EEMAG : EGOCENTRIC-EXOCENTRIC MULTIMODAL AFFORDANCE GROUNDING

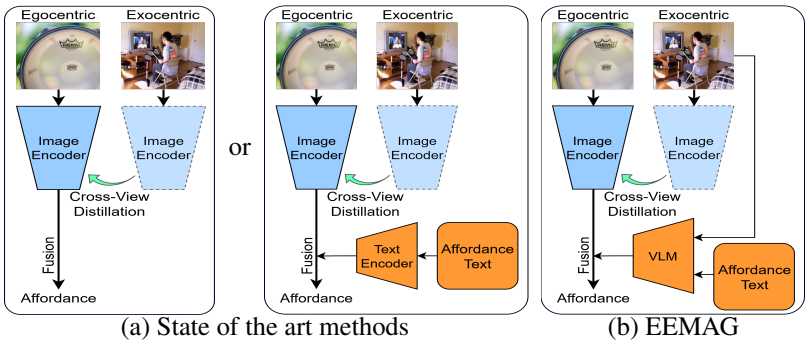
Anonymous authors
Paper under double-blind review

ABSTRACT

With the rise of intelligent systems, interactions between humans and robots need to be more precise and understandable. To this end, affordance detection, linking each affordance to precise objects' regions, has become more important than ever. This bridge between affordances and regions is learned by humans through a mix of experimentations and instructions. In this paper, a novel framework is presented based on this multimodality of affordance learning, between exocentric experiments and semantic instructions : Egocentric-Exocentric Multimodal Affordance Grounding (EEMAG). It extends the multimodal information extraction from a semantic description to the multimodal description of ongoing action. This knowledge is then transferred to an Egocentric Branch through a process of distillation and fusion. To achieve this goal, we propose a new branch to replace textual branches : the Multimodal Branch. It uses a Visual Language Model and Perceivers to extract visual and semantic information about the exocentric image. This branch effectively separates affordances and guide the Egocentric Branch towards useful visual features. Our framework is trained using the weakly supervised paradigm, using only image-level labels. Different experiments demonstrate the performance of our framework compared to state of the art affordance grounding models.

1 INTRODUCTION

Affordance, as described by Gibson (2014) in psychology, refers to all the "action possibilities" of an object. Thus, the affordance describes the uses of any object. A common example is "cut with" for a knife. Next, this definition has been extended into design (Maier & Fadel, 2009) and computer vision (Do et al., 2018). Affordance in computer vision has been the basis for many applications, such as robotic exploration (Ardón et al., 2021) or Human Machine Interaction (Li & Dai, 2025). These applications need a localization of the affordance in space. This need has focused the research towards affordance segmentation and affordance grounding.



(a) State of the art methods (b) EEMAG

Figure 1: Comparison between State of the art methods Luo et al. (2022)Xu et al. (2024) and our proposed method EEMAG. (a) State of the art methods focus on static or semantic context, while (b) EEMAG leverages useful information from exocentric images and textual information to better adapt to different affordances in inference.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Many classic approaches (Do et al., 2018)(Yin et al., 2021)(Zhao et al., 2020) for these tasks use affordance segmentation maps for training. However, segmentation maps must be rigorously annotated, and can be annotated differently by several annotators. Thus, they are prone to two problems : (1) precision errors and (2) compromises. (1) affordances often involve precise regions, making manual annotation particularly challenging. For example, the thin rim of a cup is used to "drink", making any segmentation for this affordance challenging. (2) Several affordances can be associated to a same object region. For example, the rim of the cup can be used to "drink", but also to "hold" the cup. However, 2D segmentation maps consider only one class per pixel, limiting their ability to represent multiple affordances for the same object. To lessen these problems, weakly supervised affordance grounding use heatmaps per affordances. They allow multiple affordances to be on the same pixel while lessening the burden of manual annotation. Hence, the research has focused recently on affordance grounding following the introduction of AGD20K (Luo et al., 2022) and the numerous methods built on this dataset (Li et al., 2023)(Xu et al., 2024)(Zhang et al., 2024b).

