Strategies to Leverage Foundational Model Knowledge in Object Affordance Grounding

Arushi Rai[1]*, Kyle Buettner[2]*, Adriana Kovashka[1,2]
[1]Department of Computer Science, [2]Intelligent Systems Program, University of Pittsburgh, PA, USA
{arr159, buettnerk}@pitt.edu, kovashka@cs.pitt.edu

Abstract

An important task for intelligent systems is affordance grounding, where the goal is to locate regions on an object where an action can be performed. Past weakly supervised approaches learn from human-object interaction (HOI) by transferring grounding knowledge from exocentric to egocentric views of an object. The use of HOI priors is inherently noisy and thus provides a limited source of supervision. To address this challenge, we identify that recent foundational models (i.e. VLMs and LLMs) can serve as auxiliary sources of knowledge for frameworks due to their vast world knowledge. In this work, we propose strategies to extract and leverage foundational model knowledge related to attributes and object parts to enhance an HOI-based affordance grounding framework. In particular, we propose to combine HOI and foundational model priors through (1) a spatial consistency loss and (2) heatmap aggregation. Our strategies result in mKLD and mNSS improvements, and insights suggest future directions for improving affordance grounding capabilities.

1. Introduction

Robust object understanding is an important goal for computer vision systems and is primarily evaluated using tasks like image recognition, object detection, and semantic segmentation. A higher-level problem that receives less attention is **affordance understanding**, which involves determining the actions afforded by an object or object part and localizing specific regions in an image where the action can be performed. For example, a *knife blade* is said to afford *cutting* and a *knife handle* affords *holding*. Affordance understanding is a hallmark of human intelligence, as a person can naturally recognize how to use objects through their appearance. Such understanding has essential implications as AI moves towards embodied interaction, where an agent in an environment needs similar capabilities to use objects.

*These authors contributed equally.

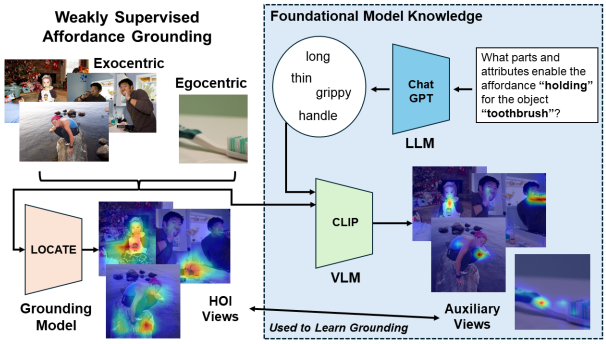

Figure 1. **Weakly supervised affordance grounding with auxiliary foundational model knowledge.** Our work's premise is that foundational model knowledge (attribute/part signals from VLMs/LLMs) can address noise in weak supervision. We particularly address noise in LOCATE [16], a framework that uses human-object interaction (exocentric views) to guide the learning of grounding in egocentric views. We add foundational model views to further enhance the learning of grounding.

A specific subtask is *affordance grounding*, where given an action label, a model should produce probabilities (i.e. heatmaps) over regions in an image to represent where an action can be performed on an object. While collecting densely annotated grounding masks for training is possible, such data acquisition is expensive. Common approaches have thus been weakly supervised, specifically leveraging human-object interaction (HOI) priors from widely available exocentric views (i.e. images of objects in action from a third-person viewpoint) to learn to ground in egocentric views (i.e. images of objects as if from one own's viewpoint) [16, 20]. Grounding is inferred from the point(s) of contact between a human and an object (*e.g.* where the human holds the tennis racket). This process can be notably noisy due to occlusion from hands and imperfect attention to affordance-specific features. In addition, grounding models tend to struggle to generalize to unseen objects.

To address these challenges, we hypothesize that the combination of vision-language models (VLMs) and large-

language models (LLMs) can provide affordance grounding support due to the broad world knowledge that the models capture. Motivated by past work which shows the value of LLMs to generate useful affordance knowledge (of key attributes/parts) for task-driven object detection [34], we aim to use *both* VLMs and LLMs to generate *auxiliary affordance views* (e.g. heatmap masks) for the grounding task. Our idea to generate views is shown in Fig. 1, where knowledge of the attributes and parts that give rise to affordances (*e.g. thin, grippy, long handle* for *holding*) is probed from ChatGPT (*gpt-3.5-turbo* [25]). This information is combined with the zero-shot capability of CLIP [28] to generate auxiliary heatmap masks.

We then test whether foundational model knowledge and HOI recognition knowledge are complementary by integrating the auxiliary affordance views with LOCATE [16], the current state-of-the-art affordance grounding framework. We propose two strategies to improve affordance grounding: (1) a *spatial consistency loss* to encourage egocentric views to consider CLIP's knowledge and (2) *heatmap aggregation* to guide exocentric combination of HOI and foundational model knowledge. Motivated by the need to address different types of noise, we empirically evaluate union-based and intersection-based forms of aggregation. Using a union-based strategy with a spatial consistency loss, we achieve an improvement of +2.7% in mKLD and +1.9% in mNSS in the seen setting vs. LOCATE.

