23,25,31–33,41,47] offer the advantage of not requiring GT, but requiring weak labels such as exocentric images with interaction text labels. Prior works [25,31,32,41] mainly align interaction-relevant object features from both egocentric and exocentric images without considering the intrinsic properties of interactions. The framework in [41] predicts object features engaged in interactions by analyzing human-object interaction videos. The works of [31, 32] preserve the correlation of affordance features from exocentric and egocentric images to learn affordances. The work of [25] enhances object feature extraction by adopting DINO-ViT [9] based Class Activation Maps (CAM) [64] and k-means clustering [35] for more explicit guidance. However, focusing solely on object features may introduce biases towards object, hindering the inference of multiple affordances for a single object. Our INTRA addresses this issue by considering the complex relationships between interactions using interaction relationship-guided contrastive loss, while ensuring the network remains attentive to the objects using object-variance mitigation loss.

2.2 Foundation Models for Affordance Grounding

Self-supervised transformer. Self-supervised transformers, extensively trained on large-scale datasets and scalability, possess robust representation power. Previous works [25,52] have explored their potential in affordance grounding. DINO-ViT [9], a vision transformer foundation model trained in a self-supervised manner, can identify both high-semantics such as overall information of the image and low-semantics such as details regarding specific object parts. This versatility has led advancements in various tasks, including classification, semantic segmentation [4, 24] and semantic correspondence [61]. LOCATE [25] leverages DINO-ViT to extract low-semantic information, resulting in performance improvements in affordance grounding. Our INTRA employed DINOv2 [46] as an image encoder to extract information about objects and their constituent parts.

Vision-language model. The Vision-Language Model (VLM) is a class of models jointly pretrained on visual and language data for various downstream tasks [45, 56, 60]. VLM text encoders, trained through contrastive learning with image-text pairs, capture representations in the joint space of the images and text [26–28, 50]. These text encoders, incorporating visual information, have demonstrated excellent performance across multiple tasks. ALBEF [28] notably enhances vision and language representation learning by aligning image and text features before fusing them. While supervised affordance grounding methods leveraging VLM text encoders [44] have been explored, their application in weakly supervised affordance grounding remains underexplored. We propose a

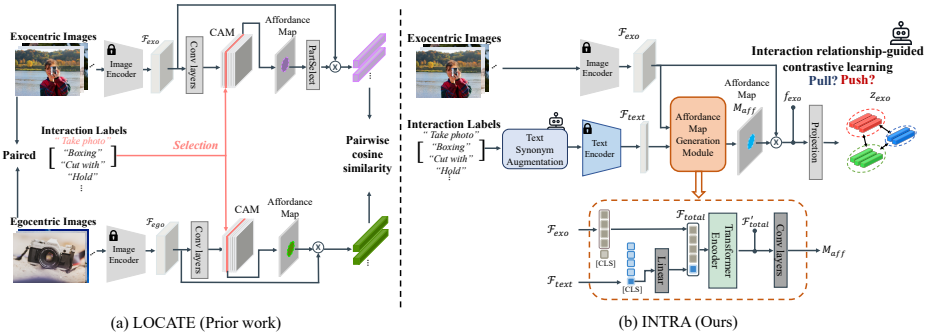


Fig. 2: Overall frameworks of (a) LOCATE [25] and (b) INTRA (Ours). LOCATE takes paired exocentric and egocentric images to generate interaction-aware affordance maps (CAMs) for predefined interactions and then selects an interaction-related CAM by the given interaction label. In contrast, INTRA takes only exocentric images and interaction labels to yield an affordance map through our affordance map generation module. Training is done via interaction relationship-guided contrastive learning on exocentric features from affordance maps. Note that all encoder parameters are frozen.

framework leveraging the text encoder of ALBEF to enable novel interactions, diverging from prior arts limited to inferring predetermined sets of affordances.

Large language model. Understanding affordance relationships is crucial for affordance grounding, as it enables extending and linking learned visual cues, and reasoning about affordances for new objects, interactions, or situations. While prior works like [19] leverage semantically similar object properties and [32] utilize affordance feature correlation, none directly exploit these relationships. We use these intricate relationships in affordance learning by adopting Large Language Models (LLMs). LLMs have gained prominence in robotics due to their profound natural language understanding, providing valuable priors about interactions and their complex relationships. Previous works [3, 29, 54, 63] focus on extracting action knowledge, deriving task-specific plans, and grounding them in the physical world. LLMs have also been widely used in previous affordance studies [37, 55], demonstrating their exceptional understanding of interactions.

3 Method

Prior arts in weakly supervised affordance grounding [25, 31, 32, 41] typically align object features of paired exocentric (interaction with object) and egocentric (object only) images to learn interaction-related features. For example, as illustrated in Fig. 2(a), LOCATE [25] generates CAMs (affordance maps) from exocentric and egocentric images for a pre-determined interaction label, extracts egocentric feature as well as exocentric object parts feature selected by PartSelect module (pseudo label), and then trains the model by optimizing cosine

similarity to align (pull) egocentric and exocentric object parts features. In contrast, we propose an alternative approach, INTRA, whose overall framework is illustrated in Fig. 2(b). Our text-conditioned affordance grounding framework of INTRA leverages VLM text encoder in affordance map generation module and employs text synonym augmentation to enhance robustness, as will be described in Sec. 3.1. Then, INTRA learn affordance grounding via our interaction relationship-guided contrastive learning, detailed in Sec. 3.2. The framework of INTRA as depicted in Fig. 2(b) clearly suggests two advantages over prior arts including LOCATE [25]: 1) it exploits exocentric images only and 2) INTRA admits novel interactions outside the pre-defined interaction set.