Nonetheless, many state of the art methods still lack adaptability to the scene and context. Most of them are mono-modal methods (Luo et al., 2022)(Do et al., 2018)(Li et al., 2023), focusing exclusively on visual aspects. Some do include contextual training through a Cross-view distillation (Luo et al., 2022)(Li et al., 2023), separating the model into two branches. But they lack context during inference and link affordances to fixed visual features with no respects to the context of the affordance. However, such context can be found in the semantics of the verbs describing the affordances. These verbs contain meaning and expectations for these affordances. This additional information has been treated differently in many papers, and has consistently led to improved results (Zhang et al., 2024b)(Zhang et al., 2022)(Jang et al., 2025). This has lead to the creation of a third branch in WSMA (Xu et al., 2024), focusing on the textual understanding of the affordance. However, this understanding stays fixed with no regards to the exocentric scene at hand. Our approach aims to leverage this exocentric scene during inference through the creation of a multimodal branch instead of a textual one.

To expand the adaptability of affordance grounding, we propose the Egocentric-Exocentric Multimodality for Affordance Grounding framework (EEMAG), as illustrated in Figure 1. This framework has two phases : (1) training and (2) inference. (1) three branches are used : Exocentric Branch, Egocentric Branch and Multimodal Branch. The first two are visual branches, used to capture useful visual features and classify affordances. The Exocentric and Egocentric Branches are similar, but with different inputs. This change of inputs is leveraged to explore and retain affordacne-related visual features. The Multimodal Branch uses a Visual Language Model (VLM) and PErceiver modules to select important features corresponding to the situation and semantic prior knowledge of affordances. This selection passes through a multimodal embedding. This embedding is fused in the Egocentric Branch to refine its results. (2) only two branches remain : Egocentric Branch and Multimodal Branch. Most of the process stay the same for both branches until the end. At the end, the output from the Egocentric Branch is fed to a Refined CAM to obtain the final grounding.

In summary, the main contributions of this article are as follows:
• The proposal of a multimodal, weakly-supervised framework (EEMAG) designed to localize affordance regions by leveraging exocentric and semantic knowledge, and effectively transferring this knowledge to the egocentric branch. A code will be available at EEMAG_git.
• The integration of Perceiver modules to format and fuse different types of information. Our approach uses two types of information : textual and visual.
• The evaluation conducted on AGD20K and HICO-IIF (Xu et al., 2024) clearly demonstrates the performance of this approach over reference methods, showing better performance for affordance grounding when exocentric images are provided.

2 Related work

Affordance Grounding is a computer-vision task aiming to locate particular regions associated with different affordances. These affordances are generally named with action verbs such as cut or drink. Originally, the task was closer to Affordance Segmentation with pixel-level annotations and supervised learning (Do et al., 2018)(Yin et al., 2021)(Zhao et al., 2020). However, these approaches were limited because of the complexity in the annotations and concurrency with Affordance Segmentation. Other approaches relied on the weakly supervised paradigm, using a restricted number

of points of reference (Sawatzky & Gall, 2017) or leveraging videos to learn affordances through human interaction (Apicella et al., 2023)(Luo et al., 2024). Since 2022, weak supervision has become the norm for Affordance Grounding, using image-level annotation with AGD20K. This has been a major focus for the domain, with approaches such as Cross-View-AG (Luo et al., 2022), LOCATE (Li et al., 2023), WSMA (Xu et al., 2024) or INTRA (Jang et al., 2025). Each of these approaches mainly used exocentric images as added information in training. In contrast, our work introduces a framework to extend the use of exocentric images to the inference case. A small survey is proposed on Cross-view Knowledge Distillation and Vision Language Models, as they are key components in our work.