To summarize the contributions of this work:

- We provide a baseline evaluation of CLIP's zero-shot affordance grounding capability using a variety of prompts (*e.g.* with parts, attributes, the affordance/action). LLM knowledge leads to improvements versus using default affordance baseline prompts.
- We propose a unique mechanism to leverage LLM and VLM knowledge in the form of auxiliary affordance grounding masks.
- We show that use of a spatial consistency loss between egocentric views from LOCATE and CLIP groundings is effective especially for improving mKLD.
- We show that aggregating HOI-based and VLM-based supervision in exocentric views is most effective in a union-based strategy, especially in tandem with the spatial loss.

2. Background and Related Work

Visual affordance learning. The goal of visual affordance understanding is to infer details about the actions that can be performed with objects in a visual input that represents the current state of the environment. Affordance understanding has motivation in robotics, as an agent needs to determine the set of actions that an environment allows. There is a close relationship to the functionality of an object, though object function is considered an immutable property of an object while affordances depend on the existing state of the environment [11] and can capture social aspects [8]. There is similarity to human-object interaction, as affordances represent what interactions can take place [12]. Affordance datasets [9, 10, 20, 21, 24, 30] have been proposed at various granularities (e.g. masks, image labels, and bounding boxes) to cover tasks such as segmentation/grounding, categorization, and detection. Affordance understanding is also involved in higher-level reasoning tasks such as task-driven object detection [32], *e.g.* "open a beer bottle". Our work is in weakly supervised affordance grounding, in particular.

Affordance grounding. Our main task of interest, *affordance grounding*, involves producing pixel-level heatmaps representing a set of action possiblities for a given object and its affordance. While it is possible to learn dense predictions with supervised masks, a goal has been to use weak supervision or self-supervision to avoid high annotation costs associated with collecting boxes or masks [31]. Another goal has been to encourage models to generalize affordance grounding from "seen" to "unseen" objects [20]. Many recent approaches opt to use human-object interactions to guide affordance grounding as these signals demonstrate where an action can take place. For example, [21] captures interactive affinity cues (i.e. contact regions) between a human and an object in images of interactions to learn to produce groundings in images with just objects. [7] alternatively uses affordances in video demonstrations, along with a novel transformer architecture and self-supervised pretraining strategy, to scalably learn grounding. [16] proposes an unsupervised part selection method to extract affordance cues from multiple exocentric views and guide weakly supervised learning of grounding in egocentric images. Across each of these methods, the potential for noise is high, due to issues like occlusion and high diversity in HOI demonstrations. *Our work hypothesizes that foundational model knowledge can address this noise.* We specifically propose to leverage the world knowledge of LLMs and VLMs, in the form of parts and attribute signals which have an important relationship with affordances [18]. We evaluate the effectiveness of this knowledge by building on the exocentric-egocentric affordance grounding knowledge transfer framework of [16].

Knowledge probing of foundational models. Benefiting from a large amount of text and paired image-caption data on the Internet, there has been a rise in large foundational models capable of zero-shot and/or few-shot transfer to various unimodal and multimodal tasks. For computer vision applications, vision-language models (VLMs) have become especially popular with notable zero-shot classification success demonstrated by CLIP [28], ALIGN [15], CoCa [36], and LiT [37]. Such VLMs have traditionally been used in tasks like classification, detection, and segmentation. *Various facets of tasks like affordance understanding remain to be studied*. Our study addresses this gap through presenting

