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 [20] 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.
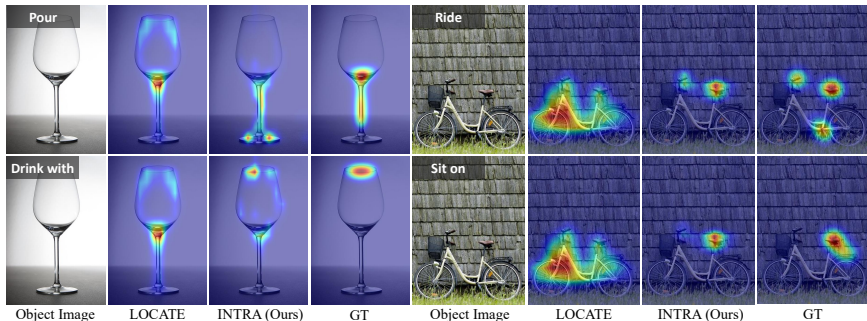
Fig. 1: Prior works on weakly-supervised affordance grounding like LOCATE [29] often failed to ground different affordances for the same object. However, our proposed INTRA yielded finer and more accurate grounding results for them that are closer to the ground truth (GT) by considering interaction relationship 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 [66]. 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, 19, 58, 74].

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, 21, 47, 69] 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 [29, 35, 36, 46] 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 [29] 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 [35,46] solved the weak supervision problem as supervised learning by pulling object features of exocentric and egocentric images closer and LOCATE [29] 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 [12,17] 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 generaliz-
ability 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 [15, 23, 27, 29, 35–37, 46, 52] offer the advantage of not requir-
ing GT, but requiring weak labels such as exocentric images with interaction
text labels. Prior works [29, 35, 36, 46] mainly align interaction-relevant object
features from both egocentric and exocentric images without considering the
intrinsic properties of interactions. The framework in [46] predicts object fea-
tures engaged in interactions by analyzing human-object interaction videos. The
works of [35, 36] preserve the correlation of affordance features from exocentric
and egocentric images to learn affordances. The work of [29] enhances object
feature extraction by adopting DINO-ViT [10] based Class Activation Maps
(CAM) [76] and k-means clustering [39] for more explicit guidance. However, fo-
cusing 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 re-
mains 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. Pre-
vious works [29, 59] have explored their potential in affordance grounding. DINO-
ViT [10], a vision transformer foundation model trained in a self-supervised man-
ner, can identify both high-semantics such as overall information of the image
and low-semantics such as details regarding specific object parts. This versa-
tility has led advancements in various tasks, including classification, semantic
segmentation [4, 28] and semantic correspondence [73]. LOCATE [29] leverages
DINO-ViT to extract low-semantic information, resulting in performance im-
provements in affordance grounding. Our INTRA employed DINOv2 [51] 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 [50, 63, 72]. VLM text encoders, trained through contrastive learning with
image-text pairs, capture representations in the joint space of the images and
text [30–32, 57]. These text encoders, incorporating visual information, have
demonstrated excellent performance across multiple tasks. ALBEF [32] notably
enhances vision and language representation learning by aligning image and
text features before fusing them. While supervised affordance grounding meth-
ods leveraging VLM text encoders [49] have been explored, their application in
weakly supervised affordance grounding remains underexplored. We propose a
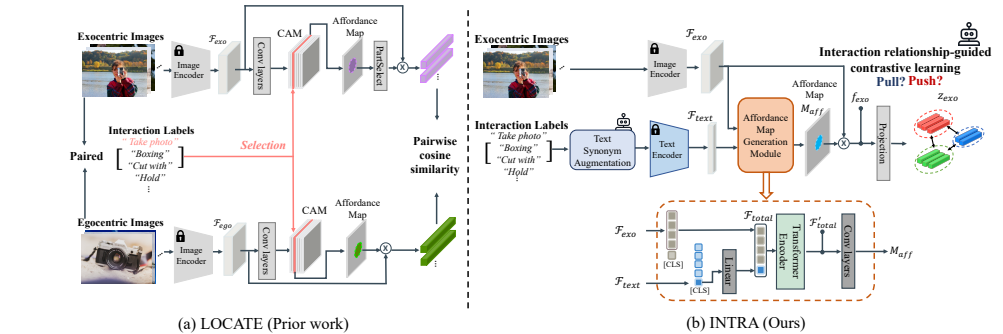
Fig. 2: Overall frameworks of (a) LOCATE [29] and (b) INTRA (Ours). LOCATE takes paired exocentric and egocentric images to generate interaction-aware affordance maps (CAMs) for predefined interactions and then selects an interaction-related CAM by the given interaction label. In contrast, INTRA takes 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 [22] leverage semantically similar object properties and [36] 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, 33, 61, 75] 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 [41, 62], demonstrating their exceptional understanding of interactions.

3   Method

Prior arts in weakly supervised affordance grounding [29, 35, 36, 46] 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 [29] 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 [29]: 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 [51], 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 [32]. 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 [34, 77]. 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 [13, 14, 70], which have demonstrated the necessity and efficiency of it.

Loss design for interaction relationship-guided contrastive learning. Supervised contrastive learning [25] 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
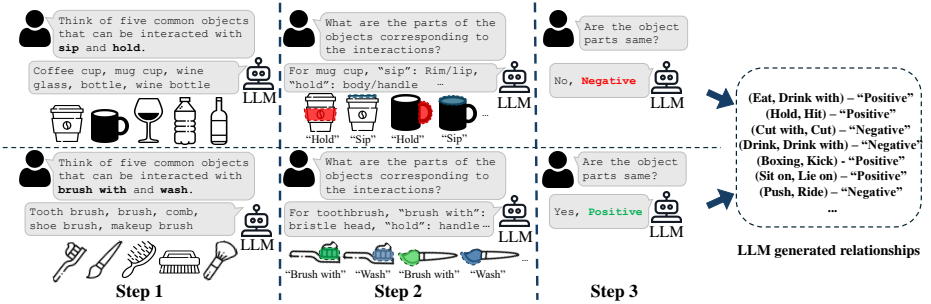
Fig. 3: The overall scheme of interaction-relationship map ($\mathcal{R}$) generation. LLM classifies all pairs of interactions in the dataset as positive or negative through chain of thoughts. This process is based on reasoning if interactions occur on same object parts.

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:

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

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_{exo}^j 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:

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

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) [36]. 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 [26] 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 [45, 48, 60] for a comprehensive evaluation. Refer to the supplementary materials for more details on the experimental settings.

Quantitative results. We evaluated previous works [18, 29, 35, 36, 40, 46, 53] 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 [18, 29, 35, 36, 40, 46, 53] 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 [40]	1.931	0.285	0.522	2.167	0.227	0.330
	SPA [53]	5.528	0.221	0.357	7.425	0.167	0.262
	TS-CAM [18]	1.842	0.260	0.336	2.104	0.201	0.151
Weakly Supervised Affordance Grounding	Exo+Ego	Hotspots [46]	1.773	0.278	0.615	1.994	0.237	0.557
		Cross-view-AG [36]	1.538	0.334	0.927	1.787	0.285	0.829
		Cross-view-AG+ [35]	1.489	0.342	0.981	1.765	0.279	0.882
		LOCATE [29]	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 [29,35,36]. 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.

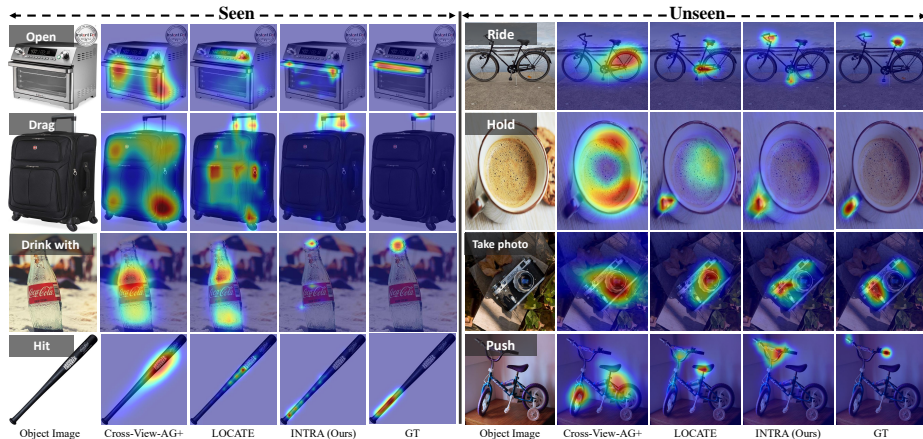|	IIT-AFF* [48]			CAD* [60]			UMD* [45]		
	mKLD↓	mSIM↑	mNSS↑	mKLD↓	mSIM↑	mNSS↑	mKLD↓	mSIM↑	mNSS↑
Cross-View-AG [36]	3.856	0.096	0.849	2.568	0.173	0.589	4.721	0.014	1.287
Cross-View-AG+ [35]	3.920	0.095	1.072	2.529	0.176	0.663	4.753	0.013	1.227
LOCATE [29]	3.315	0.115	1.709	2.528	0.187	0.558	4.083	0.026	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 [29, 35, 36] trained in the 'Seen' setting of AGD20K, on the IIT-AFF [48], CAD [60], and UMD [45] 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-

| Object Image | Cross-View-AG | Cross-View-AG+ | LOCATE | INTRA (Ours) | GT |
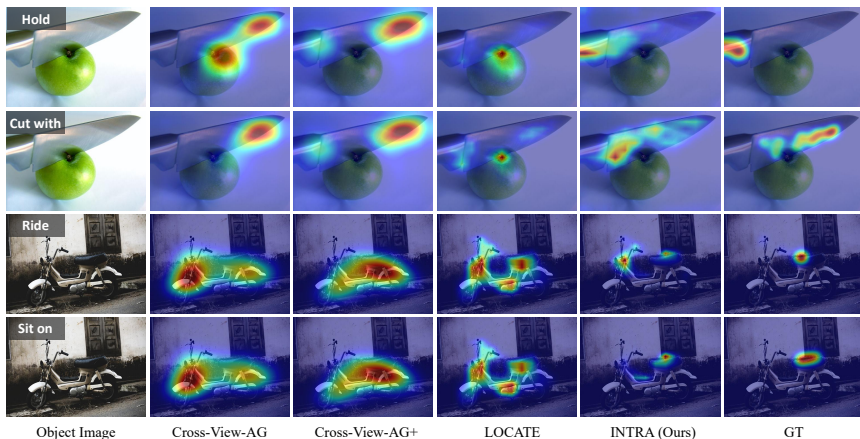
Fig. 4: Qualitative results of INTRA (Ours) and baseline models [29,35,36] on grounding affordances of multiple potential interactions on a single object. INTRA precisely localizes relevant interaction spots for each interaction. For example, with a knife, it grounds the handle for 'Hold' and the blade 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.