3.1 Text-conditioned Affordance Grounding Framework

To utilize the semantic meanings inherent in interaction labels and enable flexible inference on novel verbs, our text-conditioned affordance grounding framework generates affordance maps by conditioning image features with text features via our affordance map generation module where text and image features extracted from separately pre-trained text and image encoders are fused. In specific, as depicted in Fig. 2(b), deep features $\mathcal{F}_{exo} \in \mathbb{R}^{(h \times w) \times d}$ are obtained from the input exocentric images using DINOv2 [46], where h and w represent the height and width of the affordance map, and d refers to the dimension of the feature. The text feature $\mathcal{F}_{text}$ of the given interaction is obtained using the ALBEF text encoder [28]. See the supplementary material for further details on the rationale for employing DINOv2 and the ablation study on the text encoder.

Affordance map generation module. Before fusing text and image features, the class token of $\mathcal{F}_{text}$ passes through a single linear layer to align the separately pre-trained image and text embedding spaces and connect them, as shown to be effective in previous works [30, 65]. Subsequently, image features $\mathcal{F}_{exo}$ and the class token of text features are concatenated and processed through a transformer encoder for conditioning. The image feature part of the resulting vector is then projected using a multi-layered convolutional network and normalized using min-max normalization to obtain the affordance map $\mathcal{M}_{aff} \in \mathbb{R}^{h \times w}$. This affordance map represents the image regions in exocentric images most relevant to interactions. During inference, $\mathcal{M}_{aff}$ functions directly as an output heatmap, indicating the image regions in egocentric images most relevant to interactions.

Text synonym augmentation. To enhance the robustness of text conditioning, we integrate text synonym augmentation into our interaction embeddings. Initially, we generate k_s synonyms for each interaction label using LLM. Subsequently, any synonyms overlapping with other interaction labels are removed. These synonyms are then randomly selected to substitue the text conditioning interaction embedding, while the original interaction label is retained for interaction relationship-guided contrastive learning. This module enhances overall performance by providing models with enriched interpretations of interactions.

3.2 Interaction Relationship-guided Contrastive Learning

Our INTRA learns via interaction relationship-guided contrastive learning by comparing exocentric image features across diverse interactions. Our contrastive learning consists of two key components, 1) extracting exocentric image features with affordance map and 2) designing loss for interaction relationship-guided contrastive learning, that enable the grounding of multiple affordances on a single object.

Exocentric image feature extraction with affordance map. As described in Sec 3.1, a text-conditioned affordance map, $\mathcal{M}_{aff}$, is generated to represent interaction-relevant image regions of exocentric images. Then, the exocentric image features f_{exo} corresponding to the affordance map are extracted as follows:

$$f_{exo} = (1/hw)\Sigma_{i=1}^{h}\Sigma_{j=1}^{w}\mathcal{F}_{exo}(i,j) \cdot \mathcal{M}_{aff}(i,j) \in \mathbb{R}^{d}. \tag{1}$$

The resulting f_{exo} is then projected and normalized to obtain the exocentric image feature z_{exo} using an MLP layer, which will be used for training. This projection layer was also used in previous works [11, 12, 59], which have demonstrated the necessity and efficiency of it.

Loss design for interaction relationship-guided contrastive learning. Supervised contrastive learning [21] effectively derives good representations for each class by focusing on common characteristics in positive pairs while disregarding those in negative pairs like other classes. However, in affordance grounding tasks, treating all other interaction classes as negative pairs may be inadequate due to the complex relationship among interactions. To mitigate this issue, we propose *interaction relationship-guided contrastive loss*, $\mathcal{L}_{inter}$. Furthermore, considering the subtle meaning variations within single interaction classes depending on the object and context, we also propose *object-variance mitigation loss*, $\mathcal{L}_{obj}$. Thus, the total loss for our INTRA is formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{inter} + \lambda_{obj}\mathcal{L}_{obj} \tag{2}$$

where λ_{obj} denotes the control parameter of $\mathcal{L}_{obj}$.

Interaction relationship-guided contrastive loss. In affordance grounding, treating all other interaction classes as negative pairs is inadequate due to the intricate relationships between interactions. For example, 'Wash' and 'Brush with' toothbrush or 'Pour' and 'Seal' bottle represent distinct interactions but act on the same object parts. Manually finding these relationships is time-consuming and impractical as the number of pairs grows quadratically with the number of interaction (see the supplementary). Moreover, although linguistic relationships like synonyms or co-occurrence were considered as substitutes, they are often inadequate and degrade performance. For example, 'Sip' entails 'Hold', but they act on different part of objects, and 'Wash' and 'Cut with' a knife have different