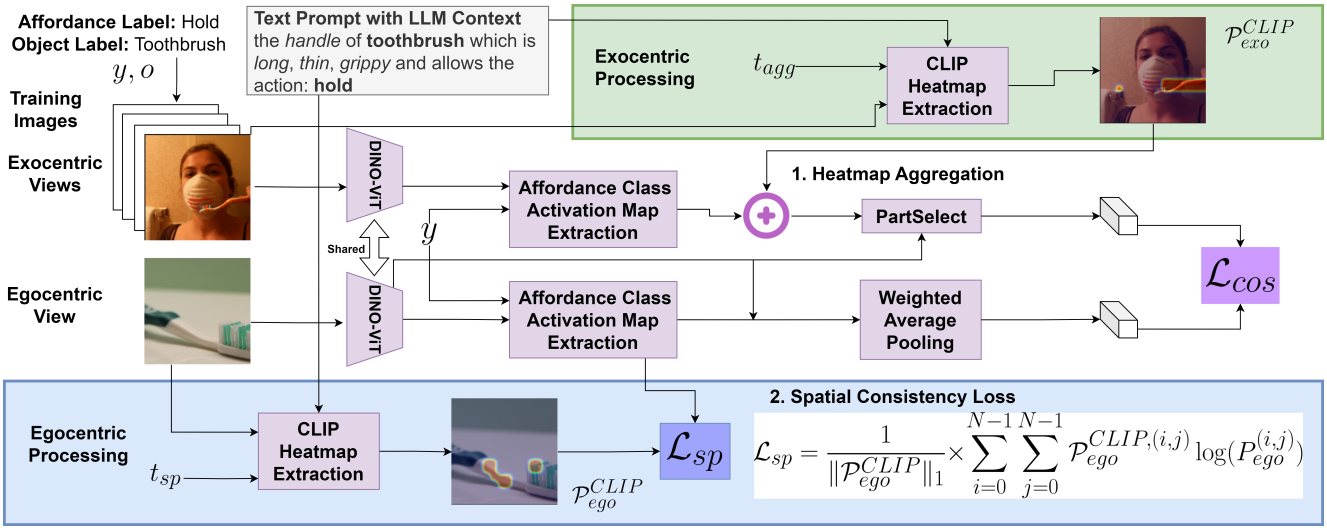

Figure 2. **Overview of our method to integrate auxiliary affordance views (heatmap masks produced with foundational model knowledge), within LOCATE, an existing affordance grounding framework**. An affordance label, object label, and LLM context of attributes/parts serve as components in a prompt used as input to the interpretability method of [6] to gather CLIP heatmaps. We follow LOCATE's exocentric-egocentric knowledge transfer and produce auxiliary views in both exocentric and egocentric processing ($\mathcal{P}_{exo}^{CLIP}$/$\mathcal{P}_{ego}^{CLIP}$). We particularly propose (1) heatmap aggregation and (2) a spatial consistency loss ($\mathcal{L}_{sp}$), respectively. Shown are relevant thresholds needed for mask creation (t_{agg}/t_{sp}) and LOCATE's own cosine embedding loss ($\mathcal{L}_{cos}$); see [16] for more details.

insights into various prompting mechanisms for affordance grounding. To gather prompt details, we use LLMs, which have shown impressive world and commonsense knowledge [1, 13, 22]. Further research has shown that LLM knowledge of objects, in the form of attribute, part, and material context, can be combined with VLMs (*e.g.* CLIP) to improve zero-shot recognition [23, 27]. With respect to affordances, knowledge from LLMs has been used to condition task-driven object detection [34], but not directly for the grounding task. We specifically explore grounding and analyze strategies to integrate auxiliary knowledge.

Explicitly leveraging attribute signals in learning. Recent works have incorporated explicit mechanisms to leverage attributes in the text of captions and scene graphs and improve in object-based tasks. [35] imposes a loss between hierarchical relationships in captions to improve visual grounding. For detection, [14] directly entangles attribute-object learning with textual scene graphs, and [2] enhances attention to attribute meaning in contrastive learning through sampling of adjective-perturbed negative captions. Our mechanism is unique in that we use LLM and VLM attribute and part knowledge in a spatial consistency loss to improve affordance grounding.

3. Approach

The main problem that we identify is that existing weakly supervised pipelines for learning affordance grounding have a limited and noisy source of supervision with human-object interaction (HOI) priors. Our overall idea is to probe key attribute and part information about affordances using an LLM (ChatGPT [25]) and then produce auxiliary heatmaps through integrating such knowledge with CLIP [28]. These heatmaps serve as *auxiliary views* (i.e. supervision) for the LOCATE framework [16], for which we explore various strategies to best integrate knowledge (i.e. a spatial consistency loss, heatmap aggregation). Fig. 2 provides an overall view of our experimental approach.

3.1. Preliminaries: LOCATE

Our work builds on top of LOCATE [16], the current state-of-the-art for weakly supervised affordance grounding. In this approach, the goal is to learn grounding for a set of affordances $\mathcal{A}$ of size C. The approach follows an exocentric-to-egocentric knowledge transfer methodology, where human-object interaction (i.e. exocentric) images guide the learning of affordance grounding in egocentric images (i.e. images where objects are depicted alone *without* interaction). Formally, given an affordance label y and object label o for an egocentric image x^{ego}, k exocentric object-affordance images $x_0^{exo}, ..., x_k^{exo}$ are sampled to supervise egocentric grounding in training. Specifically, the contact points in human-object interaction serve as weak localization for affordances. A pretrained DINO-ViT [4] is used to extract deep feature descriptors for all images. The features for the exocentric images get concatenated together as f_{exo}, while f_{ego} represents features for the egocentric image. A class-activation map layer (a projection

Affordance	Object	Attributes	Part
hold	tennis racket	long, grippy, straight	handle
	knife	long, grip, curved	handle
	baseball bat	long, grip, thin	handle
ride	bicycle	soft, saddle-shaped, comfortable	bicycle seat
beat	drum	flat, round, hard	drum surface
cut with	knife	sharp, pointed, metallic	blade
	scissors	sharp, pointy, curved	blades
look out	binoculars	clear, round, glass	lenses
hit	tennis racket	thin, stringy, crisscrossed	racket strings

Table 1. **Example knowledge gathered from an LLM (ChatGPT)**. Given an object and affordance input, the LLM produces attributes and parts that are relevant to the affordance. Observe the similarity between attributes for the same affordance, but different objects.

layer followed by two convolutions) is further trained to learn affordance-discriminative heatmaps (P_{exo} and P_{ego}). Due to the weak supervisory signal of HOI priors, these regions can be imprecise and noisy. LOCATE introduces a PartSelect module that leverages unsupervised k-means clustering to isolate object parts (f_{op}) from background and human clusters (unused) in the exocentric features for egocentric guidance. Overall, the framework's loss function is shown in Eq. 1:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda_{cos}\mathcal{L}_{cos} + \lambda_c\mathcal{L}_c \tag{1}$$

$\mathcal{L}_{cls}$ is a cross-entropy loss over the set of affordance classes, $\mathcal{L}_c$ is a loss to encourage concentrated connected components in grounding, and $\mathcal{L}_{cos}$ is a cosine embedding loss that is used to align egocentric features with the features of the parts collected with PartSelect.