| Object Image | Cross-View-AG+ | LOCATE | INTRA (Ours) | GT | Object Image | Cross-View-AG+ | LOCATE | INTRA (Ours) | GT |

Fig. 5: Qualitative results comparison between our approach and other baselines [29, 35,36]. 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+ [35]	2.897	3.022	2.732
LOCATE [29]	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 [44], Word2Vec [43], respectively. $\mathcal{L}_{Co-occur.}$ used co-occurrence probability in GloVe [54].

|	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+ [35], LO-CATE [29], 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
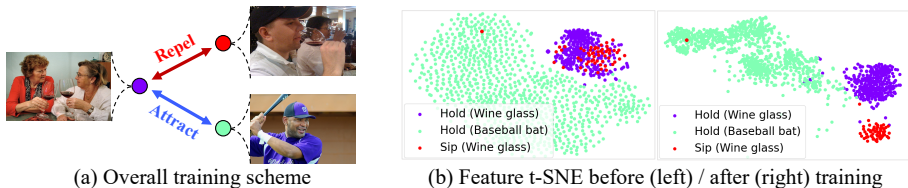
Fig. 6: An illustration of interaction relationship-guided contrastive learning and t-SNE [38] visualization of feature distribution. (a) In interaction relationship-guided contrastive learning, positive interaction pairs attract each other, while others repel. (b) t-SNE visualization of DINOv2 [51] class token and f_{exo} from INTRA, showing that features of positive interaction pairs become closer as learning progresses.
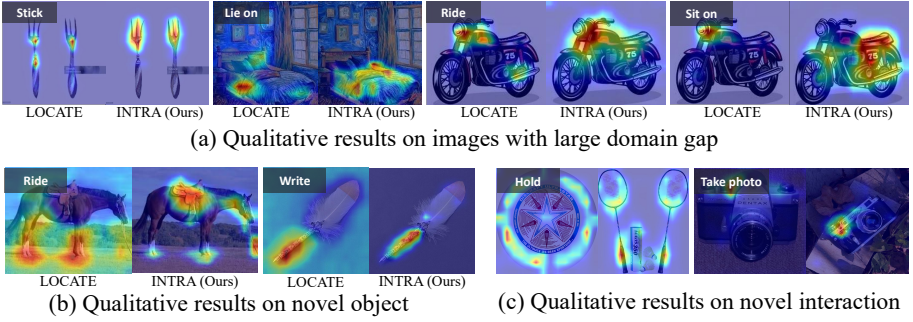
choice. We measured the similarity of interaction pairs using WordNet [44] and Word2Vec [43], or computed co-occurence probability of interaction pairs with Glove [54]. Based on these measurements, we created an Interaction-relationship Map and trained the INTRA framework. The results are in the Tab. 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.

(a) Qualitative results on images with large domain gap

(b) Qualitative results on novel object

(c) Qualitative results on novel interaction

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 [29], 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

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. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., Durand, F.: What do different evaluation metrics tell us about saliency models? arXiv preprint arXiv:1604.03605 (2016)
10. 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)
11. Chao, Y.W., Wang, Z., He, Y., Wang, J., Deng, J.: HICO: A benchmark for recognizing human-object interactions in images. In: ICCV (2015)
12. Chen, J., Gao, D., Lin, K.Q., Shou, M.Z.: Affordance grounding from demonstration video to target image. In: CVPR. pp. 6799–6808 (2023)
13. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: ICML. pp. 1597–1607 (2020)

14. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR. pp. 15750–15758 (2021)
15. Cornia, M., Baraldi, L., Serra, G., Cucchiara, R.: A deep multi-level network for saliency prediction. In: ICPR. pp. 3488–3493 (2016)
16. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018)
17. 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)
18. 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)
19. 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)
20. Gibson, J.: The Ecological Approach to Visual Perception. Resources for ecological psychology, Lawrence Erlbaum Associates (1986)
21. Hadjivelichkov, D., Zwane, S., Agapito, L., Deisenroth, M.P., Kanoulas, D.: One-shot transfer of affordance regions? affcorrs! In: CoRL. pp. 550–560 (2023)
22. Hou, Z., Yu, B., Qiao, Y., Peng, X., Tao, D.: Affordance transfer learning for human-object interaction detection. In: CVPR. pp. 495–504 (2021)
23. 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)
24. Jiang, H., Ma, X., Nie, W., Yu, Z., Zhu, Y., Anandkumar, A.: Bongard-hoi: Benchmarking few-shot visual reasoning for human-object interactions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19056–19065 (2022)
25. 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)
26. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)
27. Kümmerer, M., Wallis, T.S., Bethge, M.: Deepgaze ii: Reading fixations from deep features trained on object recognition. arXiv:1610.01563 (2016)
28. 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)
29. 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)
30. 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)
31. 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)
32. 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)
33. 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)

34. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. NeurIPS (2023)
35. Luo, H., Zhai, W., Zhang, J., Cao, Y., Tao, D.: Grounded affordance from exocentric view. arXiv:2208.13196 (2022)
36. Luo, H., Zhai, W., Zhang, J., Cao, Y., Tao, D.: Learning affordance grounding from exocentric images. In: CVPR. pp. 2252–2261 (2022)
37. 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)
38. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research (2008)
39. 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)
40. 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)
41. Mees, O., Borja-Diaz, J., Burgard, W.: Grounding language with visual affordances over unstructured data. In: ICRA. pp. 11576–11582 (2023)
42. Michael J. Swain, D.H.B.: Color indexing. IJCV (1991)
43. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. ICLR (2013)
44. Miller, G.A.: Wordnet: a lexical database for english. Communications of the ACM pp. 39–41 (1995)
45. Myers, A., Teo, C.L., Fermüller, C., Aloimonos, Y.: Affordance detection of tool parts from geometric features. In: ICRA. pp. 1374–1381. IEEE (2015)
46. Nagarajan, T., Feichtenhofer, C., Grauman, K.: Grounded human-object interaction hotspots from video. In: ICCV. pp. 8688–8697 (2019)
47. Nguyen, A., Kanoulas, D., Caldwell, D.G., Tsagarakis, N.G.: Detecting object affordances with convolutional neural networks. In: IROS. pp. 2765–2770 (2016)
48. 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)
49. 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)
50. 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)
51. 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)
52. 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)
53. 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)
54. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP. pp. 1532–1543 (2014)
55. Peters, R.J., Iyer, A., Itti, L., Koch, C.: Components of bottom-up gaze allocation in natural images. Vision Research pp. 2397–2416 (2005)
56. Pratt, S., Yatskar, M., Weihs, L., Farhadi, A., Kembhavi, A.: Grounded situation recognition. ArXiv **abs/2003.12058** (2020)

57. 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)
58. Rana, K., Haviland, J., Garg, S., Abou-Chakra, J., Reid, I., Suenderhauf, N.: Say-plan: Grounding large language models using 3d scene graphs for scalable task planning. In: CoRL (2023)
59. 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)
60. Sawatzky, J., Srikantha, A., Gall, J.: Weakly supervised affordance detection. In: CVPR. pp. 2795–2804 (2017)
61. 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)
62. Tang, J., Zheng, G., Yu, J., Yang, S.: Cotdet: Affordance knowledge prompting for task driven object detection. In: ICCV. pp. 3068–3078 (2023)
63. Wan, B., Tuytelaars, T.: Exploiting clip for zero-shot hoi detection requires knowledge distillation at multiple levels. In: WACV. pp. 1805–1815 (2024)
64. Wang, F., Liu, H.: Understanding the behaviour of contrastive loss. In: CVPR. pp. 2495–2504 (2021)
65. Wang, S., Yap, K.H., Ding, H., Wu, J., Yuan, J., Tan, Y.P.: Discovering human interactions with large-vocabulary objects via query and multi-scale detection. In: ICCV (2021)
66. Warren, W.: Perceiving affordances: Visual guidance of stair climbing. Journal of experimental psychology. Human perception and performance pp. 683–703 (1984)
67. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. NeurIPS pp. 24824–24837 (2022)
68. Wu, Z., Palmer, M.: Verb semantics and lexical selection. In: 32nd Annual Meeting of the Association for Computational Linguistics. pp. 133–138. Association for Computational Linguistics, Las Cruces, New Mexico, USA (Jun 1994)
69. 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)
70. Xue, Y., Gan, E., Ni, J., Joshi, S., Mirzasoleiman, B.: Investigating the benefits of projection head for representation learning. In: ICLR (2024)
71. Yang, Y., Zhai, W., Luo, H., Cao, Y., Luo, J., Zha, Z.J.: Grounding 3d object affordance from 2d interactions in images. arXiv:2303.10437 (2023)
72. Yu, S., Seo, P.H., Son, J.: Zero-shot referring image segmentation with global-local context features. In: CVPR. pp. 19456–19465 (2023)
73. 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)
74. 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)
75. 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)
76. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR. pp. 2921–2929 (2016)

77. 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)

Supplementary Material for
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

A Additional Experimental Details

A.1 Dataset

Dataset description. AGD20K [36] is an affordance grounding dataset, consisting of 20,061 exocentric images and 3,775 egocentric images that are categorized based on object and interaction labels. During evaluation, egocentric images and interaction labels are provided to identify the most relevant regions of interaction in object images. Note that Ground Truth (GT) masks of AGD20K were annotated by the interactions between humans and objects in the OPRA dataset [17].

Seen setting. The 'Seen' setting of the AGD20K dataset comprises 36 *interaction* labels and the train and test sets each contains 50 object categories.

Unseen setting. The 'Unseen' setting of the AGD20K dataset includes 25 *affordance* categories. Unlike the 'Seen' setting, the object classes in train and test sets do not overlap so that verification of whether the network can infer affordances for previously unseen objects is possible. There are 35 object classes in the train set and 12 object classes in the test set as the following object categories:

- **Train set:** *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 set:** *axe, banana, bed, bicycle, broccoli, camera, cup, golf clubs, laptop, refrigerator, skis, soccer ball*

---

* Authors contributed equally. † Corresponding author.