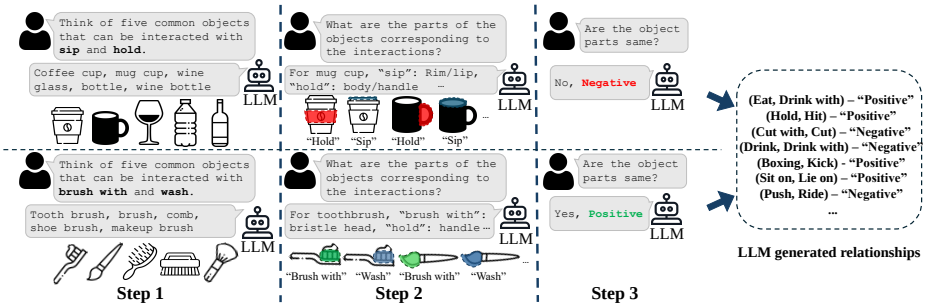


Fig. 3: The overall scheme of interaction-relationship map ($\mathcal{R}$) generation. LLM classifies all pairs of interactions in the dataset as positive or negative through chain of thoughts. This process is based on reasoning if interactions occur on same object parts.

meanings, but they act on the same blade. To mitigate this, we leverage LLM to determine if interaction pairs act on the same object part. Through Chain of Thoughts (CoT), interaction pairs are categorized as positive or negative in three steps as described in Fig. 3. In Step 1, LLM deduces five different objects where both interactions could be performed. In Step 2, LLM identifies object parts where these interactions could occur by considering five objects one by one, not simultaneously. In Step 3, if the identified parts of the interaction pair are the same, the pair is classified as positive; otherwise, negative. Positive pairs are assigned 1 in the interaction-relationship map $\mathcal{R}$, and negative pairs are assigned 0. We propose interaction-relationship guided contrastive loss by integrating $\mathcal{R}$ into supervised contrastive learning as follows:

$$\mathcal{L}_{inter} = \sum_{i=1}^{2N} \frac{-1}{2N_{y_i} - 1} \sum_{j=1}^{2N} \mathcal{R}_{(y_i, y_j)} \cdot \log \frac{\exp\left(z^i_{exo} \cdot z^j_{exo} / \tau\right)}{\sum_{k=1}^{2N} \mathbf{1}_{i \neq k} \cdot \exp\left(z^i_{exo} \cdot z^k_{exo} / \tau\right)} \tag{3}$$

where i, j are sample indices, y_i, y_j are class labels, N_{y_i} is the number of samples in the batch labeled with y_i, N is the total number of distinct samples in the batch, z^j_{exo} is the exocentric image feature vector of each sample, τ is the temperature, and $\mathcal{R}_{(y_i, y_j)}$ is the value of (y_i, y_j) pair in interaction-relationship map.

Object-variance mitigation loss. In the context of affordance, the interpretation of the same interaction can vary significantly based on the object and context. For instance, 'Hold' a baseball bat and a cup may seem similar since both involve grasping an object. However, the former involves gripping the bat's slender part, while the latter entails holding the cup's rounded, protruding part. To address this variance within the same interaction category, we implemented

an object-variance mitigation loss $\mathcal{L}_{obj}$ as follows:

$$\sum_{i=1}^{2N} \frac{-1}{2N_{o_i} - 1} \sum_{j=1}^{2N} \mathbf{1}_{o_i = o_j} \cdot \log \frac{\exp\left(z_{exo}^{i} \cdot z_{exo}^{j} / \tau\right)}{\sum_{k=1}^{2N} \mathbf{1}_{i \neq k} \cdot \exp\left(z_{exo}^{i} \cdot z_{exo}^{k} / \tau\right)} \tag{4}$$

where o_i, o_j denote object class of i and j.

4 Experiments

4.1 Experimental Setting

Dataset and metrics. We conducted an evaluation of our method using the Affordance Grounding Dataset (AGD20K) [32]. AGD20K comprises both exocentric and egocentric images, with 20,061 exocentric images and 3,755 egocentric images labeled with 36 affordances. The dataset support evaluation under two settings: 1) the 'Seen' setting, where the object categories of the training and testing sets are identical, and 2) the 'Unseen' setting, where no objects overlap between the training and test sets. Our approach only used exocentric images in training for all experiments, while other approaches were trained using both egocentric and exocentric images. We employed three evaluation metrics commonly employed in previous affordance grounding methodologies: 1) Kullback-Leibler Divergence (KLD), 2) Similarity (SIM), 3) and Normalized Scanpath Saliency (NSS). These metrics were utilized to quantify the similarity between the distributions of ground truth heatmaps and predicted affordance grounding.

Implementation details. We employed DINOv2 as the image encoder and ALBEF, fine-tuned with RefCOCO+, as the text encoder. ChatGPT-4 [1] served as the LLM. Images were resized to 384×384, then cropped to 336×336. Training utilized the Adam optimizer [22] with a learning rate of 2e-4 and a batch size of 256. The hyperparameter λ_{obj} was set to 4, and all experiments were conducted on a single NVIDIA A100 GPU. More details are provided in the supplementary.

4.2 Comparison to State-of-the-art Methods

To comprehensively assess our method, we conduct quantitative and qualitative comparisons with state-of-the-art weakly-supervised grounding methods, incorporating a user study. We further expand our experiments to include additional datasets [40, 43, 53] for a comprehensive evaluation. Refer to the supplementary materials for more details on the experimental settings.