Our approach to improve LOCATE is to provide *auxiliary views using foundational model knowledge*. These views are represented in the form of heatmaps. As such, our improvements primarily impact which parts are used to learn grounding (f_{op}), and consequently $\mathcal{L}_{cos}$. We therefore further outline this loss in Eq. 2, with α being a margin:

$$\mathcal{L}_{cos} = \max(1 - \frac{f_{op}\dot{f}_{ego}}{|f_{op}||f_{ego}|} - \alpha, 0) \tag{2}$$

We refer the reader to [16] for more specific details about the LOCATE framework.

3.2. Foundational Model Knowledge Probing

This section outlines our approach to gather foundational model knowledge for grounding. We first probe affordance information from an LLM, and then extract heatmaps from a VLM using various prompting strategies.

Gathering LLM attribute and part knowledge. Intuitively, the reason an object affords an action is because it possesses key properties to enable that action. For instance, for the affordance *cut with* to be in an image, there should be something *sharp*, which is often *metallic*, like a *blade*. [34] shows that such information can help in task-driven object detection, but we alternatively wish to see if it can address noise in weakly supervised affordance grounding. As such, the first part of our approach is to gather knowledge about which attributes and/or parts enable an affordance. We prompt ChatGPT (*gpt-3.5-turbo*) for object part and attribute information using the following template:

> Output a text string that answers: "What part of the object affords this action?". Answer the question with just the answer, concisely. The answer should be understandable by a ten year old. It is possible for the entire object to afford the action, so use the exact original object term in such case. On the line following your answer, please add 3-5 concise visual attributes/words that would allow a ten year old to point to that object part in an image.
> EX: What part of ping pong paddle affords hitting?
> A: "paddle head"
> Attributes: "flat", "round", "slim"
> Q: What part of {object} affords {affordance}?

Notably, we provide an example to stimulate the in-context learning ability of the language model. We ask for 3-5 attributes since multiple attributes may contribute to an affordance. In general, we find the outputs to be fairly reasonable. We show examples (for AGD20K [20]) in Table 1.

Combining affordance knowledge with CLIP. We leverage LLM knowledge in concert with the CLIP VLM by prompting CLIP and producing grounding maps. We test a variety of different prompt templates to thoroughly evaluate how to best extract CLIP's knowledge:

- "{affordance}"
- "{object}"
- "the region of {object} which allows the action {affordance}"
- "{part}"
- "{attributes}"
- "{part} of {object}"
- "{part} of {object} which is {attributes}"
- "{part} of {object} which allows the action: {affordance}"
- "{part} of {object} which is {attributes} and allows the action: {affordance}"

Our motivation for testing the various prompts is to gauge some naive ways of probing affordance knowledge (*e.g.* just using the affordance and object names) vs. more specified versions with attribute and part knowledge.

We specifically use the ViT-B/32 version of CLIP for inference. To produce grounding maps, we leverage the strategy of [6], which produces visual explanations (relevancy map outputs) by considering modality interactions through attention layers. Notably this extraction approach is competitive vs. gradient approaches like GradCAM [33].

3.3. Integrating Auxiliary Views with LOCATE

We leverage the LLM knowledge-based maps probed with CLIP as *auxiliary affordance views* within the LOCATE framework. We identify that these views may be able to help as auxiliary supervision for both exocentric and egocentric images. In this section, we outline the two main approaches we use to explore integration: a spatial consistency loss and heatmap aggregation.

Spatial consistency loss. We design a **spatial consistency loss** to improve the learning of egocentric affordance grounding. We reason that part and attribute knowledge can be effective as part of a direct constraint with egocentric images, for example, due to object properties like "sharp" and "blade" being clearly visible for objects like "knife". To construct this loss, we define $\mathcal{F}_{ego}^{CLIP}$ as the normalized CLIP relevance map extracted for an egocentric image by using the approach [6] with a prompt $text_{aff}$. We use "{part} of {object} which is {attributes} and allows the action: {affordance}" as $text_{aff}$ due to its high performance in the zero-shot setting. $\mathcal{P}_{ego}^{CLIP}$ is further defined as the relevance map thresholded by a spatial parameter t_{sp}, shown in Equation 3:

$$\mathcal{P}_{ego}^{CLIP} = \mathcal{F}_{ego}^{CLIP} \geq t_{sp} \tag{3}$$

As the LOCATE framework similarly produces a thresholded map $\mathcal{P}_{ego}$ of size $N \times N$, our goal is to enforce a level of consistency between these two maps. Specifically, we want $\mathcal{P}_{ego}$ to ground *at least* the same regions as $\mathcal{P}_{ego}^{CLIP}$, as $\mathcal{P}_{ego}^{CLIP}$ is precise, but can have low recall over the entire