A.2   Metrics

Unlike segmentation tasks where GT is usually a binary mask, affordance grounding involves mapping the probability of an action on an object, thereby necessitating a probabilistic representation for GT. Following previous works [12, 29, 35–37, 46, 71], we employ Kullback-Leibler divergence (KLD) [9], similarity (SIM) [42], and normalized scanpath saliency (NSS) [55] as metrics to evaluate our method where KLD quantifies the discrepancy between two probability distributions, SIM measures the similarity between two distributions, and NSS evaluates the correspondence between two maps. The details for computing these metrics are as follows. Firstly, we resize the GT mask $\mathcal{M}_{GT}$ and the model's predicted attention map $\mathcal{M}_{aff}$ to 224×224 using bilinear interpolation so that $\{\mathcal{M}_{GT}, \mathcal{M}_{aff}\} \in \mathbb{R}^{224 \times 224}$. Then, they are min-max normalized and each element is divided by the sum of all elements to obtain $\hat{\mathcal{M}}_{GT}$ and $\hat{\mathcal{M}}_{aff}$ as follows:

$$\hat{\mathcal{M}}_{GT} = \mathcal{M}_{GT} / \sum \mathcal{M}_{GT}, \ \hat{\mathcal{M}}_{aff} = \mathcal{M}_{aff} / \sum \mathcal{M}_{aff}. \tag{1}$$

Lastly, Using $\hat{\mathcal{M}}_{GT}$ and $\hat{\mathcal{M}}_{aff}$, KLD and SIM are calculated as following:

$$\mathrm{KLD}(\hat{\mathcal{M}}_{GT} || \hat{\mathcal{M}}_{aff}) = \sum \hat{\mathcal{M}}_{GT} \cdot \log\left(\frac{\hat{\mathcal{M}}_{GT}}{\hat{\mathcal{M}}_{aff}}\right), \tag{2}$$

$$\mathrm{SIM}(\hat{\mathcal{M}}_{GT}, \hat{\mathcal{M}}_{aff}) = \sum \min(\hat{\mathcal{M}}_{GT}, \hat{\mathcal{M}}_{aff}). \tag{3}$$

The following calculations are conducted to compute NSS:

$$\tilde{\mathcal{M}}_{GT} = \mathbf{1}(\mathcal{M}_{GT} > 0.1), \ \tilde{\mathcal{M}}_{aff} = \frac{\mathcal{M}_{aff} - \mu_{\mathcal{M}_{aff}}}{\sigma_{\mathcal{M}_{aff}}} \tag{4}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function, $\mu_{\mathcal{M}_{aff}}$ and $\sigma_{\mathcal{M}_{aff}}$ represent the mean and standard deviation of $\mathcal{M}_{aff}$, respectively. From these, we calculate NSS as follows:

$$\mathrm{NSS}(\tilde{\mathcal{M}}_{GT}, \tilde{\mathcal{M}}_{aff}) = \frac{1}{\sum \tilde{\mathcal{M}}_{GT}} \sum \tilde{\mathcal{M}}_{GT} \cdot \tilde{\mathcal{M}}_{aff}. \tag{5}$$

A.3   User study

Affordances can be ambiguous due to several factors, including context dependence, perceptual limitations, and subjective interpretations. Therefore, we undertook a comprehensive assessment of our INTRA's prediction output through a user study. The study addressed three aspects: 'validity', which evaluates fidelity of the affordance grounding map, 'finesse', which evaluates granularity and detail of the presented affordance map, and 'separability', which evaluates the model's ability to appropriately ground the same object in different locations for various interactions. Eight interactions ('push', 'cut with', 'take photo', 'sip', 'open', 'sit on', 'pour', 'hold') and nine objects ('motorcycle', 'scissors', 'camera', 'cup', 'microwave', 'refrigerator', 'bicycle', 'wine glass', 'knife') were

chosen at random. Respondents assigned scores ranging from 1 to 5 to the affordance maps generated by prior arts (Cross-View-AG+ [35], LOCATE [29]), GT itself, and our proposed INTRA (Ours). The sequence of results from each model was randomized for each questionnaire. A total of 104 respondents evaluated the 4 models (including the oracle or GT) across the 9 items.

### A.4   Experiment on additional datasets

In addition to AGD20K, our INTRA was evaluated on three additional datasets: IIT-AFF [48], CAD [60], and UMD [45]. While these datasets were originally created for affordance segmentation (not specifically for affordance grounding), they can still provide valuable evaluations and comparisons on how well our IN-TRA performs affordance grounding in terms of accuracy and finesse compared to other prior arts on generalized datasets. We made the binary segmentation mask using a threshold of 0, thus setting the mask value to be 1 if the value in the mask exceeds 0. Since prior arts in affordance grounding [29, 35, 36] can not predict the affordances that were not part of the training data, the affordances that were not included in the train set were excluded from the dataset for fair comparisons (advantageous for prior works). With the modified dataset, we assessed the models with a total of 9,797 images from IIT-AFF, 20,016 images from CAD, and 39,846 images from UMD. Since the majority of objects presented in these datasets are also included in the AGD20K dataset, we conducted evaluations utilizing a model trained under the 'Seen' setting. All remaining evaluation processes were carried out in the same manner as described in Sec A.2.

B   Additional Pipeline Details

B.1   Additional details on network architecture

Design choices for network architecture. $\mathcal{F}_{exo}$ was extracted using the pre-trained DINOv2-base [51] as an image encoder, with a patch size of 14 and a feature dimension of 768. For $\mathcal{F}_{text}$, we employed the text encoder of ALBEF [32] based on BERT-base-uncased [16]. In the affordance map generation module, a transformer encoder architecture stacked with 4-layers and 4-head attention was utilized. Finally, the 2-layer convolution network in the module projects 768-channel feature into a single-channel $\mathcal{M}_{aff}$. For the projection layer, a 3-layer MLP with the output dimension of 128 was adopted to generate z_{exo}, which was later utilized in contrastive learning.

Rationale for adopting DINO. DINOv2 [51] was selected as the image encoder for our INTRA due to its superior performance in extracting low-level features from images. To visualize the ability of DINOv2, we projected DINOv2 image features onto 3 principal components using 3-dimensional PCA, and then normalized each dimension from 0 to 1. We then mapped the values to integers
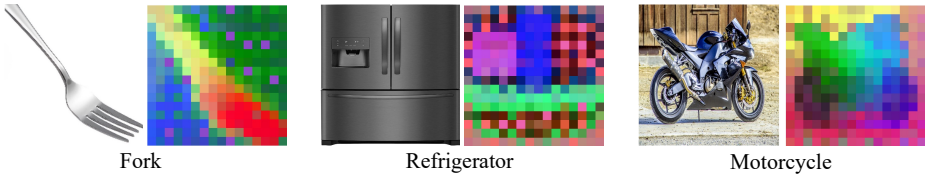
Fig. S1: PCA analysis of DINOv2 [51] features on three objects: fork, refrigerator, motorcycle. As depicted in the figure, the DINOv2 is capable of distinguishing the tip and handle of the fork; the door of the refrigerator; the handle, saddle, and the wheel of the motorcycle.

from 0 to 255 to obtain an RGB image. As depicted in Fig. S1, DINOv2 is capable of distinguishing the object parts: the tip and handle of a fork, the door of a refrigerator, and the handle, saddle, and the wheel of a motorcycle.

B.2   Additional details on usage of LLM

Text synonym augmentation. In text synonym augmentation, LLM was used to find synonyms. Examples of queries for finding synonyms can be found in Tab. S1. To provide the exact context of interaction, examples of interaction and object pairs were given. Three synonyms were found, and those that overlapped with labels of other classes were replaced before usage. The interaction names used in conditioning were replaced with generated synonyms with a 20% probability, while the names of classes used in interaction relationship-guided contrastive learning remained unchanged. Examples of the synonyms used are depicted in Tab. S2. ChatGPT-4 [1] was adopted as LLM, and no hyperparameters were used.

Query for interaction relationship map. We utilized LLM to determine whether the given affordance pair acts on the same part of an object, *i.e.*, whether the pair is positive or not. LLM was provided with a few examples, comprising 1.4% of the whole interaction pair, for the task and then asked to classify the given pair with Chain-of-Thoughts (CoT) [67]. The full query for the generation of interaction relationship map is provided in Tab. S3. ChatGPT-4 [1] was adopted as LLM, and no hyperparameter was used.

Table S1: An example of LLM query and answer to find synonyms for text synonym augmentation.

Example of LLM query for text synonym augmentation

User	Give me three synonyms for each verbs in form of ['word1', 'word2', 'word3']. 1. 'beat' in context of 'beat drum' 2. 'boxing' in context of 'boxing punching bag' 3. 'brush with' in context of 'brush with toothbrush' 4. 'carry' in context of 'carry suitcase' 5. 'catch' in context of 'catch baseball'
LLM	1. ['strike', 'hit', 'pound'] 2. ['punching', 'sparring', 'training'] 3. ['clean with', 'scrub with', 'sweep with'] 4. ['transport', 'haul', 'convey'] 5. ['grab', 'snag', 'snatch']

Table S3: An example of LLM query and answer to determine whether ['drink with', 'sip'] pair is 'positive' or 'negative' pair.

Example of LLM query for interaction-relationship map generation

User	You should clarify each verb's relation whether it is positive pair or negative pair. Here's some example of the examples of clarification. ex1) 'hold' and 'carry' is 'positive' pair. Because if we do the given interaction to the object like suitcase, ski, or snowboard, we interacts with the same part of the object. The side part of the object for 'hold' and 'carry'. ex2) 'hit' and 'carry' is 'positive' pair. Because if we do the given interaction to the object like baseball bat or badminton racket, we interacts with the same part of the object. The thin part of the object for 'hit' and 'carry'. ex3) 'cut with' and 'hold' is 'negative' pair. Because if we 'cut with' the object like scissor, or knife, we interacts with the different part of the object. Sharp part of the object for 'cut with' and 'hold' for handle or dull part. ex4) 'sip' and 'hold' is 'negative' pair. Because if we do the given interaction to the object like cup, bottle, or wine glass, we interacts with the different part of the object. Rim of the object for 'sip', and 'hold' for handle or round part. Think of 5 objects that can be commonly interacted with. For each object, describe the interactions and then list the object parts that should be interacted with step by step. Use this information to create criteria for classification based on the given example. Specify the final answer with short format and it should be one of [positive, negative]. ['drink with', 'sip'] → ['positive', 'negative']?

LLM   1. Cup

- Drink with: When using a cup to drink, the action involves bringing the rim of the cup to your lips and tilting it to allow the liquid to flow into your mouth.
- Sip: Sipping from a cup also specifically involves using the rim. The key distinction is in the amount of liquid and the care with which it's consumed, but it's the same part of the cup — the rim — that facilitates the action.

2. Straw

- Drink with: Drinking with a straw involves placing one end of the straw in your mouth and the other end in the beverage. The interaction is exclusively with the straw as the medium through which the beverage is drawn up.
- Sip: Sipping through a straw also involves the same exact use of the straw. The action might be more gentle or controlled, but it still requires using the straw to draw the liquid up in the same manner.

3. Glass

- Drink with: Using a glass to drink involves directly interacting with the rim of the glass. You typically tilt the glass to pour the beverage into your mouth, engaging directly with the rim.
- Sip: Sipping from a glass, much like drinking, involves delicate interaction with the rim of the glass. The rim is the point of contact for controlling the intake of the liquid, even if the sipping is slower or in smaller quantities.

4. Bottle

- Drink with: Drinking from a bottle involves placing your lips around the opening or the rim of the bottle. The beverage is consumed by tilting the bottle and allowing the liquid to flow into your mouth through the opening.
- Sip: Sipping from a bottle uses the same opening. The primary interaction is still with the rim or the mouth of the bottle, with the main difference being the amount of liquid consumed and the rate at which it is consumed.