Quantitative results. We evaluated previous works [15, 25, 31, 32, 36, 41, 48] and our method based on the metrics mentioned above. Tab. 1 shows the quantitative comparison results of our method with prior arts. In both 'Seen' and 'Unseen' setting, our approach surpasses the baseline performances across all three metrics: KLD, SIM, and NSS, thereby setting a new state-of-the-art.

Table 1: Quantitative results of ours and other baselines [15, 25, 31, 32, 36, 41, 48] on the AGD20K dataset. ↑ / ↓ indicates that higher / lower the metric is, the better the model performs. INTRA outperformed all baselines, despite being trained only with exocentric images, whereas other models incorporated both exocentric and egocentric images during training.

Prior works			Seen			Unseen		
			mKLD↓	mSIM↑	mNSS↑	mKLD↓	mSIM↑	mNSS↑
Weakly Supervised Object Localization		EIL [36]	1.931	0.285	0.522	2.167	0.227	0.330
		SPA [48]	5.528	0.221	0.357	7.425	0.167	0.262
		TS-CAM [15]	1.842	0.260	0.336	2.104	0.201	0.151
Weakly Supervised Affordance Grounding	Exo+Ego	Hotspots [41]	1.773	0.278	0.615	1.994	0.237	0.557
		Cross-view-AG [32]	1.538	0.334	0.927	1.787	0.285	0.829
		Cross-view-AG+ [31]	1.489	0.342	0.981	1.765	0.279	0.882
		LOCATE [25]	1.226	0.401	1.177	1.405	0.372	1.157
	Exo	**INTRA (Ours)**	1.199	0.407	1.239	1.365	0.375	1.209

Table 2: Quantitative results on the modified IIT-AFF, CAD, and UMD dataset for our method and other baselines [25,31,32]. Models were trained in the 'Seen' setting of AGD20K and tested on the datasets without additional training. INTRA outperformed all baselines on all metrics across all datasets. * Objects with affordances that prior works are unable to predict were eliminated from the datasets for fairness, wheares our method can infer affordances on novel interactions.

	IIT-AFF* [43]			CAD* [53]			UMD* [40]		
	mKLD↓	mSIM↑	mNSS↑	mKLD↓	mSIM↑	mNSS↑	mKLD↓	mSIM↑	mNSS↑
Cross-View-AG [32]	3.856	0.096	0.849	2.568	0.173	0.589	4.721	0.014	1.287
Cross-View-AG+ [31]	3.920	0.095	1.072	2.529	0.176	0.663	4.753	0.013	1.227
LOCATE [25]	3.315	0.115	1.709	2.528	0.187	0.558	4.083	0.026	2.699
INTRA(Ours)	2.663	0.148	2.511	2.095	0.243	1.259	3.081	0.062	4.195

Results on additional datasets. We evaluated the generalization and robustness of the INTRA framework, along with previous works [25, 31, 32] trained in the 'Seen' setting of AGD20K, on the IIT-AFF [43], CAD [53], and UMD [40] datasets. The experiment was conducted in the 'Seen' setting due to overlapping objects between these datasets and AGD20K. Each GT was processed in the same way as when evaluating the AGD20K test set. Despite significant domain gaps across datasets, INTRA outperformed in all metrics on all datasets, demonstrating its superior generalizability as shown in Tab. 2. Further details of the experiment can be found in the supplementary material.

Qualitative results. Fig. 4 and Fig. 5 show our superior grounding precision compared to the baselines, being closer to the GT and finer in granularity. INTRA precisely identifies the exact object part for a given affordance, unlike the baselines, which ground the same parts regardless of the affordances provided.

User study. Affordance grounding can be ambiguous depending on context and interpretation, thus relying solely on metrics for evaluation has limita-

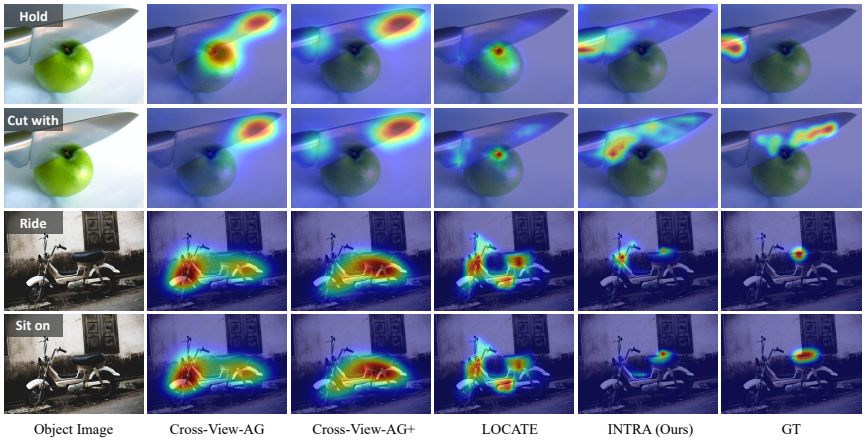


Fig. 4: Qualitative results of INTRA (Ours) and baseline models [25,31,32] on grounding affordances of multiple potential interactions on a single object. INTRA precisely localizes relevant interaction spots for each interaction. For example, with a knife, it grounds the handle for 'Hold' and the blade for 'Cut with'. For a motorcycle, it accurately grounds the saddle for 'Sit on'. Additionally, for 'Ride', it grounds both the handle and saddle, slightly deviating from the GT but still producing reasonable results, as we usually interacts with handle and saddle to 'Ride' a motorcycle.