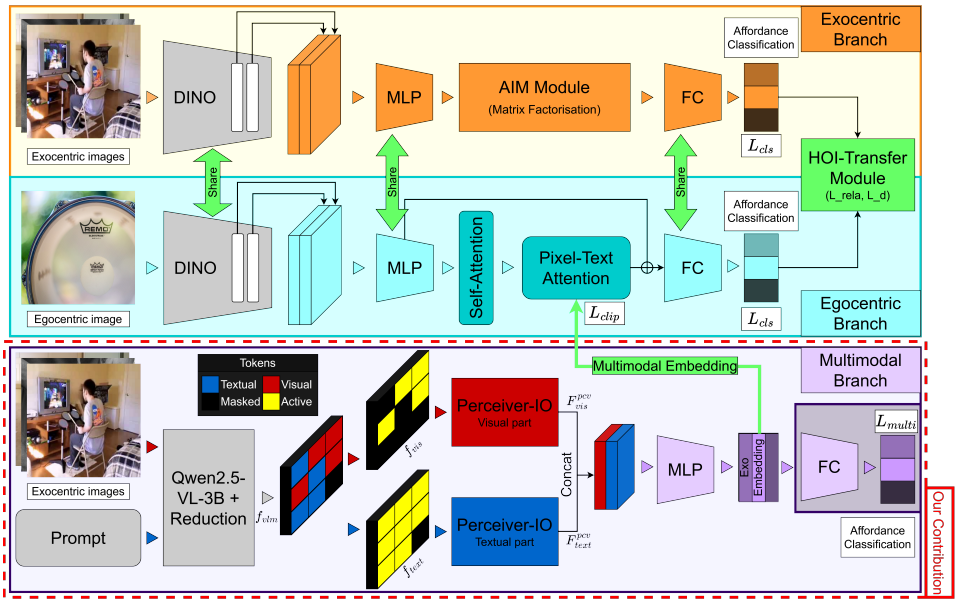


Figure 2: Overview of our proposed framework in training phase : EEMAG (in red dashed line). During training, the model is separated between three branches : (top-yellow) Exocentric, (middle-blue) Egocentric and (bottom-purple) Multimodal. The Exocentric branch transfers action related features to the Egocentric branch. The Multimodal branch transfers action related knowledge from the exocentric images to the Egocentric branch. The Egocentric branch extracts features and fuses them with the multimodal embedding to refine its choice. Any module in color is trained. Perceivers are related to the Multimodal branch while following the tokens color scheme. Interaction between branches are shown in green.

2.1 Cross-view Knowledge Distillation

Knowledge distillation is a deep-learning training technique that aims to transfer model knowledge from a teacher model to a student model (Moslemi et al., 2024). It has been the basis for different domains: size reduction, self-supervision, and cross-view knowledge. This last one treats the distillation of knowledge between different perspectives. It is open to changes in angle but also situation, hence its usage in exocentric to egocentric distillation. It has been notably the case for Cross-View-AG (Luo et al., 2022), LOCATE (Li et al., 2023), WSMA (Xu et al., 2024) and INTRA (Jang et al., 2025). This paper uses the basis of WSMA for its Cross-view Knowledge distillation process, which itself is based on Cross-View-AG's distillation process.

2.2 Vision Language Models and Large Language Models

Vision Language Models (VLM) have evolved rapidly recently (Zhang et al., 2024a). They aim for comprehension in visual and textual tasks, but more importantly at their interactions. They are separated into two main groups: direct VLMs and VLMs fused with Large Language Models (LLM). (1) The best example of a direct VLM is CLIP (Radford et al., 2021). They aim to create a strong nexus between visual and textual information, consisting of a textual and visual encoder trained

jointly at the end. (2) VLMs fused with Large Language Models (LLM) vary in nature. They can be double-tower, single-tower, or fused toward the LLM architectures. They are trained differently to fit multiple tasks with textual outputs. VLMs of all sorts have already impacted affordance grounding. Indeed, CLIP has been used in many works, mostly for its text encoder (Xu et al., 2024)(Zhang et al., 2024b). However, while more complex VLMs are used (Qian et al., 2024), they are used only for textual comprehension and semantic prior knowledge with the egocentric image. Our work aims to use VLMs not only for semantic prior knowledge, but also for exocentric comprehension at inference. It leads to better adaptation relying on dynamic action description, compared to other methods relying on affordance specific prompts.