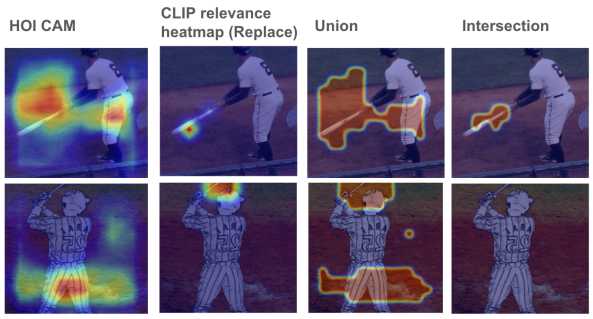

Figure 3. **Heatmap aggregation strategies of exploration.** Shown is the affordance "hit" with human-object interaction (HOI) class-activation maps as produced by [16] and CLIP relevance heatmaps computed with [6]. The CLIP heatmaps can be precise but often lack coverage of the full affordance region.

affordance region. For this goal, we compute binary cross-entropy (BCE) over each feature map element of the learned P_{ego} and auxiliary $\mathcal{P}_{ego}^{CLIP}$. We soften BCE to not penalize $\mathcal{P}_{ego}$ unfairly for correct predictions; thus we compute only the positive term of binary-cross entropy in Eq. 4:

$$\mathcal{L}_{sp} = -\frac{1}{\|\mathcal{P}_{ego}^{CLIP}\|_1} \times \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} \mathcal{P}_{ego}^{CLIP,(i,j)} \log(P_{ego}^{(i,j)}) \tag{4}$$

This loss serves as an auxiliary constraint in LOCATE to ensure the parts and attributes get adequate attention in egocentric grounding.

Heatmap aggregation. Exocentric-to-egocentric knowledge transfer is an important source of supervision in affordance grounding [16, 20] due to wide availability of interaction vs. egocentric images. However, this process can introduce errors because it depends on using imperfect class activation maps (CAMs) from the exocentric view to identify areas related to an affordance. These CAMs may accidentally include unrelated background areas or might not be accurate due to commonly encountered hand occlusion.

We experiment with **heatmap guidance**, as we reason that attribute and part knowledge can complement knowledge transfer from exocentric views. By default, for a given affordance a, an exocentric localization map $\mathcal{P}_{exo}$ is produced by thresholding the min-max normalized activations produced by DINO ($\mathcal{F}_{exo}$). The width and height dimensions of these features make up a 14×14 grid. For additional supervision, we also produce $\mathcal{P}_{CLIP}$, which is the min-max normalized and thresholded relevance heatmap gathered from CLIP using a fully specified prompt (with affordance, object, part, and attributes info). We average pool features to the same size as DINO (14×14). We further define a threshold t_{agg} as a parameter to specify the activation

CLIP Prompt	mKLD↓	mSIM↑	mNSS↑
{affordance} EX: cut with	1.713	0.297	0.688
{object} EX: knife	1.534	0.333	0.875
the region of {object} which allows the action: {affordance} EX: the region of knife which allows the action: cut with	1.510	0.337	0.904
{part} EX: blade	1.581	0.319	0.854
{attributes} EX: sharp, pointed, metallic	1.662	0.305	0.768
{part} of {object} EX: blade of a knife	**1.478**	<u>0.341</u>	0.967
{part} of {object} which is {attributes} EX: blade of a knife which is sharp, pointed, metallic	1.483	<u>0.341</u>	**0.973**
{part} of {object} which allows the action: {affordance} EX: blade of a knife which allows the action: cut with	1.484	0.340	0.950
{part} of {object} which is {attributes} and allows the action: {affordance} EX: blade of a knife which is sharp, pointed, metallic and allows the action: cut with	<u>1.479</u>	**0.342**	<u>0.972</u>

Table 2. **Zero-shot affordance grounding with CLIP, using groundings extracted with [6] for ViT-B/32.** AGD20k-Seen test (Egocentric). Lower mKLD is better, and higher mSIM/mNSS is better. **Bold** = top. <u>Underlined</u> = 2nd best.

level needed from CLIP relevance maps (similar to Eq. 3).

To investigate potentially complementary benefits of auxiliary affordance knowledge and HOI knowledge, we specifically test taking the *union* ($\mathcal{P}_{CLIP} \cup \mathcal{P}_{exo}$) and *intersection* ($\mathcal{P}_{CLIP} \cap \mathcal{P}_{exo}$) between the two heatmaps. We reason that union can provide more coverage over the affordance region, while intersection may be able to isolate specific regions on object parts that are relevant to affordances. Both aggregation methods have potential drawbacks: (1) union runs the risk of adding noise from either heatmap, and (2) intersection might lack coverage of the affordance region. Therefore, experimentation with both strategies can capture tradeoffs. We notably also test a simple baseline *replacing* LOCATE's map (i.e. *CLIP-guidance only*) to test CLIP's effects alone. Figure 3 shows each of these methods.