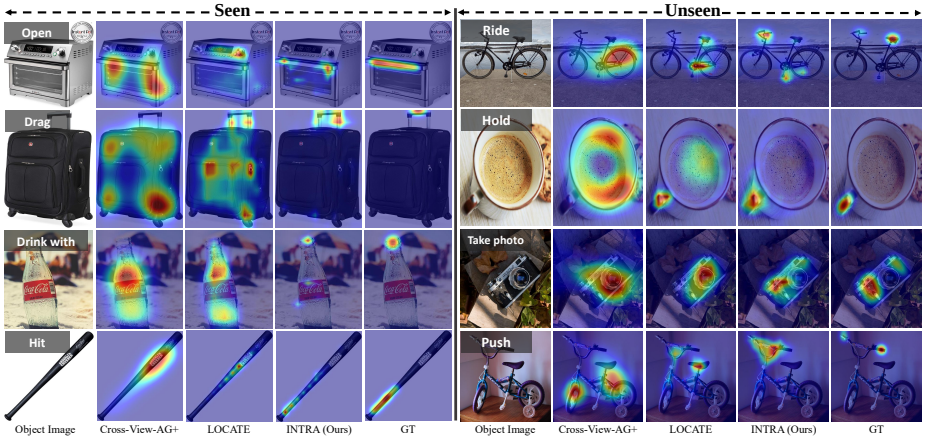


Fig. 5: Qualitative results comparison between our approach and other baselines [25, 31,32]. Our approach, INTRA, demonstrates superior precision and detail in grounding affordances compared to the baselines. For instance, in the example of 'Drag', while baselines either fail to localize the handle or erroneously ground several other parts, INTRA accurately identifies and grounds the handle of a suitcase with finesse.

Table 3: The result of user study on validity, finesse, and separability. Users were asked to score a 5-point scale, and we averaged it for mean opinion score (MOS).

	Validity	Finesse	Separability
Cross-View-AG+ [31]	2.897	3.022	2.732
LOCATE [25]	3.054	2.573	2.651
INTRA (Ours)	3.134	3.112	3.221
Ground Truth	2.905	3.334	3.160

Table 4: Quantitative results of ablation study on our loss design. We incrementally added each component of the losses to examine their impact.

	Seen			Unseen		
	mKLD↓	mSIM↑	mNSS↑	mKLD↓	mSIM↑	mNSS↑
baseline	1.678	0.338	0.891	1.581	0.300	1.100
$\mathcal{L}_{inter}$	1.439	0.334	1.031	1.569	0.292	1.133
$\mathcal{L}_{obj}$	1.336	0.387	1.218	1.521	0.334	1.042
$\mathcal{L}_{inter}$+$\mathcal{L}_{obj}$	1.199	0.407	1.239	1.365	0.375	1.209

Table 5: Quantitative results of ablation study on different $\mathcal{R}$. $\mathcal{L}_{WordNet}$, $\mathcal{L}_{Word2Vec}$ are calculated using word similarity from WordNet [39], Word2Vec [38], respectively. $\mathcal{L}_{Co-occur.}$ used co-occurrence probability in GloVe [49].

	Seen			Unseen		
	mKLD↓	mSIM↑	mNSS↑	mKLD↓	mSIM↑	mNSS↑
$\mathcal{L}_{WordNet}$	1.701	0.282	0.710	1.698	0.277	0.937
$\mathcal{L}_{Co-occur.}$	1.519	0.309	0.988	1.639	0.274	1.101
$\mathcal{L}_{Word2Vec}$	1.547	0.302	0.958	1.679	0.270	0.980
$\mathcal{L}_{inter}$**(Ours)**	**1.439**	**0.334**	**1.031**	**1.569**	**0.292**	**1.133**

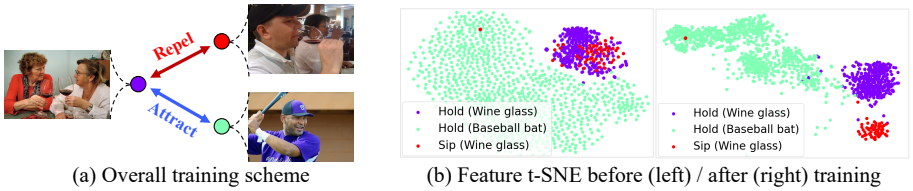
tions. Hence, we conducted a user study comparing Cross-View-AG+ [31], LOCATE [25], GT, and INTRA (Ours) across three categories: 1) Validity: assessing heatmap reasonableness, 2) Finesse: measuring heatmap detail, 3) Separability: determining the accuracy of the heatmap when different affordances are assigned to the same object. A total of 936 responses were collected for randomly selected samples from 104 respondents. Results presented in Tab. 3 demonstrate that our approach outperforms baselines and par on GT based on human perception.

4.3 Ablation Studies

We validate our pipeline design choices and parameters with ablation studies. This section includes ablation studies on loss design, adoption of LLM, and text synonym augmentation. Refer to the supplementary for further ablation studies.