3 PROPOSED FRAMEWORK : EEMAG

Our framework is illustrated in Figures 2 for training, and Figure 4 for inference respectively. This framework is designed as an extension of WSMA's Cross-View Distillation process. As such, first subsection will shortly introduce WSMA architecture, variables and losses. Then, our multimodal contribution will be detailed.
We present some notations used in this paper. n : number of affordances, depends on the dataset. n_{exo} : number of exocentric images processed per pass. b_{dino} : number of blocks in the feature extraction model in the egocentric and exocentric branches. b_{vlm} : number of blocks in the VLM in the multimodal branch. b_{inter} : number of the intermidiary block used in the multimodal branch.

3.1 SHARED ARCHITECTURE WITH WSMA

The proposed framework builds upon the Cross-View Distillation architecture introduced by Xu et al. (2024). Since the present work extends rather than replaces this framework, we first summarize its core components before detailing the proposed multimodal branch.
Feature extraction : Both Egocentric and Exocentric Branches use the same feature extraction from their featuring model, DINO-VIT (Caron et al., 2021), noted $M(\cdot)$. It provides visual features from an image I : $M(I) = (f^1, f^2, ..., f^{b_{dino}-1}, f^{b_{dino}})$. In the proposed framework, both $f^{b_{dino}-1}$ and $f^{b_{dino}}$ are fused using an MLP. These final features are noted f_{exo} or f_{ego} according to branch.
Exocentric Branch : The feature extraction is done over n_{exo} exocentric images. This process yields $(f^1_{exo}, f^2_{exo}, ..., f^{n_{exo}}_{exo})$. These features are then fed to the Affordance Invariance Mining module (AIM) Module (Luo et al., 2022), using matrix factorization (Lee & Seung, 2000). It generalizes each features into : $f^i_{exo} = W_{exo} \times H^i_{exo} + E^i_{exo}$. f_{exo} and $(W_{exo} \times H^i_{exo})_i$ are fed to a residual pass and a fully connected layer. This yields a final score c_{exo}.
Egocentric Branch : The feature extraction is done over a single egocentric image, resulting in f_{ego}. This extraction is then followed by a Pixel-Text Fusion Module (Xu et al., 2024). It uses f_{ego} with a self attention layer and multimodal features f_{multi} to create a score Z_{clip} and an attention map f_{att}. Z_{clip} is used for $L_{clip} = CrossEntropy(Z_{clip})$. f_{att} is then used in a residual pass with f_{ego} to create the final egocentric features F_{ego}. A fully connected layer is used to create the final score c_{ego}, completing $L_{cls} = CrossEntropy(c_{exo}) + CrossEntropy(c_{ego})$.
HOI-Transfer : Both c_{exo} and c_{ego} are used in the HOI-Transfer Module (Xu et al., 2024). To target precise interaction regions, the module relies on the Class Activation Mapping method (CAM) (Minh, 2023). They are applied for each expected affordance classes to create two feature maps per affordance, Y_{ego} and Y_{exo}. These are then used for two losses : $L_d = ||Y^C_{ego} - Y^C_{exo}||$ and $L_{rela} = Cosine(Corr(Y_{ego}), Corr(Y_{exo}))$.

3.2 PROPOSED MULTIMODAL BRANCH

Compared to a textual branch (Xu et al., 2024), the multimodal branch is more dynamic thanks to the use of exocentric images. As shown in Figure 2, this branch has two different inputs: a prompt and exocentric images. For simplicity, the prompt is neither learned nor averaged, unlike other works (Kawaharazuka et al., 2024). It is kept as a general prompt asking for the affordance, with a template response. The used prompt is provided in Section 4.2. The exocentric images are the same as those used in the exocentric branch.

3.2.1 VLM BACKBONE

This Multimodal Branch uses a VLM, Qwen2.5 VLM (Li et al., 2026), as its backbone. This allows for a deep context understanding from the exocentric image. This VLM is composed of a visual encoder and an LLM. The output of the visual encoder is fed to the LLM along with the prompt, and the final processing is performed by the LLM. This processing yields the following output: $(f^1_{vlm}, f^2_{vlm}, ..., f^{b_{vlm}}_{vlm})$. An intermediary block is used to maintain a better balance between textual and spatial features. Thus, the model uses $f_{inter} = f^{b_{inter}}_{vlm}$. These features are then passed through a projection layer to obtain smaller embeddings per token, as follows $f_{vlm} = GELU(LayerNorm(Linear(f_{inter})))$.