5. Tea Cup

- Drink with: Drinking from a tea cup involves using the rim of the tea cup, typically held by a handle for ease. The rim is the point of contact with the mouth, and the tea is consumed by tilting the cup.
- Sip: Sipping from a tea cup also involves the rim. While sipping might suggest a more refined or controlled action, especially with hot beverages like tea, it still requires interaction with the same part of the tea cup — the rim.

In each case, drink with and sip utilize the same part of these objects, confirming that they form a positive pair based on their interaction with the same part of the object.

Table S2: An example of synonyms generated by LLM for text synonym augmentation.

Examples of synonyms for text synonym augmentation	
beat	strike, tap, pound
boxing	punching, sparring, kicking
brush with	clean with, scrub with, sweep with
carry	transport, haul, convey
catch	grab, snag, snatch
cut	slice, chop, carve
cut with	slice with, carve with, cut with
drag	pull, haul, tow
drink with	drink from, quaff from, imbibe from
eat	consume, devour, ingest
hit	strike, smack, blow
hold	grasp, embrace, grip
jump	skate, glide, roll
kick	kick out, boot, strike
lie on	rest on, recline on, lie upon
lift	hold, raise, hoist
look out	watch out, gaze out, observe
open	unzip, unpack, reveal
pack	bundle, assemble, prepare
peel	peel off, strip off, skin
pick up	pick up, collect, gather
pour	flow, spill, dispense
push	shove, press, thrust
ride	cycle, pedal, roll
sip	nibble, salute with, taste
sit on	sit upon, rest on, perch on
stick	pike with, thrust with, jab with
stir	mix, whisk, blend
swing	swipe, swat, clout
take photo	photograph with, capture with, shoot with
talk on	speak on, communicate on, converse on
text on	message on, compose on, write on
throw	toss, fling, hurl
type on	key in, enter with, type with
wash	rinse, scrub, cleanse
write	record with, scrible with, jot with

C   Additional Ablation Studies

C.1   Ablation study on interaction relationship map

The ablation study in the main paper includes quantitative results with various relationship maps. $\mathcal{R}$ generated with LLM (Ours), was substituted with $\mathcal{R}_{WordNet}$, $\mathcal{R}_{Co-Occur.}$, and $\mathcal{R}_{Word2Vec}$. $\mathcal{R}_{WordNet}$ was generated by calculating Wu-Palmer similarities [68] for each interaction pair using WordNet [44]. $\mathcal{R}_{Co-Occur.}$ was generated using the co-occurence probability of each interaction pair, which is the inner product of word vectors from the GloVe [54] 840B model. $\mathcal{R}_{Word2Vec}$ contains the similarity of each interaction pair calculated with Word2Vec [43]. All matrices were converted to binary with threshold of 0.5.

C.2   Ablation study on text encoder

A text encoder was employed to enhance the flexibility to input interactions and improve robustness against unseen interactions by leveraging VLM's text encoder, recognizing the intimate relationship between affordance grounding and visual information. We assessed the performance of various text embedding methods by integrating each encoder into our architecture, as detailed in Tab. S4. It is evident that while random embedding under nearly orthogonal conditions yields satisfactory performance, it struggles to infer novel interactions. On the other hand, employing BERT [16] enables the inference of novel interactions, although the ALBEF [32] text encoder demonstrates superior performance.

Table S4: Quantitative results of ablation study on various text encoders. For the 'Random', a 768-dimensional random vector was initialized from a Gaussian distribution for each interaction. Just as in the INTRA model, only the class token of the BERT [16] embedding was employed for 'BERT'. The results indicate that the ALBEF [32] text encoder outperforms others, because it embeds visual information in the text.

	Encoder	mKLD↓	mSIM↑	mNSS↑
Seen	Random	1.230	0.392	1.209
	BERT [16]	1.286	0.389	1.138
	ALBEF [32]	1.199	0.407	1.239
Unseen	Random	1.520	0.319	1.118
	BERT [16]	1.458	0.321	1.190
	ALBEF [32]	1.365	0.375	1.209

C.3   Ablation study on the number of projection layers

We conducted ablation experiments on the number of projection layers. The quantitative results can be found in Tab. S5. The importance of the projection

layer in contrastive learning is already well-known in prior works [13, 14, 70]. Based on these experimental results, we used three projection layers for our pipeline.

Table S5: Quantitative results of ablation study on the number of projection layers. We changed the number of projection layers, trained INTRA and compared the performance. Based on this experimental results, we used three projection layers for our pipeline.

	# of proj. layers	mKLD↓	mSIM↑	mNSS↑
Seen	1	1.380	0.393	1.085
	2	1.260	0.401	1.157
	3	1.199	0.407	1.239
	4	1.223	0.400	1.198
Unseen	1	1.680	0.298	0.891
	2	1.593	0.320	0.931
	3	1.365	0.375	1.209
	4	1.497	0.324	1.101

C.4   Ablation study on text synonym augmentation

The detailed quantitative results of ablation study on the text synonym augmentation are provided in Tab. S6 and Tab. S7. As shown in the table, text synonym augmentation enhances overall performance as well as the performance on novel verbs. The subsets 'hold', 'swing' and 'take photo' were selected based on their proportion in the whole dataset to demonstrate that text synonym augmentation improves performance on novel verbs regardless of their proportion. 'hold', 'swing' and 'take photo' each accounts for 24.17%, 3.82% and 2.45% of the train set.

Table S6: Quantitative results of ablation study on the text synonym augmentation. Text synonym augmentation enhances overall performance, especially in 'Unseen' setting. 'w/o aug.' denotes that the inference was done without text synonym augmentation and 'w/ aug.' denotes that the inference was done with text synonym augmentation.

	Seen			Unseen		
	mKLD↓	mSIM↑	mNSS↑	mKLD↓	mSIM↑	mNSS↑
w/o aug.	1.288	0.386	1.151	1.563	0.322	1.058
w/ aug.	1.199	0.407	1.239	1.365	0.375	1.209

Table S7: Quantitative results of ablation study on the text synonym augmentation. Text synonym augmentation improves performance in novel verb inference, irrespective of their proportion in the train set.

	'hold'			'swing'			'take photo'		
	mKLD↓	mSIM↑	mNSS↑	mKLD↓	mSIM↑	mNSS↑	mKLD↓	mSIM↑	mNSS↑
w/o aug.	1.533	0.317	0.825	1.828	0.212	0.940	0.922	0.472	1.070
w/ aug.	1.344	0.356	1.304	1.789	0.216	1.031	0.648	0.555	1.320

Table S8: Quantitative results of ablation study on the temperature of interaction relationship-guided contrastive loss. In temperature-scaled contrastive loss, selecting the appropriate temperature is crucial for achieving optimal model performance. Utilizing this property of contrastive loss, we conducted experiments to find a suitable value of temperature and trained INTRA with $\tau = 0.2$.

|	Temp. (τ)	mKLD↓	mSIM↑	mNSS↑
Seen	0.07	1.283	0.396	1.143
	0.1	1.263	0.384	1.173
	0.2	1.199	0.407	1.239
	0.4	1.291	0.382	1.137
Unseen	0.07	1.462	0.333	1.126
	0.1	1.527	0.336	1.074
	0.2	1.365	0.375	1.209
	0.4	1.550	0.324	1.025

C.5 Ablation study on hyperparameter

Temperature of interaction relationship-guided contrastive loss. In temperature-scaled contrastive loss, temperature is one of the important hyperparameters that determines the gap between positive and confusing negative samples. It is also well-known that the performance of the model varies depending on this value [64]. Specifically, in the affordance grounding task, setting the gap between hard negatives and positives is crucial due to instances where different parts of an object need to be localized in the same input image. Through exhaustive experiments, we have found a suitable value of temperature, and some of the results are presented in Tab. S8.

Ratio of object-variance mitigation loss. The object-variance mitigation loss aims to mitigate the dissimilarity among object classes within a single interaction class. For example, the 'hold' interaction class contains 21 different objects, such as 'axe', 'badminton racket', 'baseball bat', 'book', 'bottle', 'bowl' and 'frisbee', each exhibiting distinct visual characteristics. Through experiments, we have determined an appropriate ratio for the object-variance mitigation loss. As depicted in Tab. S9, there was no significant difference in performance, but $\lambda_{obj} = 4$ yields the best results for our network.

Table S9: Quantitative results of ablation study on the ratio of object-variance mitigation loss. Although this coefficient, λ_{obj}, does not have significant effect on performance of our model, $\lambda_{obj} = 4$ works best for our network, as shown in the table.

	λ_{obj}	mKLD↓	mSIM↑	mNSS↑
Seen	1	1.288	0.384	1.160
	2	1.249	0.395	1.196
	4	1.199	0.407	1.239
	8	1.288	0.390	1.155
Unseen	1	1.554	0.302	1.083
	2	1.629	0.305	0.980
	4	1.365	0.375	1.209
	8	1.624	0.294	0.939

(a) Affordance grounding on 'adjust (tie)'   (b) Affordance grounding on 'carry (chair)'
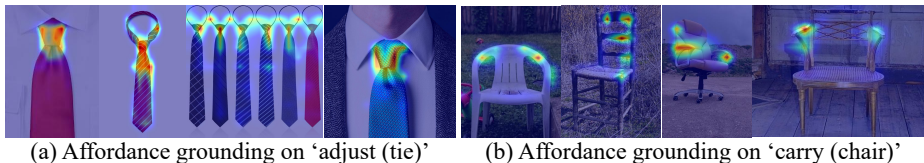
Fig. S2: Qualitative results of INTRA trained with HICO-DET [11]. (a) 'adjust (tie)' accurately grounds the knot of the tie which human usually interacts with tie to 'adjust' it. (b) 'carry (chair)' precisely highlights the arm rest and back of the chair which are part of object facilitate for carrying.