Ablation study on loss design. To assess the individual impact of the components comprising loss on its overall performance, we analyzed by incrementally adding components. We started with the most basic element: a normal supervised contrastive loss. Subsequently, we sequentially added an interaction relationship-guided loss and an object-variance mitigation loss. The performance outcomes of these incremental modifications were thoroughly evaluated to understand their contributions, as represented in the Tab. 4.

Ablation study on adoption of LLM. Adopting LLM to create the relationship map was essential given the intricate nature of affordances. We experimented with various methods to create the relationship map to validate this



(a) Overall training scheme (b) Feature t-SNE before (left) / after (right) training

Fig. 6: An illustration of interaction relationship-guided contrastive learning and t-SNE [34] visualization of feature distribution. (a) In interaction relationship-guided contrastive learning, positive interaction pairs attract each other, while others repel. (b) t-SNE visualization of DINOv2 [46] class token and f_{exo} from INTRA, showing that features of positive interaction pairs become closer as learning progresses.

choice. We measured the similarity of interaction pairs using WordNet [39] and Word2Vec [38], or computed co-occurence probability of interaction pairs with Glove [49]. Based on these measurements, we created an Interaction-relationship Map and trained the INTRA framework. The results are in the Tab. 5.

Ablation study on text synonym augmentation. We conducted an ablation study on the effectiveness of text synonym augmentation on overall performance. We compared performance with and without the module. The module improved performance by up to 21.93%, particularly in the 'Unseen' setting, enriching models with varied meanings of interactions. Additionally, to test its effectiveness on novel verb inference, we deliberately omitted the subset 'Hold' (24.17% of training data) and then performed inference on 'Hold'. The module boosted performance for novel verbs by up to 58.06%. Similar tendencies were observed for other verbs. Detailed results are available in the supplementary.

5 Discussion

5.1 Effect of Interaction Relationship-guided Contrastive Loss

Our rationale for learning affordance grounding solely with exocentric images relies on the consistent presence of humans within these images. By repelling common features of negative pairs, such as human parts, the images effectively exclude irrelevant elements. Conversely, positive pairs, sharing the desired feature of the object—specifically, the rim of the object near the face—facilitate learning by attracting these relevant features (see Fig. 6(a)). To visualize the effectiveness of our loss in learning interaction-relevant features in similar images, we examine the feature distributions of 'Hold' and 'Sip' a wine glass, involving distinct affordances. Prior to training, these distributions overlap. However, after training with our loss function, the feature distribution for 'Hold wine glass' aligns more closely with 'Hold baseball bat' than with 'Sip wine glass'. This indicates that our loss function effectively discriminates between the characteristics of different interactions without exhibiting bias towards objects (see Fig. 6(b)). Detailed explanation is illustrated in the supplementary.

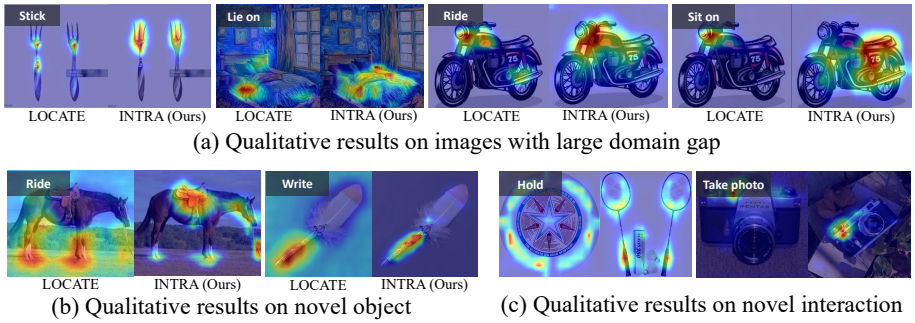


Fig. 7: Qualitative results of feasibility study: (a) Inference on diverse images with significant domain gap such as pixel arts and paintings. (b) Inference on novel objects that were not in the training data. (c) Inference on unseen novel interactions. INTRA demonstrates superior grounding accuracy in (a)-(c) compared to LOCATE [25], showing proper affordance region inference without explicit training.

5.2 Feasibility Study on Generalization Property of INTRA

INTRA excels in affordance grounding on images with large domain gaps, such as pixel art and paintings, as illustrated in Fig. 7(a). Furthermore, our method showcases strong generalization abilities for novel objects like a horse and quill, not present in the training set, as shown in Fig. 7(b). Additionally, despite deliberately not being trained on specific interaction classes like 'Hold' and 'Take photo' for experiment, INTRA successfully infers their affordances, as depicted in Fig. 7(c). More results and detailed experimental settings are in the supplementary. One possible explanation for this generalization property is that our INTRA employs VLM so that diverse domains and novel object can be dealt with without explicitly tuning for them. Another explanation is INTRA's contrastive training that may achieve better representation learning.