3.2.2 VISUAL AND TEXTUAL SEPARATION

Through their training, VLMs' outputs are highly semantic. However, this model leverages a full exocentric comprehension. Hence, to balance both visual and semantic elements, the visual and textual tokens are separated in f_{vlm}. This is done using the input's special tokens. These tokens include the image token, a token made to be replaced by the features from the VLM's visual encoder. Two attention masks are created from this : **Textual attention mask** att_{text} : includes all tokens outside of padding. **Visual attention mask** att_{vis} : includes only special image token. They are respectively used to create : $f_{text} = f_{vlm}[att_{text}]$ and $f_{vis} = f_{vlm}[att_{vis}]$.
While these attention masks are created from the VLM's input, VLMs keep a similar layout in between its blocks. This separation is the basis of the branch. It highlights both clues and raises them to the same level for further processes. However, these features still have sparse imbalance because of the different masks.

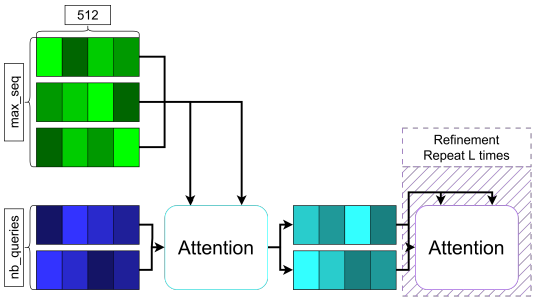


Figure 3: Overview of Perceiver blocks used in EEMAG. Serves to adapt inputs with different sizes or attention masks. Uses a set of queries as a base to further help the model.

3.2.3 QUERY ADAPTATION

To address this imbalance and generalize the approach towards other modality, the framework includes a query-based adaptation built on Perceiver (Jaegle et al., 2021). The queries are used to adapt the VLM's output to the number of affordances n and create a simple embedding for the rest of the model. A set of n queries is created per type of acquired information or modality. We have two sets of features : f_{text} and f_{vis}. Hence, our approach uses two sets of n queries, q_{text} and q_{vis}. As presented in Figure 3, the queries are passed through a multi-headed attention layer, with the modality output used as keys and values. This layer is then refined through the use of n_{refine} self-attention layers. With i refering to textual ($text$) or visual modality (vis), each perceiver outputs as follows : $F^{pcv}_i = SelfAttention^{n_{refine}}_i \circ ... \circ SelfAttention^1_i \circ Attention_i(q_i, f_i)$. Function composition is noted by $\circ$.

From this point, the VLM's output has been adapted to the query size and no longer has imbalance. Fusion of modalities is performed using a linear layer : $f_{multi} = Linear(concat(f_{text}, f_{vis}))$. This simple fusion keeps the order and query-based structure. Hence, f_{multi} has appropriate dimensions to be used in the Pixel-Text Fusion module. f_{multi} is also used to compute a classification score for the Multimodal Branch. It is reduced in dimension using a depth-wise linear layer and, finally, a fully connected layer to produce a score c_{multi}. This score is used with a cross-entropy loss to compute the multimodal loss $L_{multi} = CrossEntropy(c_{multi})$.