4. Experimental Results

4.1. Setup

Dataset. We evaluate on **AGD20K** [20], consisting of 20,061 exocentric interaction images drawn from HICO [5] and COCO [19], and 3,755 egocentric images. The goal is to learn with exocentric knowledge to perform inference on egocentric views. There are two class settings for evaluation: "seen"/"unseen". In the "seen" split, there are 36 affordance classes, with potentially multiple objects for each class. Every affordance-object combination is used for training and testing. In the "unseen" split, there are 25 affordance classes, though certain object classes are reserved for training and others for testing. The "unseen" setting is thus used to test grounding generalizability across objects. For all images, ground-truth maps are produced by applying Gaussian blur to keypoints in densely annotated maps of affordance regions, followed by image normalization.

Evaluation metrics. Like prior work [3, 16, 20], we use three commonly used metrics: Kullback-Leibler divergence (KLD) [3], similarity or histogram intersection (SIM) [17, 29], and normalized scanpath saliency (NSS) [26].

Training. Each model is trained on 1 NVIDIA RTX A5000 GPU with 24 GB of memory. Notably, testing on seen/unseen splits requires training separate models. All models are trained for 15 epochs and use LOCATE's default settings. Python 3.7.1, PyTorch version 1.7.1, and Cuda 11.0 are relevant software versions. Reproduced results differ from reported results in [16], so we report both sets.

Modeling. ChatGPT outputs are gathered with *gpt-3.5-turbo*, at settings of max tokens 100 and temperature 0.7. The ViT-B/32 version of CLIP is used for grounding.

4.2. Results and Analysis

In this section, we outline results for how we optimize CLIP's utility in affordance grounding, as well as the strategies used to integrate auxiliary knowledge within LOCATE.

How can affordance groundings effectively be probed from CLIP? We establish a baseline for CLIP's zero-shot grounding capabilities, across a variety of prompting strategies. As outlined in Sec. 3.2, we gather part and attribute information from an LLM to complement known affordance and object information, and use the map extraction method of [6] to produce groundings with the ViT-B/32 encoder. Table 2 demonstrates the results on AGD20k-Seen, for which

Strategy	t_{sp}	Seen			Unseen		
		mKLD↓	mSIM↑	mNSS↑	mKLD↓	mSIM↑	mNSS↑
Reported [16]	-	1.226	0.401	1.177	1.405	**0.372**	1.157
Reproduced [16]	-	1.227	0.400	1.200	1.417	0.367	1.153
SP Loss	0.2	1.211	0.398	1.197	1.420	0.362	1.144
	0.4	**1.198**	**0.403**	**1.210**	**1.402**	0.369	1.151
	0.6	1.203	**0.403**	1.206	**1.402**	0.369	**1.159**

Table 3. **Spatial consistency loss in LOCATE.** Spatial consistency loss is tested at various threshold values for t_{sp}. The prompt used in heatmap extraction is "{part} of {object} which is {attributes} and allows the action: {affordance}". **Bold** = top in column. Underlined = 2nd best. Spatial consistency is most effective in seen setting with respect to mKLD.

Strategy	SP Loss	t_{agg}	Seen			Unseen		
			mKLD↓	mSIM↑	mNSS↑	mKLD↓	mSIM↑	mNSS↑
Reported [16]		-	1.226	0.401	1.177	1.405	**0.372**	1.157
Reproduced [16]		-	1.227	0.400	1.200	1.417	0.367	1.153
SP Loss (t_{sp}=0.4)	✓	-	1.198	**0.403**	1.210	**1.402**	0.369	1.151
Union		0.2	1.224	0.397	1.209	1.422	0.361	1.161
		0.4	1.226	0.398	1.205	1.423	0.365	1.152
	✓	0.2	**1.194**	0.400	**1.223**	1.407	0.362	1.170
	✓	0.4	1.200	0.401	1.209	**1.402**	0.367	1.159
Intersection		0.2	1.225	0.401	1.199	1.435	0.367	1.124
		0.4	1.232	0.401	1.189	1.465	0.365	1.087
	✓	0.2	1.206	0.402	1.197	1.423	0.367	1.121
	✓	0.4	1.205	**0.403**	1.192	1.440	0.363	1.105
Replace		0.2	1.223	0.402	1.170	1.420	0.362	1.152
		0.4	1.223	0.400	1.198	1.437	0.365	1.120
	✓	0.2	1.199	0.400	1.216	1.404	0.362	**1.171**
	✓	0.4	1.201	0.402	1.201	1.426	0.363	1.128

Table 4. **Aggregation strategies and spatial consistency loss in LOCATE.** Spatial consistency loss is at threshold 0.4, while the aggregation strategies are tested over values of t_{agg}. Highlighted in light gray are notable combinations with multiple top/second-best values overall in terms of mKLD and/or mNSS. **Bold** = top in column. Underlined = 2nd best.

we make the following general observations: (1) Specifying just the affordance (*e.g.* "cut with") is the least successful approach. (2) The object name (*e.g.* "knife") provides a strong prior for grounding. Part information is further helpful, especially in combination with the object name (*e.g.* "blade of knife"), reaching the top mKLD (1.478). (3) Attribute information can be complementary to object and part knowledge, as shown in the approach with top mNSS (0.973). (4) The approach with *all* information is either the best or second-best across all metrics, demonstrating benefits to a high level of specification. Overall, the benefits of the full specification are shown in improvements of 0.234 mKLD, 0.045 mSIM, and 0.284 mNSS vs. the default affordance prompt.