D   Additional Experiments on Loss Design

Effectiveness of LLM usage. To generate an interaction relationship map, a comprehensive prior knowledge about interaction is required, as explained previously. While it is technically feasible to utilize a manually annotated matrix, it is suboptimal due to its scalability. For a given number of interactions, denoted as N_{inter}, the number of pairs that need to be determined grows quadratically as $N_{\text{inter}}C_2$, which follows a growth function of $O(N^2)$. For instance, while the AGD20K dataset features 36 interactions, necessitating determination of approximately 630 pairs, the HICO-DET [11] dataset entails 6,786 pairs for 117 interactions, and 79,800 pairs for the SWiG-HOI [56, 65] with 400 interactions. Therefore, the LLM-generated interaction relationship map is essential for the dataset scalability of the INTRA framework and the consistency of the pair classification. To demonstrate scalability of our method and the crucial role of LLM generated interaction-relationship map for dataset scalability, we trained our INTRA model with HICO-DET dataset. Since egocentric images or GT for quantitative evaluation are unavailable, we evaluated our trained network on crawled egocentric images containing interactions not present in AGD20K, but can be found in HICO-DET. As shown in Fig. S2, our network exhibits scalability for larger datasets and effectively identifies interactions such as 'adjust', grounding the knot of a tie, and 'carry', grounding the armrest and back of the
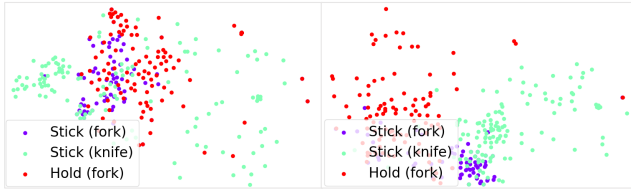
(a) t-SNE before (left) / after (right) training (Stick, Hold)
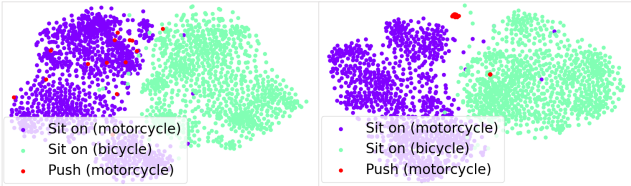
(b) t-SNE before (left) / after (right) training (Sit on, Push)

Fig. S3: Feature distribution comparison before and after training was conducted using t-SNE. For t-SNE before training, the DINOv2 class token was utilized. For t-SNE after training, f_{exo}, which is simply the weighted sum of DINOv2 features, was used. As depicted in figure, features of interaction images implying different affordances (*i.e.*, 'stick' and 'hold' fork or 'push' and 'sit on' bicycle) become separate as a result of interaction relationship-guided contrastive loss, while features of images containing similar object features still remain close to each other.

chair. If our method was not based on LLM-generated interaction-relationship map, we had to manually annotate 6,786 interaction pairs to generate the map.

Effectiveness of interaction relationship-guided contrastive loss. As seen in Fig. S3 (a) before training, DINOv2 features of 'stick (fork)', 'stick (knife)', and 'hold (fork)' overlap, as 'stick' usually comes with 'hold'. However, since 'stick' and 'hold' fork imply different affordances, their features should be separated in the training. As training progresses, features of 'stick' and 'hold' forks being well separated, while 'stick' forks and 'stick' knife exhibit close feature distribution due to the closeness in their interactions, showing the effectiveness of interaction relationship-guided contrastive learning. Meanwhile, the object features contained in the images are well preserved, resulting in f_{exo} of 'stick' and 'hold' fork still remain close to each other. The same tendency is also depicted in Fig. S3 (b), where 'sit on' motorcycle becomes separate from 'push' motorcycle after training.

Table S10: Quantitative results on the new version of AGD20K for our method and LOCATE [29]. INTRA still outperformed LOCATE on all metrics.

	Seen			Unseen		
	mKLD↓	mSIM↑	mNSS↑	mKLD↓	mSIM↑	mNSS↑
LOCATE [29]	1.277	0.389	1.370	1.410	0.358	1.372
INTRA (Ours)	**1.209**	**0.443**	**1.450**	**1.388**	**0.385**	**1.384**

E   Additional Results

E.1   Additional results on extended AGD20K testset

In the main paper, we compared the performance using the existing AGD20K test set to maintain consistency with the experimental settings of previous works. However, we also compared the performance with LOCATE [29] using the recently extended AGD20K test set. Tab S10 shows the quantitative comparison results of our method with LOCATE [29]. The results show that INTRA still outperformed in all metrics.

E.2   Affordance map visualization on exocentric images

To verify that our proposed method, INTRA, focuses on the interaction-relevant object features in exocentric images during training, we conducted affordance map visualization on exocentric images. As illustrated in Fig. S4, even in complex scenes involving people interacting, it is evident that only the portion of the object engaged in interaction with people is grounded.

E.3   Visualization of affordance grounding on the same objects with different interactions

Just as humans can interact with different parts of same objects for distinct interactions, it's crucial to accurately discern affordance grounding based on these interactions for further applications in real situations. We achieve this through interaction-guided contrastive learning and present various results in this section. As shown in Fig. S5, previous approaches struggle to accurately identify the object part that corresponds to the interaction, wheares INTRA (Ours) successfully grounds the relevant object part. We conducted the affordance grounding predictions for pairs such as 'drink with' and 'pour', 'open' and 'pour', 'ride' and 'sit on', and 'cut with' and 'stick', comparing the performance of previous methods with INTRA (Ours).

E.4   Additional qualitative results on AGD20K dataset

In Fig. S6, Fig. S7, Fig. S8 and Fig. S9, we present additional qualitative results and compare them with state-of-the art methods [29, 35, 36]. In both the

Table S11: Comparison to previous arts on different object scales. Results of * are taken from [29]. The test set is divided to 'Big', 'Medium' and 'Small' based on the ratio of the mask to the whole image. INTRA (Ours) outperforms other baselines [18, 29,35,36,40,46,53] significantly in the 'Small' subsets, while demonstrating competitive or superior performance in the 'Medium' and 'Big' subsets compared to the baselines.

	Method	Big			Medium			Small		
		KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
Seen	EIL* [40]	1.047	0.461	0.389	1.794	0.284	0.710	3.057	0.123	0.231
	SPA* [53]	5.745	0.317	0.222	4.990	0.228	0.440	6.076	0.118	0.297
	TS-CAM* [18]	1.039	0.424	0.166	1.814	0.248	0.401	2.652	0.132	0.352
	Hotspots* [46]	0.986	0.448	0.408	1.738	0.265	0.672	2.587	0.149	0.683
	Cross-view-AG* [36]	0.766	0.533	0.652	1.485	0.322	1.040	2.373	0.175	0.927
	Cross-view-AG+* [35]	0.787	0.521	0.660	1.481	0.314	1.089	2.381	0.167	0.959
	LOCATE* [29]	0.676	0.580	0.706	1.178	0.390	1.316	2.029	0.216	1.349
	INTRA (Ours)	0.695	0.579	0.782	1.193	0.394	1.300	1.826	0.239	1.587
Unseen	EIL* [40]	1.199	0.393	0.271	1.906	0.246	0.482	3.082	0.113	0.116
	SPA* [53]	8.299	0.259	0.254	6.938	0.186	0.333	7.784	0.095	0.144
	TS-CAM* [18]	1.238	0.351	0.072	1.970	0.208	0.236	2.766	0.113	0.124
	Hotspots* [46]	1.015	0.425	0.548	1.872	0.242	0.605	2.693	0.134	0.544
	Cross-view-AG* [36]	0.884	0.500	0.728	1.595	0.303	0.945	2.558	0.147	0.692
	Cross-view-AG+* [35]	0.867	0.485	0.776	1.658	0.279	0.988	2.630	0.133	0.754
	LOCATE* [29]	0.571	0.629	0.956	1.302	0.373	1.257	2.223	0.189	1.071
	INTRA (Ours)	0.662	0.573	0.955	1.288	0.378	1.249	2.032	0.230	1.299

'Seen' and 'Unseen' setting, baselines struggles to accurately ground affordance when multiple objects are present in the images. Additionally, their heatmaps are also coarse and do not seem to handle occlusion that occurs in exocentric images. In contrast, even in complex scenes, INTRA (Ours) successfully grounds affordances.

E.5 Visualization of feasibility study on generalization property of INTRA

Domain gap. Our framework, INTRA, demonstrates excellent grounding results when there is a significant domain gap between training and inference. We conducted inference using egocentric images from various domains to assess the robustness against this gap. We obtained synthetic images of objects present in the train set from the internet, including pen illustrations, pixel art, or object images from instructions. Compared to other baseline model [29], our model accurately grounds affordances in those images. Especially, we can see in a picture of drum and chair of Fig. S10 that our model successfully identifies all interaction-relevant regions of objects. Moreover, in images of pen-illustrated chair and camera, our model generates finer heatmaps.

Novel object. To analyze the generalization property of the affordance grounding, we conducted experiments on novel objects, which were not in the train set.

We tested our model and baseline [29] on novel objects that share common properties with objects in the train set, but have never been seen before. IN-TRA (Ours) successfully grounds the most interaction-relevant object parts, as demonstrated in examples of 'wallet' and 'door' in Fig. S11.

Novel interaction. With VLM text embedding, INTRA (Ours) successfully predicts affordance maps for novel interactions, as illustrated in our main paper. We designed and conducted experiments to analyze the generalization ability adopted by the VLM text encoder. During training, we excluded specific interaction classes and inferred affordance grounding on these excluded interaction classes to assess the grounding performance of the unlearned interactions. Fig. S12 is qualitative results of affordance grounding when the 'hold', 'pour', and 'kick' classes are unseen during the training.

Novel object and novel interaction. We conducted experiments to assess whether our model can infer affordance even in cases where both the interaction and object are novel. Since the baseline model can only infer predetermined interaction types, for comparison, we asked LLM to find the closest interaction of the novel interaction and inferred affordance grounding on the closest interaction in the baseline model. Specifically, 'brew-pour', 'wipe-brush with', 'pull-hold', 'drive-ride' were selected to substitute novel interaction in the baseline model. In Fig. S13, it is evident that our INTRA can accurately ground affordance even when both the interaction and object are unseen, and the novel objects have many tractable parts. For instance, Ours grounded all relevant parts of 'Drive' in photos of car interior, whereas the baseline could not. Also, Ours accurately grounded all relevant parts that can be pulled in a wine opener and wagon.

Comparison on different scales. To investigate how varying affordance region scales impact the model, we follow [29, 35, 36], dividing the test set into three subsets: 'Big', 'Medium' and 'Small'. These subsets are determined by the ratio of the mask, with thresholds set at more than 0.1, between 0.03 and 0.1, and less than 0.03, respectively. INTRA (Ours) outperforms other baselines [18, 29, 35, 36, 40, 46, 53] significantly in the 'Small' subset, demonstrating its ability to generate finer heatmaps. Additionally, performance in the 'Medium' and 'Big' subsets is either comparable to or surpasses the baselines.