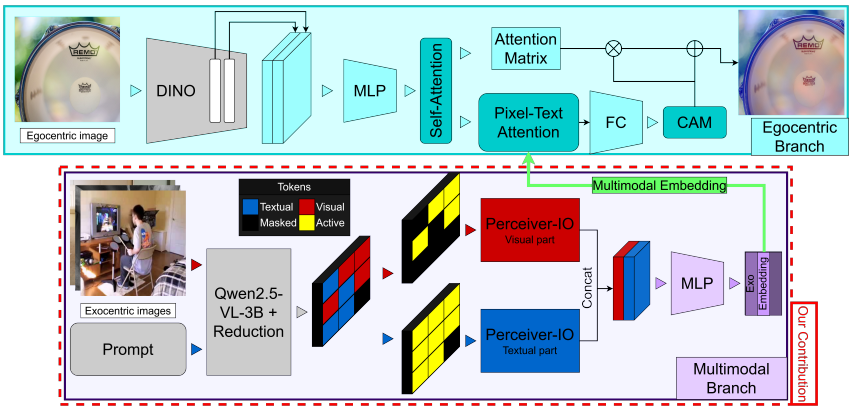


Figure 4: Description of the inference for EEMAG's framework. Our contribution is highlighted in red. Egocentric and Multimodal branches are retained for this use. Multimodal branch's classification is put aside, with a focus on the Egocentric branch. It uses additional knowledge from the fusion to refine its prediction.

3.3 Inference

As seen in Figure 4, the inference model is not exactly the same as the training model. Indeed, during inference the task changes, and the model must produce an activation map for a single egocentric image and affordance. To match this change, only the multimodal and egocentric branches are kept, although the final multimodal score is discarded. Otherwise, the architecture stays the same until the egocentric score. With the CAM method, the heatmap H corresponding to the desired affordance is generated. To enhance the resulting heatmap, this paper proposes to use the CAM Refined module created by Xu et al. (2024). In this module, the attention map Q from the self-attention layer is used on top of H. A threshold t is then used to narrow down the grounding and create the final heatmap H', as follows $H' = Q \times H[H > t] + H$.

4 Experimental Results

4.1 Dataset and Metrics

The datasets employed are AGD20K (Luo et al., 2022) and HICO-IIF (Xu et al., 2024). AGD20K has become a staple of the domain because of its multiple viewpoints, extensive data volume, including 20061 exocentric and 3755 egocentric images, and its unique Seen/Unseen split. The Seen split contains training and testing data drawn from the same object categories, whereas the Unseen split uses training data on different categories than those used for testing (e.g., scissors are employed to train the "cut_with" affordance but knife is used for testing it). Experiments were also conducted on HICO-IIF to compare this approach with WSMA.
To evaluate model performance against dataset ground truths, three metrics are applied: Kullback-Leibler Divergence (KLD), Similarity (SIM), and Normalized Scanpath Saliency (NSS). These metrics are well established in the field and are further defined in prior works (Rezatofighi et al., 2019).

4.2 Implementation

For both the egocentric and exocentric branches, a pre-trained DINO-VIT-S (Caron et al., 2021) is used for the feature extraction. The VLM backbone in the multimodal branch is based on a pre-trained Qwen2.5-Instructor-3B (Li et al., 2026). Both models are frozen throughout training. Exocentric images are processed in parallel within each branch, while multimodal processing follows a separate pipeline. A full review of hyper-parameters is provided in Table 1. The general prompt used is provided in Figure 5.

```
Analyze the scene and return a single
 JSON object using the provided
template in English. Describe the
overall scene, including the number
of people, their actions, the objects
 present, the room or location, and
the main ongoing activities. When
describing ongoing activities, be as
precise as possible. Do not include
any text outside the JSON output. {"
overall_scene": "", "people": {"count
": , "actions": [""]}, "objects
":[""], "location": "","
main_ongoing_activities": ""}
```

Figure 5: General prompt used.

Parameter	Seen	Unseen	HICO-IIF
$\lambda_{cls}, \lambda_{clip}, \lambda_{multi}$	1	1	1
$\lambda_d, \lambda_{rela}$	0.5	0.5	0.5
t	0.2	0.2	0.2
n_{refine}	1	1	0
n	36	25	10
n_{exo}	3	3	3
b_{dino}	12	12	12
b_{vlm}	32	32	32
b_{inter}	28	28	28

Table 1: Summary of main hyper-parameters per dataset. AGD20K-Seen and AGD20K-Unseen are shortened to Seen and Unseen.