How effective is an explicit spatial consistency loss as an enhancement strategy? We experiment with a spatial consistency loss to facilitate additional supervision from CLIP for egocentric affordance grounding. We present results in Table 3 over three threshold values for t_{sp}, and compare to the reproduced and reported LOCATE baselines. We observe the most significant differences in mKLD, where we maximally see 0.029 mKLD improvement over the reproduced baseline in the seen setting (row 4 vs. 2). Using a threshold of t_{sp} of 0.4 offers improvements over the reproduced setting in 5 out of 6 metric settings. The spatial consistency loss therefore shows added benefit from foundational model knowledge. The gains are more pronounced in the seen setting, perhaps since the attribute and part knowledge used in training is tailored towards objects, and objects at test time are seen in training. Still, since objects for the same affordance can share attributes (*e.g.* one can hold a *golf club* and a *tennis racket* with its *long, grippy handle*), the added object, attribute, and part guidance still benefits unseen mKLD despite an object not being seen in training.

How effective is heatmap aggregation as an enhancement strategy? As outlined in Sec. 3.3, the aggrega-

tion strategy used to combine CLIP and the HOI heatmaps can either alleviate undercoverage of an affordance region (*union*) or eliminate background noise (*intersection*). There can be tradeoffs with these strategies as shown earlier in Figure 3. To investigate potential tradeoffs, we perform experiments over various forms of aggregation, including a CLIP-only *replace* strategy to test CLIP's effects alone. We also test aggregation strategies in tandem with the spatial consistency loss strategy. Table 4 demonstrates these results over thresholds for t_{agg}. We find that the overall best setting in terms of seen mKLD and mNSS is union-based (t_{agg}=0.2, with spatial consistency loss). In particular, this union strategy achieves seen improvements over the reproduced baseline of 0.033 mKLD and 0.023 mNSS, while maintaining SIM. There are thus mKLD and mNSS benefits to the foundational model knowledge enhancement. In unseen, the improvements are 0.010 mKLD and 0.017 mNSS, though there is a 0.005 drop in SIM.

Comparing this method's numbers (row 6) to just using a spatial loss (row 3), the effects, especially in mNSS, can be complementary: there are gains of +0.013 mNSS (1.210 to 1.223) in seen and +0.017 mNSS (1.153 to 1.170) in unseen. The replace strategy (t_{agg}=0.2, with spatial consistency loss) is similarly effective, indicating CLIP's coverage of parts and attributes is beneficial to affordance grounding and that *lacking coverage of the affordance region is a problem in exo-to-ego knowledge transfer*. We reason that LOCATE's PartSelect module reduces the impact of overcoverage decently.

We find that aggregation strategies alone have limited effectiveness. Union (row 4/5) helps slightly vs. the reproduced baseline in mNSS (row 2), but the benefits are more noticeable in conjunction with the spatial consistency loss (row 6/7). We find that intersection does not add benefits vs. the baselines in both the seen and unseen setting. These results indicate that there is likely a lack of coverage of the key parts for an affordance in this scenario.

How do results compare to a model with a CLIP backbone? One relevant comparison is to replace the DINO backbone with CLIP, specifically to test if the combination of extracting foundational knowledge plus using the DINO pretrained model (our approach) is more effective than just using the foundational model (CLIP backbone). In Table 5, we compare a CLIP backbone model (CLIP BB) with our approach using spatial consistency loss and the union strategy. We find a considerable drop when swapping the DINO backbone with CLIP; this may be because self-supervised vision transformers (ViTs) like DINO contain explicit semantic segmentation information [4]. We speculate that this capability may be an artifact of visual self-supervision in ViTs compared to CLIP's language supervision. Overall, there is value in combining DINO and CLIP.

Strategy	Seen			Unseen		
	mKLD↓	mSIM↑	mNSS↑	mKLD↓	mSIM↑	mNSS↑
Repo. [16]	1.226	0.401	1.177	1.405	0.372	1.157
Repr. [16]	1.227	0.400	1.200	1.417	0.367	1.153
CLIP BB	1.432	0.352	0.980	1.594	0.313	1.019
Ours	1.194	0.400	1.223	1.407	0.362	1.170

Table 5. **Overall comparison, including LOCATE with CLIP backbone.** We show the benefits of our HOI and foundational knowledge combination. Ours is with spatial consistency loss (t_{sp}=0.4) and union aggregation (t_{agg}=0.2). The prompt used in heatmap extraction is "{part} of {object} which is {attributes} and allows the action: {affordance}". BB=backbone, Repo=reported, Repro=reproduced.