F   Limitation

While INTRA produces good grounding results even with novel interactions or objects, it finds it challenging to learn interactions in exocentric images where there is no contact between the object and the human. Although such interactions do not exist in AGD20K, it has been observed that classes without direct contact, such as 'watch (clock)' or 'direct (airplane),' are difficult to learn when training with the HOI datasets [11, 24, 56, 65]. Thus, inferrring instructions that

contain implicit meanings can be challenging. For example, grounding is possible for 'turn on (air-conditioner),' but not for abstract instructions like 'I'm cold.' Additionally, grounding may fail when the points of interaction are visually similar. For instance, when operating a microwave, it is difficult to identify the button with the affordance for 'heat' among multiple buttons.
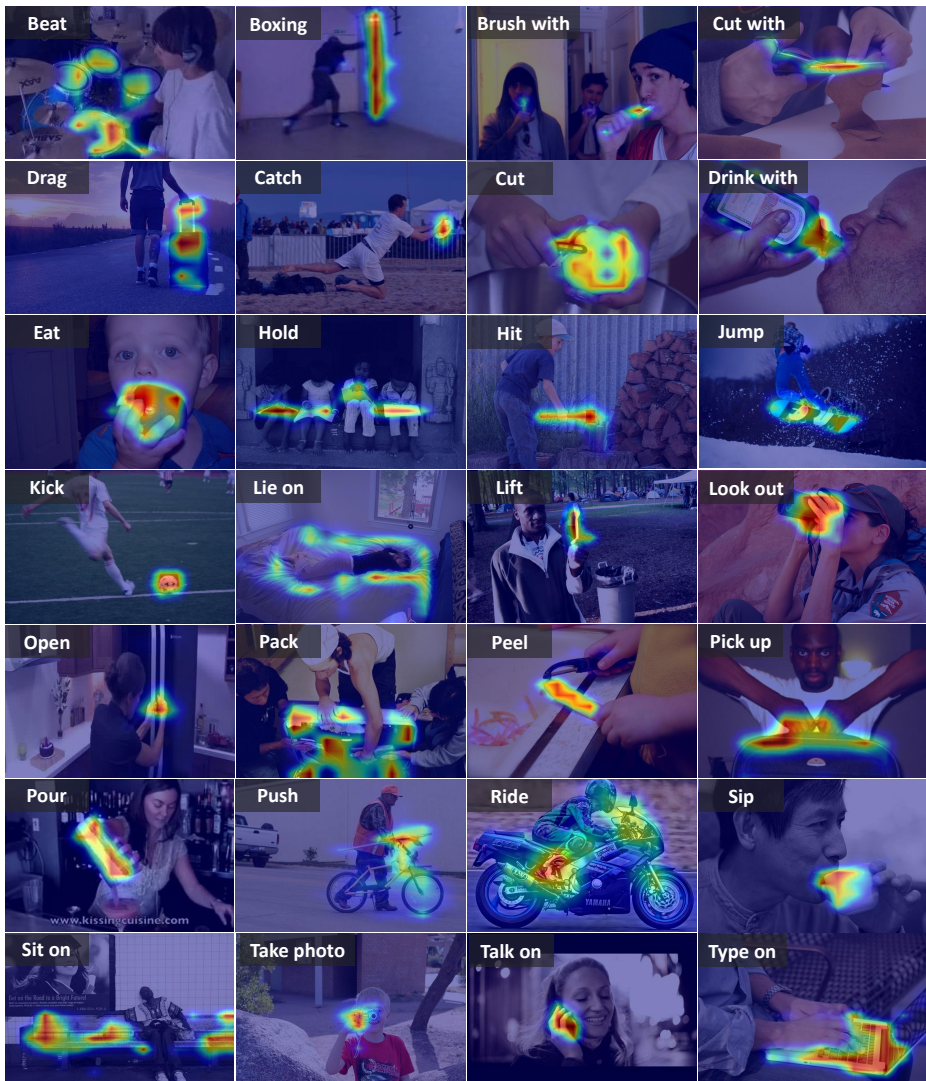
Fig. S4: Affordance map visualization results on exocentric images. The model precisely localizes to the interacting part of the object rather than the person overall. Accurately pinpointing the part of the object involved in interaction is important, especially when occlusion occurs. Our model handles this well, as shown in instances such as 'hit', 'look out', 'open', 'push', 'ride', and 'talk on'.
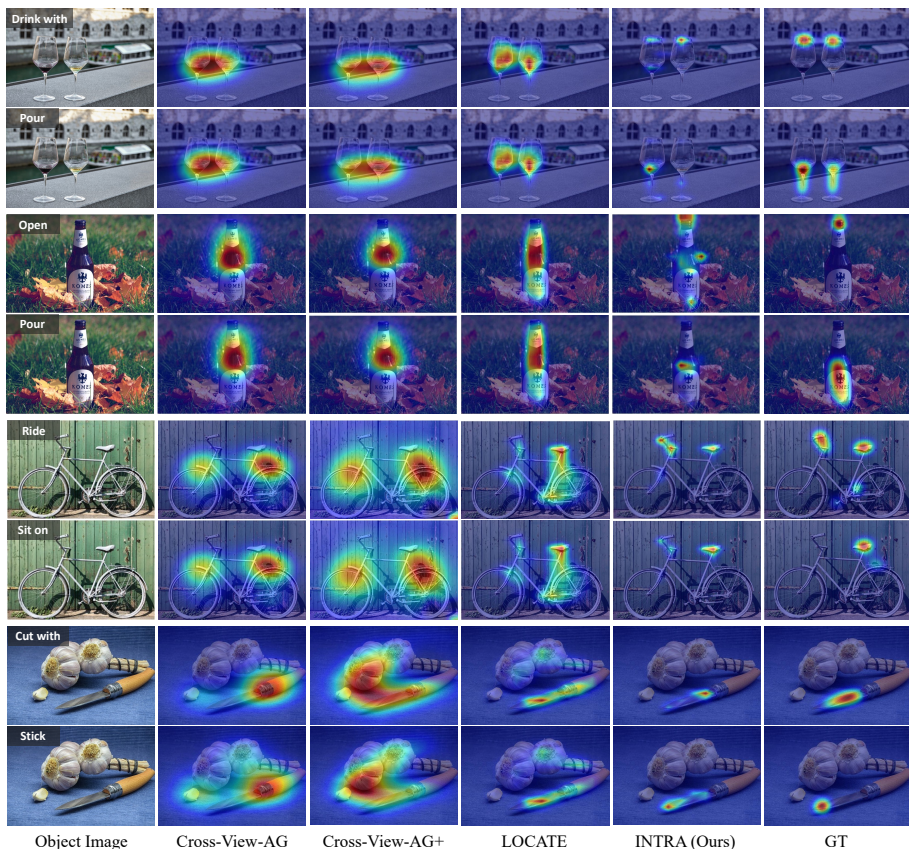
Fig. S5: Visualization of affordance grounding on the same objects with different interactions compared to results of previous arts [29, 35, 36]. We performed affordance grounding for 'wine glass', 'bottle', 'bicycle', and 'knife', where two interactions require different parts of objects to be grounded. The model grounded the rim of 'wine glass' for 'drink with' and the handle for 'pour'. For 'bottle' and 'bicycle', INTRA (Ours) precisely identifies the object parts corresponding to the given interactions. Particularly, for 'knife', we observe that INTRA (Ours) grounds the blade for 'cut with' and the tip of the knife for 'stick'.
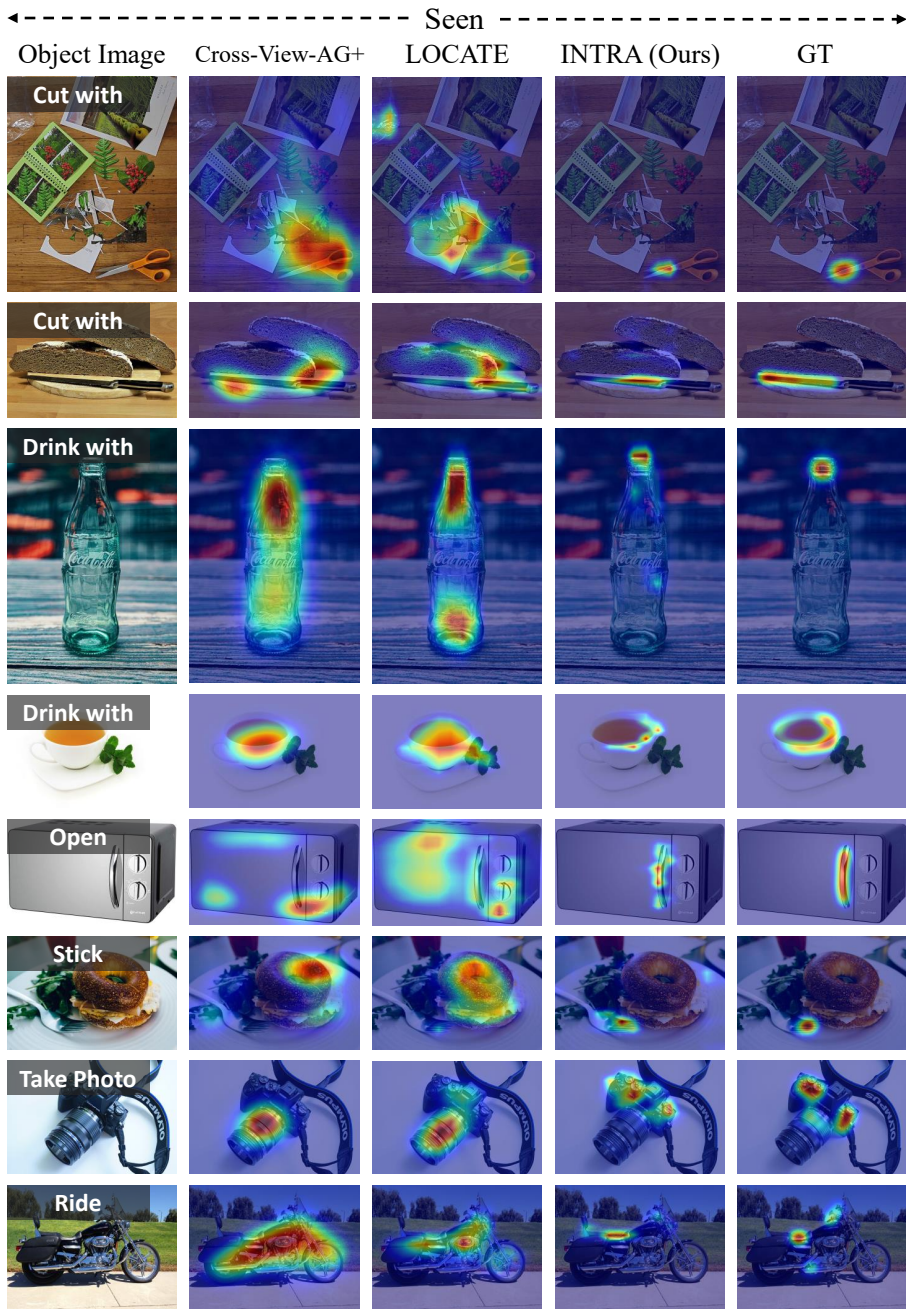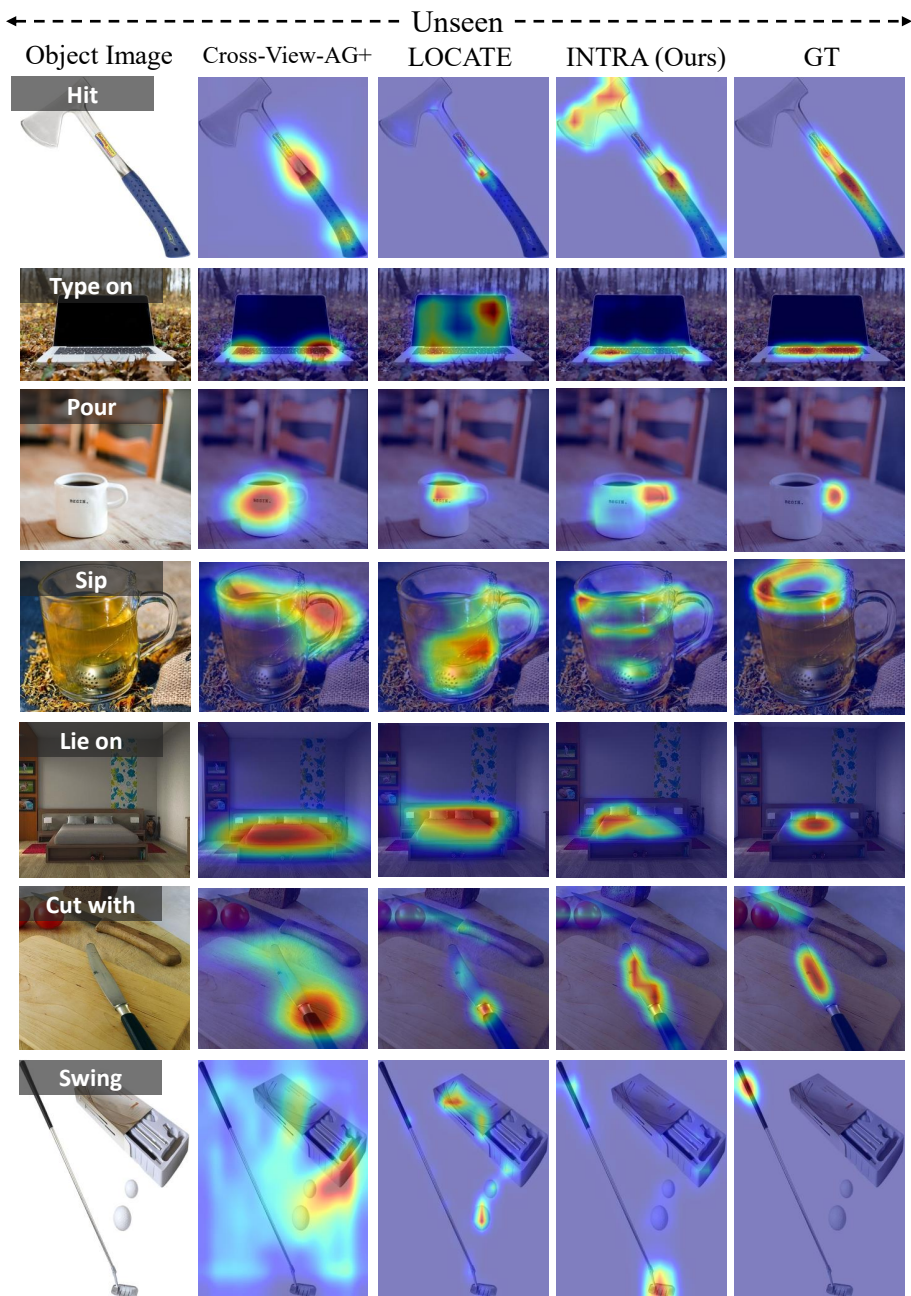
Fig. S6: Additional qualitative results comparison between INTRA (Ours) and other baselines [29, 35, 36] on 'Seen' testset of AGD20K. INTRA (Ours) grounds affordance accurately when the images are clustered with many objects. For example, grounding results of 'cut with' and 'stick' accurately mark the blade of scissors, knife and the tip of fork. Unlike other baselines that tend to generate coarse heatmaps, our heatmaps are fine and localize only the relevant parts of the interactions.