6 Conclusion

In this paper, we introduce INTRA, a novel framework reformulating the weakly supervised affordance grounding with representation learning. We suggest interaction relationship-guided contrastive learning, informed by affordance knowledge from LLM. Furthermore, INTRA actively leverages VLM text embedding in proposed text-conditioned affordance map generation for flexible affordance grounding, further bolstered by text synonym augmentation for robustness. INTRA achieves state-of-the-art performance across diverse datasets, relying solely on exocentric images for training, unlike prior methods that also use egocentric images. Moreover, our method demonstrates generalization feasibility on novel objects, interactions, and images with significant domain gaps.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)], the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. NRF-2022M3C1A309202211) and AI-Bio Research Grant through Seoul National University. Also, the authors acknowledged the financial support from the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Ho, D., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jang, E., Ruano, R.J., Jeffrey, K., Jesmonth, S., Joshi, N., Julian, R., Kalashnikov, D., Kuang, Y., Lee, K.H., Levine, S., Lu, Y., Luu, L., Parada, C., Pastor, P., Quiambao, J., Rao, K., Rettinghouse, J., Reyes, D., Sermanet, P., Sievers, N., Tan, C., Toshev, A., Vanhoucke, V., Xia, F., Xiao, T., Xu, P., Xu, S., Yan, M., Zeng, A.: Do as i can and not as i say: Grounding language in robotic affordances. In: arXiv:2204.01691 (2022)
3. Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., et al.: Do as i can, not as i say: Grounding language in robotic affordances. arXiv:2204.01691 (2022)
4. Amir, S., Gandelsman, Y., Bagon, S., Dekel, T.: Deep vit features as dense visual descriptors. arXiv:2112.05814 (2021)
5. Ardón, P., Pairet, È., Lohan, K.S., Ramamoorthy, S., Petrick, R.: Affordances in robotic tasks–a survey. arXiv:2004.07400 (2020)
6. Ardón, P., Pairet, E., Petrick, R.P., Ramamoorthy, S., Lohan, K.S.: Learning grasp affordance reasoning through semantic relations. RA-L pp. 4571–4578 (2019)
7. Bahl, S., Mendonca, R., Chen, L., Jain, U., Pathak, D.: Affordances from human videos as a versatile representation for robotics. In: CVPR. pp. 13778–13790 (2023)
8. Burke, C.J., Tobler, P.N., Baddeley, M., Schultz, W.: Neural mechanisms of observational learning. PNAS pp. 14431–14436 (2010)
9. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV. pp. 9650–9660 (2021)
10. Chen, J., Gao, D., Lin, K.Q., Shou, M.Z.: Affordance grounding from demonstration video to target image. In: CVPR. pp. 6799–6808 (2023)
11. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML. pp. 1597–1607 (2020)
12. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR. pp. 15750–15758 (2021)
13. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: A deep multi-level network for saliency prediction. In: ICPR. pp. 3488–3493 (2016)

14. Fang, K., Wu, T.L., Yang, D., Savarese, S., Lim, J.J.: Demo2vec: Reasoning object affordances from online videos. In: CVPR. pp. 2139–2147 (2018)
15. Gao, W., Wan, F., Pan, X., Peng, Z., Tian, Q., Han, Z., Zhou, B., Ye, Q.: Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In: ICCV. pp. 2886–2895 (2021)
16. Geng, Y., An, B., Geng, H., Chen, Y., Yang, Y., Dong, H.: Rlafford: End-to-end affordance learning for robotic manipulation. In: ICRA. pp. 5880–5886 (2023)
17. Gibson, J.: The Ecological Approach to Visual Perception. Resources for ecological psychology, Lawrence Erlbaum Associates (1986)
18. Hadjivelichkov, D., Zwane, S., Agapito, L., Deisenroth, M.P., Kanoulas, D.: One-shot transfer of affordance regions? affcorrs! In: CoRL. pp. 550–560 (2023)
19. Hou, Z., Yu, B., Qiao, Y., Peng, X., Tao, D.: Affordance transfer learning for human-object interaction detection. In: CVPR. pp. 495–504 (2021)
20. Huang, Y., Cai, M., Li, Z., Sato, Y.: Predicting gaze in egocentric video by learning task-dependent attention transition. In: ECCV. pp. 754–769 (2018)
21. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. NeurIPS **33**, 18661–18673 (2020)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)
23. Kümmerer, M., Wallis, T.S., Bethge, M.: Deepgaze ii: Reading fixations from deep features trained on object recognition. arXiv:1610.01563 (2016)
24. Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L.M., Shum, H.Y.: Mask dino: Towards a unified transformer-based framework for object detection and segmentation. In: CVPR. pp. 3041–3050 (2023)
25. Li, G., Jampani, V., Sun, D., Sevilla-Lara, L.: Locate: Localize and transfer object parts for weakly supervised affordance grounding. In: CVPR. pp. 10922–10931 (2023)
26. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv:2301.12597 (2023)
27. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML. pp. 12888–12900 (2022)
28. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. NeurIPS pp. 9694–9705 (2021)
29. Liang, J., Huang, W., Xia, F., Xu, P., Hausman, K., Ichter, B., Florence, P., Zeng, A.: Code as policies: Language model programs for embodied control. In: ICRA. pp. 9493–9500 (2023)
30. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. NeurIPS (2023)
31. Luo, H., Zhai, W., Zhang, J., Cao, Y., Tao, D.: Grounded affordance from exocentric view. arXiv:2208.13196 (2022)
32. Luo, H., Zhai, W., Zhang, J., Cao, Y., Tao, D.: Learning affordance grounding from exocentric images. In: CVPR. pp. 2252–2261 (2022)
33. Luo, H., Zhai, W., Zhang, J., Cao, Y., Tao, D.: Learning visual affordance grounding from demonstration videos. IEEE Transactions on Neural Networks and Learning Systems (2023)
34. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research (2008)

35. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. vol. 1, pp. 281–297. Oakland, CA, USA (1967)
36. Mai, J., Yang, M., Luo, W.: Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In: CVPR. pp. 8766–8775 (2020)
37. Mees, O., Borja-Diaz, J., Burgard, W.: Grounding language with visual affordances over unstructured data. In: ICRA. pp. 11576–11582 (2023)
38. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. ICLR (2013)
39. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM pp. 39–41 (1995)
40. Myers, A., Teo, C.L., Fermüller, C., Aloimonos, Y.: Affordance detection of tool parts from geometric features. In: ICRA. pp. 1374–1381 (2015)
41. Nagarajan, T., Feichtenhofer, C., Grauman, K.: Grounded human-object interaction hotspots from video. In: ICCV. pp. 8688–8697 (2019)
42. Nguyen, A., Kanoulas, D., Caldwell, D.G., Tsagarakis, N.G.: Detecting object affordances with convolutional neural networks. In: IROS. pp. 2765–2770 (2016)
43. Nguyen, A., Kanoulas, D., Caldwell, D.G., Tsagarakis, N.G.: Object-based affordances detection with convolutional neural networks and dense conditional random fields. In: IROS. pp. 5908–5915 (2017)
44. Nguyen, T., Vu, M.N., Vuong, A., Nguyen, D., Vo, T., Le, N., Nguyen, A.: Open-vocabulary affordance detection in 3d point clouds. In: IROS. pp. 5692–5698 (2023)
45. Ning, S., Qiu, L., Liu, Y., He, X.: Hoiclip: Efficient knowledge transfer for hoi detection with vision-language models. In: CVPR. pp. 23507–23517 (2023)
46. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv:2304.07193 (2023)
47. Pan, J., Ferrer, C.C., McGuinness, K., O'Connor, N.E., Torres, J., Sayrol, E., Giro-i Nieto, X.: Salgan: Visual saliency prediction with generative adversarial networks. arXiv:1701.01081 (2017)
48. Pan, X., Gao, Y., Lin, Z., Tang, F., Dong, W., Yuan, H., Huang, F., Xu, C.: Unveiling the potential of structure preserving for weakly supervised object localization. In: CVPR. pp. 11642–11651 (2021)
49. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP. pp. 1532–1543 (2014)
50. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
51. Rana, K., Haviland, J., Garg, S., Abou-Chakra, J., Reid, I., Suenderhauf, N.: Sayplan: Grounding large language models using 3d scene graphs for scalable task planning. In: CoRL (2023)
52. Rashid, A., Sharma, S., Kim, C.M., Kerr, J., Chen, L.Y., Kanazawa, A., Goldberg, K.: Language embedded radiance fields for zero-shot task-oriented grasping. In: CoRL (2023)
53. Sawatzky, J., Srikantha, A., Gall, J.: Weakly supervised affordance detection. In: CVPR. pp. 2795–2804 (2017)
54. Singh, I., Blukis, V., Mousavian, A., Goyal, A., Xu, D., Tremblay, J., Fox, D., Thomason, J., Garg, A.: Progprompt: Generating situated robot task plans using large language models. In: ICRA. pp. 11523–11530 (2023)
55. Tang, J., Zheng, G., Yu, J., Yang, S.: Cotdet: Affordance knowledge prompting for task driven object detection. In: ICCV. pp. 3068–3078 (2023)

56. Wan, B., Tuytelaars, T.: Exploiting clip for zero-shot hoi detection requires knowledge distillation at multiple levels. In: WACV. pp. 1805–1815 (2024)
57. Warren, W.: Perceiving affordances: Visual guidance of stair climbing. Journal of experimental psychology. Human perception and performance pp. 683–703 (1984)
58. Xu, R., Chu, F.J., Tang, C., Liu, W., Vela, P.A.: An affordance keypoint detection network for robot manipulation. IEEE RA-L pp. 2870–2877 (2021)
59. Xue, Y., Gan, E., Ni, J., Joshi, S., Mirzasoleiman, B.: Investigating the benefits of projection head for representation learning. In: ICLR (2024)
60. Yu, S., Seo, P.H., Son, J.: Zero-shot referring image segmentation with global-local context features. In: CVPR. pp. 19456–19465 (2023)
61. Zhang, J., Herrmann, C., Hur, J., Cabrera, L.P., Jampani, V., Sun, D., Yang, M.H.: A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. arXiv:2305.15347 (2023)
62. Zhang, X., Wang, D., Han, S., Li, W., Zhao, B., Wang, Z., Duan, X., Fang, C., Li, X., He, J.: Affordance-driven next-best-view planning for robotic grasping. In: CoRL (2023)
63. Zhao, X., Li, M., Weber, C., Hafez, M.B., Wermter, S.: Chat with the environment: Interactive multimodal perception using large language models. arXiv:2303.08268 (2023)
64. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR. pp. 2921–2929 (2016)
65. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv:2304.10592 (2023)