Which affordance classes are best aided by foundation model knowledge? We conduct a classwise analysis to see which affordances are most aided by the foundational knowledge. We find that in the seen setting for the method with union heatmap aggregation (t_{agg}=0.2) and spatial consistency loss (t_{sp}=0.4), the following classes have greater than 0.1 KLD improvements (which are decreases) vs. the reproduced baselines: *brush with* (+0.160), *drag* (+0.568), *pick up* (0.207), *push* (+0.429), *ride* (+0.128), *text on* (+0.143), and *write* (+0.249). Some notable affordance-object combinations are: *brush with toothbrush* (+0.568), *lie on surfboard* (+0.402), *open bottle* (+0.281), *open suitcase* (+0.353), *push bicycle* (+0.515), *push motorcycle* (+0.292), *ride bicycle* (+0.277), *sit on bicycle* (+0.247), and *write pen* (+0.249). Auxiliary information, like "bristles" for *brush with toothbrush*, appears especially helpful in these cases.

5. Conclusion

Conclusions. We have provided insights into affordance grounding with LLMs+VLMs and shown benefits from incorporating foundational knowledge of affordances into an existing pipeline. Such insights can inspire future methods to ensure robust, scalable, and generalizable grounding.

Future work. While we demonstrate improvements with foundational model knowledge, there can be more effective ways to integrate the knowledge. Attributes and parts are often shared across objects for the same affordance. A method that strongly conditions predictions on attribute/part information may achieve better seen-to-unseen generalization. It may be desirable to extend past the use of exocentric-to-egocentric knowledge transfer. VLMs can also be adapted themselves for better zero-shot affordance grounding.

Acknowledgement. This work was in part supported by a National Science Foundation Grant No. 2006885.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020. 3

[2] Kyle Buettner and Adriana Kovashka. Investigating the role of attribute context in vision-language models for object recognition and detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5474–5484, 2024. 3

[3] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):740–757, 2019. 6

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 3, 8

[5] Yu-Wei Chao, Yunfan Liu, Michael Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 381–389, 2017. 6

[6] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 387–396, 2021. 3, 5, 6

[7] Joya Chen, Difei Gao, Kevin Qinghong Lin, and Mike Zheng Shou. Affordance grounding from demonstration video to target image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6799–6808, 2023. 2

[8] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 975–983, 2018. 2

[9] Shengheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3D AffordanceNet: A benchmark for visual object affordance understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1778–1787, 2021. 2

[10] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J Lim. Demo2vec: Reasoning object affordances from online videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2139–2147, 2018. 2

[11] Mohammed Hassanin, Salman Khan, and Murat Tahtali. Visual affordance and function understanding: A survey. *ACM Computing Surveys (CSUR)*, 54(3):1–35, 2021. 2

[12] Zhi Hou, Baosheng Yu, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Affordance transfer learning for human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 495–504, 2021. 2

[13] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In *International Conference on Machine Learning*, pages 9118–9147. PMLR, 2022. 3

[14] Achiya Jerbi, Roei Herzig, Jonathan Berant, Gal Chechik, and Amir Globerson. Learning object detection from captions via textual scene attributes. *arXiv preprint arXiv:2009.14558*, 2020. 3

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2

[16] Gen Li, Varun Jampani, Deqing Sun, and Laura Sevilla-Lara. LOCATE: Localize and transfer object parts for weakly supervised affordance grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10922–10931, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[17] Jia Li, Changqun Xia, Yafei Song, Shu Fang, and Xiaowu Chen. A data-driven metric for comprehensive evaluation of saliency models. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 190–198, 2015. 6

[18] Yong-Lu Li, Yue Xu, Xinyu Xu, Xiaohan Mao, Yuan Yao, Siqi Liu, and Cewu Lu. Beyond object recognition: A new benchmark towards object concept learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20029–20040, 2023. 2

[19] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, 2014. 6

[20] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2252–2261, 2022. 1, 2, 4, 5, 6

[21] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Leverage interactive affinity for affordance learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6809–6819, 2023. 2

[22] Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. Language models of code are few-shot commonsense learners. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Abu Dhabi, UAE, 2022. 3

[23] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *International Conference on Learning Representations, ICLR*, 2023. 3

[24] Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1374–1381. IEEE, 2015. 2

[25] OpenAI. ChatGPT: Optimizing language models for dialogue, 2022. 2, 3

[26] Robert J. Peters, Asha Iyer, Laurent Itti, and Christof Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45:2397–2416, 2005. 6

[27] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? Generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701, 2023. 3

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 2, 3

[29] Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin, and Thierry Dutoit. Saliency and human fixations: State-of-the-art and study of comparison metrics. *2013 IEEE International Conference on Computer Vision*, pages 1153–1160, 2013. 6

[30] Anirban Roy and Sinisa Todorovic. A multi-scale CNN for affordance segmentation in RGB images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 186–201. Springer, 2016. 2

[31] Johann Sawatzky, Abhilash Srikantha, and Juergen Gall. Weakly supervised affordance detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[32] Johann Sawatzky, Yaser Souri, Christian Grund, and Jurgen Gall. What object should I use?- Task Driven Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7605–7614, 2019. 2

[33] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017. 5

[34] Jiajin Tang, Ge Zheng, Jingyi Yu, and Sibei Yang. CoTDet: Affordance knowledge prompting for task driven object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3068–3078, 2023. 2, 3, 4

[35] Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5945–5954, 2017. 3

[36] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[37] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 2