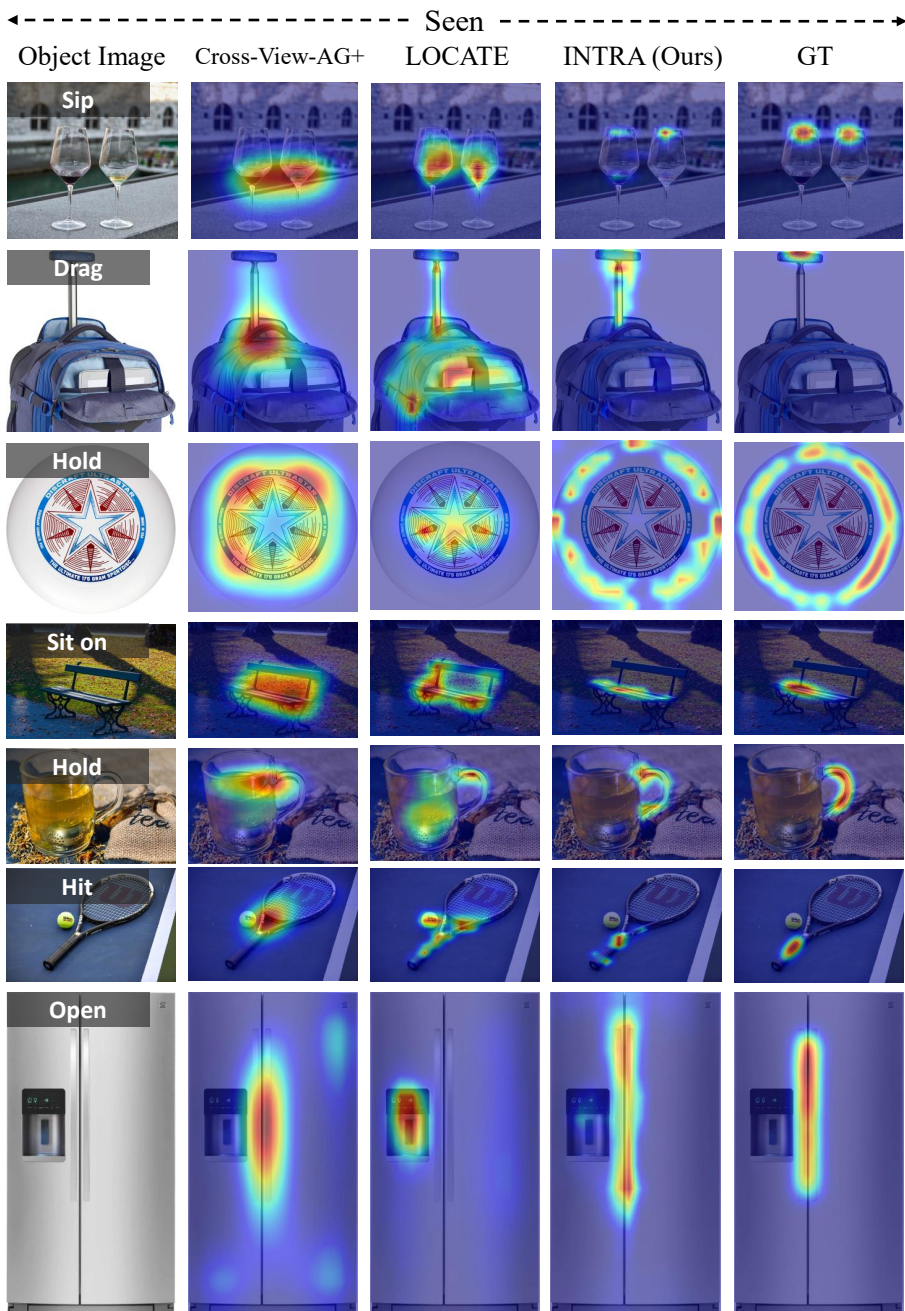
Fig. S7: Additional qualitative results comparison between INTRA (Ours) and other baselines [29, 35, 36] on 'Seen' testset of AGD20K. INTRA (Ours) grounds affordance accurately and generates finer heatmaps. For example, grounding results of 'drag' and 'open' accurately mark the handle of suitcase, and the door handle of the refrigerator.
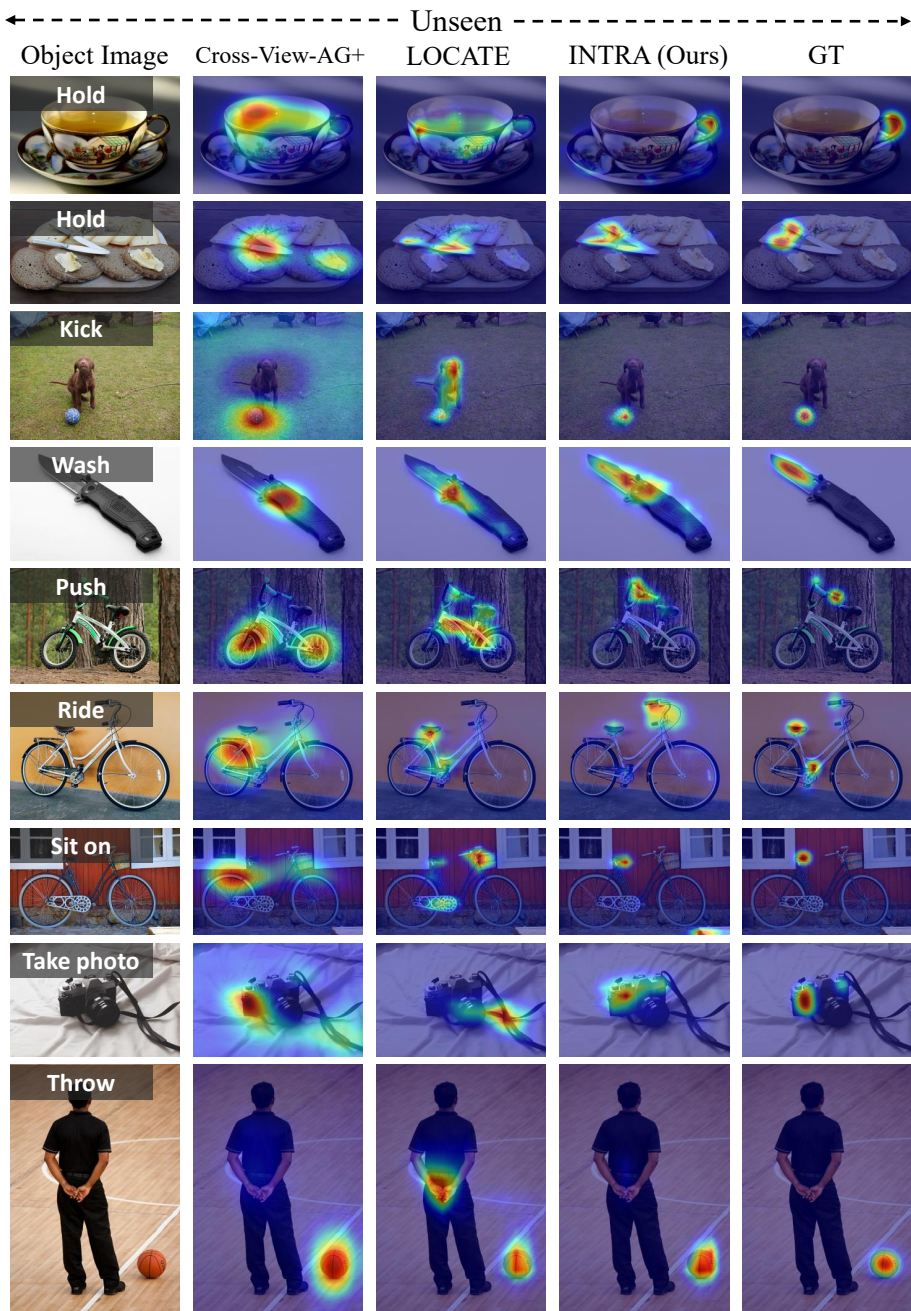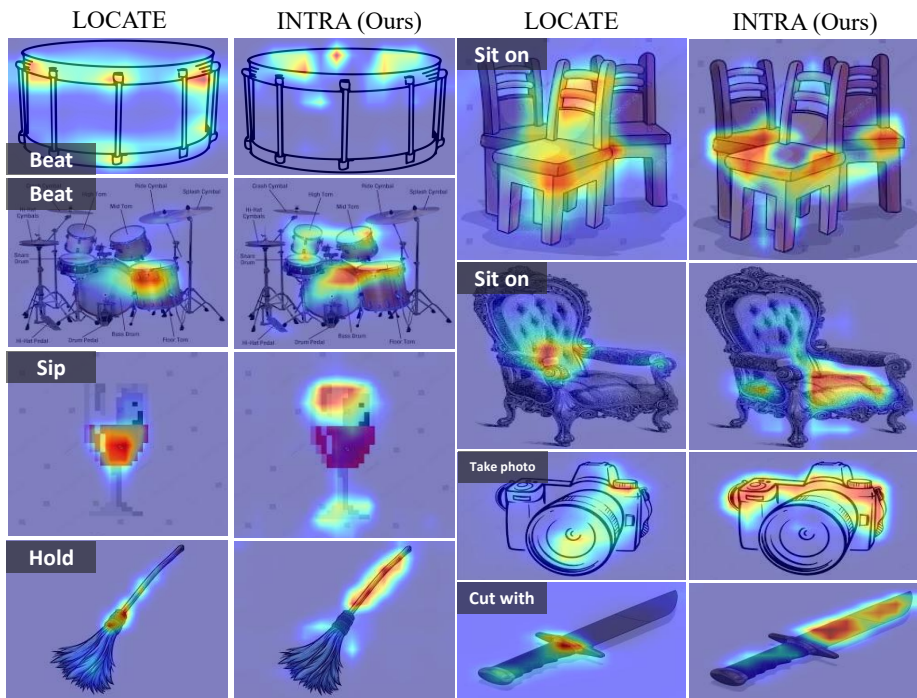
Fig. S8: Additional qualitative results comparison between INTRA (Ours) and other baselines [29,35,36] on 'Unseen' testset of AGD20K. Qualitative comparison of grounding results in 'Unseen' setting also shows that accuracy of our grounding results outperforms others. Especially, grounding result of 'sit on' not only localizes bicycle saddle, but also the wooden chair that is partially shown in the image.
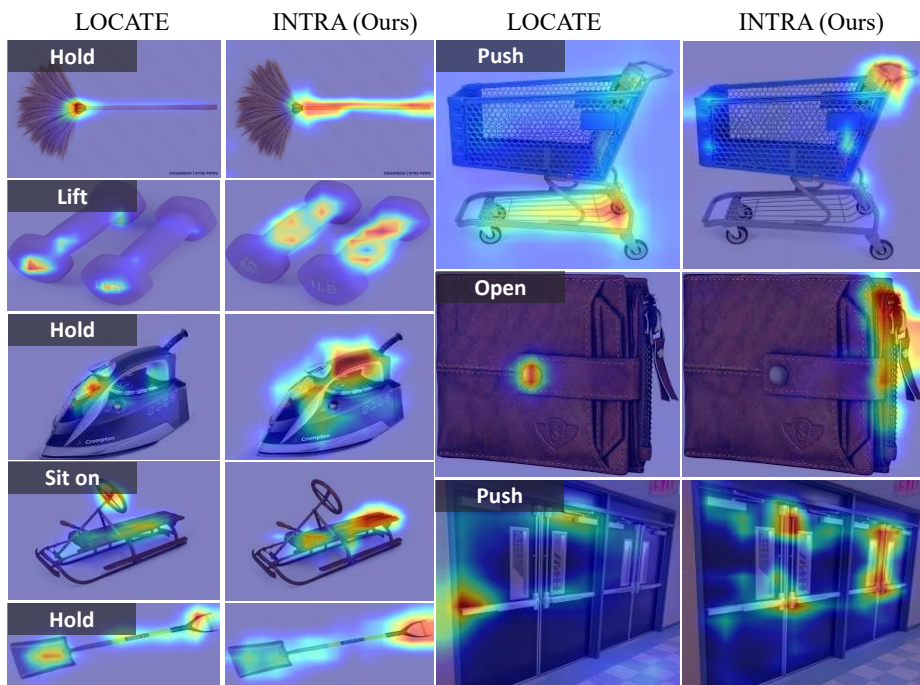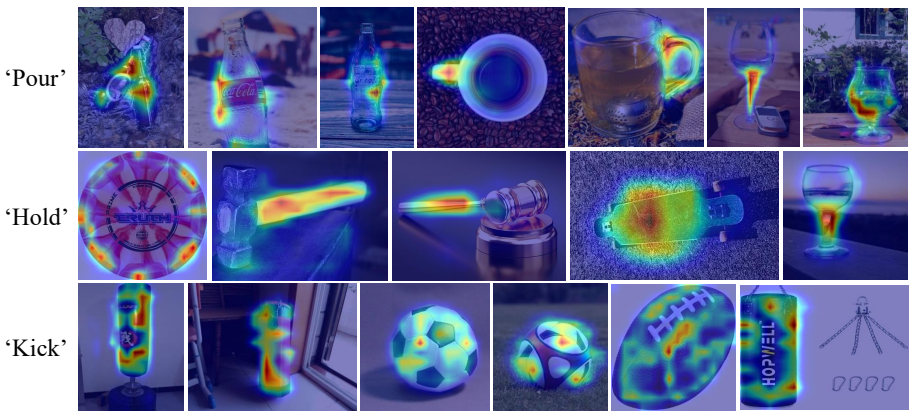
Fig. S9: Additional qualitative results comparsion between INTRA (Ours) and other baselines [29,35,36] on 'Unseen' testset of AGD20K. Qualitative comparison of grounding results in 'Unseen' setting also shows that accuracy of our grounding results outperforms others. Especially, the grounding result of 'hit' not only localizes the handle of the axe, but also includes the blade of the axe, which can be interpreted as an integral part of the object incorporated with 'hit'.
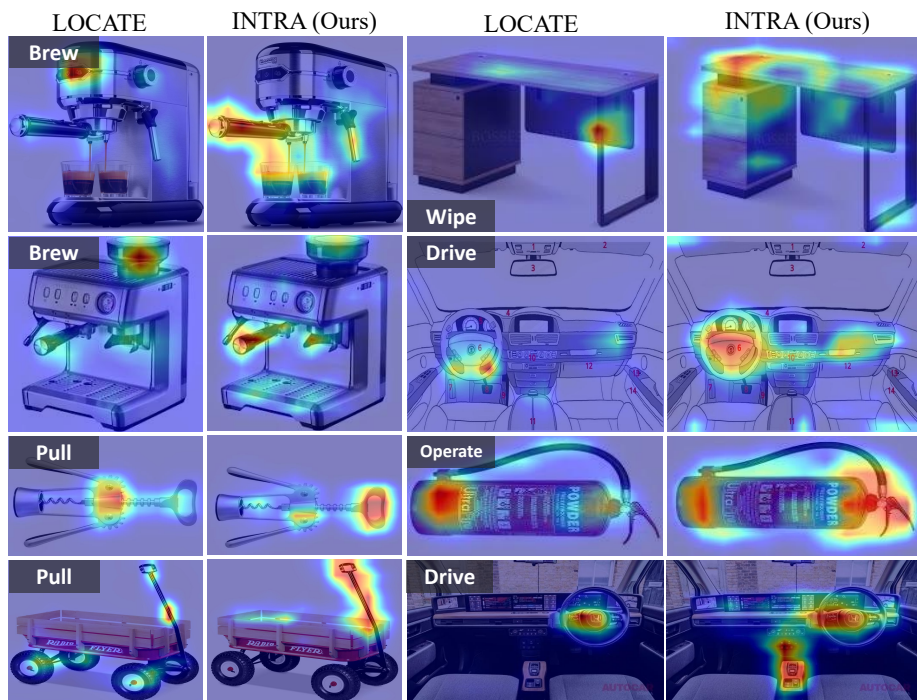
Fig. S10: Additional qualitative results comparison between INTRA (Ours) and other baseline [29] on affordance grounding in object images with significant domain gap. IN-TRA (Ours) accurately and finely grounds affordances in pen-illustrated chairs, cameras and pixel-art wine glasses while other baseline can't. Also in case of 'beat (drum)', while baseline model is inaccurately grounding the side of the drum, our model grounds center, top side of drum accurately. In case of drum set, our model grounded all drums that we can 'beat' while the baseline grounded on the side of base drum. The examples of 'hold (broomstick)' and 'cut with (knife)' shows that although there were significant domain gap between training images, INTRA (Ours) grounds the parts that are incorporated by interactions accurately.