Method	AGD20K-Unseen			AGD20K-Seen			HICO-IIF		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
Cross-View-AG	1.787	0.285	0.829	1.538	0.334	0.927	1.779	0.263	0.946
LOCATE	1.405	0.372	1.157	1.226	0.401	1.177	1.593	0.327	0.966
AffordanceLLM	1.661	0.361	0.947	1.463	0.377	1.070	-	-	-
INTRA	1.365	0.375	1.209	1.199	0.407	1.239	-	-	-
WSMA	1.335	0.382	1.220	1.176	0.416	1.247	1.465	0.358	1.012
Selective-CL	**1.243**	**0.405**	**1.368**	**1.124**	**0.433**	**1.280**	**1.358**	**0.378**	1.231
EEMAG (Our)	1.261	0.399	1.292	1.152	0.419	1.251	1.360	0.372	**1.338**

Table 2: Comparison with affordance grounding state-of-the-art models : Cross-View AG (Luo et al., 2022), LOCATE (Li et al., 2023), WSMA (Xu et al., 2024), AffordanceLLM (Qian et al., 2024), INTRA (Jang et al., 2025) and Selective-CL (Moon et al., 2025). The best results are bolded, second best are underlined.

A critical hyperparameter is the learning rate. This model is highly sensitive to the learning-rate schedule within it. Because of its architecture, the multi-task loss L_{multi} focuses on a narrow set of 36 classifications compared to thousands of parameters needed for the queries. Thus the queries get small gradients across the different losses, and struggle to learn appropriate affordances. Consequently, the multimodal branch utilizes a specialized optimizer with a distinct layer ordering.

4.3 Comparison to State of the Art

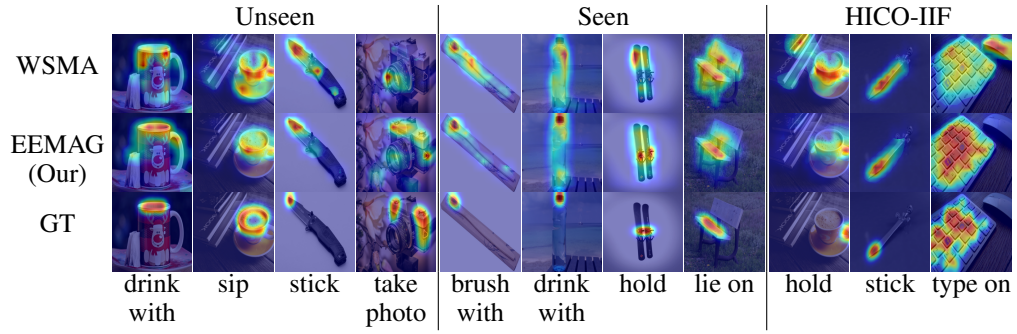
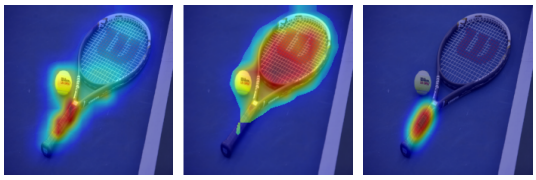


Figure 6: Visual comparisons between EEMAG (our approach) with the the base model WSMA (Xu et al., 2024).

Quantitative results : A benchmark against five other affordance-grounding models is presented in Table 2. The compared methods include Cross-View AG (Luo et al., 2022), LOCATE (Li et al., 2023), WSMA (Xu et al., 2024), INTRA (Jang et al., 2025), AffordanceLLM (Qian et al., 2024) and Selective-CL (Moon et al., 2025). The proposed model outperforms most of these approaches

across all three metrics, with EEMAG surpassing WSMA in every dataset. It compares closely to Selective-CL (Moon et al., 2025) overall. This result underscores the value of a more comprehensive understanding of affordances beyond mere semantic information.

Visual results : A qualitative assessment of the models is illustrated in Figure 6. These results highlight the more precise grounding. Moreover, they also show a better understanding of each affordance, leading to better region separation. For example, the toothbrush and bottle are treated more as global objects by WSMA, grounding the "grasp" affordance at the same time. However, our model better separates features for "grasp" and other affordances. Similarly, this is the case with other affordances, yielding more precise results consistently.



(a) AGD20K (b) HICO-IIF (c) GT

Figure 7: Visual grounding of tennis rackets' "swing" affordance. (a) is grounding from EEMAG trained on AGD20K. (b) is a grounding from EEMAG trained on HICO-IIF. (c) is the ground truth.