| LOCATE | INTRA (Ours) | LOCATE | INTRA (Ours) |

Fig. S11: Additional qualitative results comparison between INTRA (Ours) and other baseline [29] on affordance grounding in novel objects. Our INTRA, as seen in examples like 'open (wallet)' or 'push (door)', accurately grounds more important interaction points such as center of doors or zipper of the wallet. Also, for example of 'hold (shovel)', 'hold (iron)' and 'push (shopping cart)', it accurately captures the exact interaction points which are handle of the object. Although the train set does not contain images of weights or iron, INTRA (Ours) successfully grounds the interaction points.

Fig. S12: Additional Qualitative results of affordance grounding when interactions are unseen. The affordance grounding output alongside each interaction demonstrates the inference results for interactions that were not part of the training data. For example, in case of 'pour', all the exocentric images related to 'pour' were excluded during training, yet our model still exhibits fine grounding quality when inferring 'pour' on 'bottle', 'cup', and 'wine glass'.

Fig. S13: Additional qualitative results comparison between INTRA (Ours) and other baseline [29] on affordance grounding when both the interaction and object are unseen during the training. Our approach accurately grounds affordances even when objects have many tractable parts, as observed in cases such as 'coffee machine', 'wine opener' or 'car interior'. For instance, in the case of 'brew', INTRA (Ours) captures the handle of the portafilter, while LOCATE [29] focuses on the bean container. Similarly, for the example of 'wipe', INTRA grounds on the flat part of the desk, wheares LOCATE focuses on the desk leg.