Affordance understanding : When looking at the results of our framework over different datasets, an unexpected problem can be observed. For the "swing" affordance, the results for the same objects vary widely between HICO-IIF and AGD20K. As shown in Figure 7, groundings over HICO-IIF focus on the racket's net, while groundings over AGD20K focus around the handle. This change can be linked to the affordance understanding. Indeed, in HICO-IIF, the "swing" affordance is trained and tested with tennis rackets, whereas, in AGD20K, the "swing" affordance is trained and tested with tennis rackets and baseball bats. This change introduces a shift in the affordance definition.

4.4 Ablative Study

To further validate the proposed approach, additional experiments were performed. Detailed outcomes are reported in Tables 3. Several non-redundant tests were excluded to avoid overlap with WSMA (Xu et al., 2024). These studies are focusing on the losses used to train the Multimodal Branch and the number of refining heads for the perceiver layers.

n_{refine}		AGD20K-Unseen			AGD20K-Seen			HICO-IIF		
		KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
0		1.599	0.330	0.953	1.172	**0.421**	1.235	**1.360**	**0.372**	**1.338**
1		**1.261**	**0.399**	**1.292**	**1.152**	0.420	**1.251**	1.942	0.226	0.824
2		1.292	0.392	1.247	1.175	0.419	1.203	1.387	0.369	1.278
4		1.273	0.394	1.290	1.166	0.419	1.236	1.475	0.353	1.193
L_{multi}	L_{clip}	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
0	1	1.270	0.400	1.267	1.162	0.427	1.249	1.512	0.346	**1.216**
1	0	1.473	0.345	1.128	1.211	0.398	1.239	2.130	0.211	0.593
0.5	1	1.328	0.388	1.218	**1.152**	0.418	1.253	**1.501**	**0.354**	1.213
2	1	1.262	**0.404**	**1.292**	1.156	**0.428**	**1.258**	1.932	0.225	0.857
1	1	**1.261**	0.399	**1.292**	**1.152**	0.419	1.251	1.942	0.226	0.824

Table 3: Comparison of number of refining heads on performance and of losses impact on Multimodal Branch with $n_{refine} = 1$. The best results are bolded, second best are underlined.

Ablative study - Refinement : For n_{refine}, the related ablation study is provided in Table 3 and in Figure 8. These experiments show a significant impact from the number of refining heads. Each dataset and split has its own best number of refining heads, although a catastrophic failure is occuring for $n_{refine} = 1$ with HICO-IIF and for $n_{refine} = 0$ with AGD20K-Unseen. These results open the analysis in three dimensions :

• *Catastrophic failures* : These failures can be explained by the difficult access to queries weights.

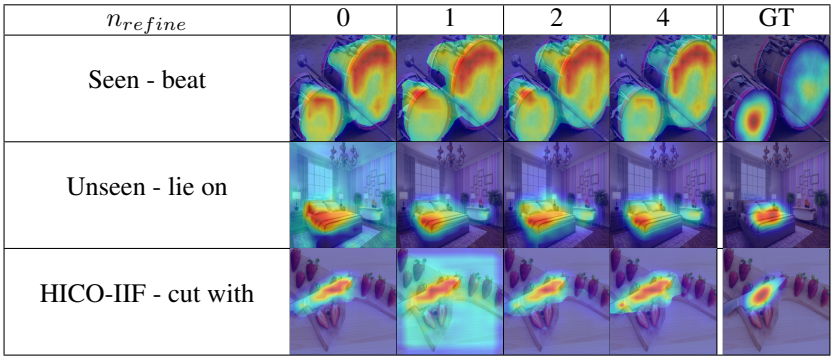


Figure 8: Visual comparison between number of refining heads. Ground truths are presented at the bottom.

This means that they are very dependant on initialization weights and first batches or epochs. However, in light of the visual results, these failures seem to retain important grounding but mixing it with global features, blurring the overall results.

• *Disparity between datasets* : The performance varies widely between different n_{refine}. This variation is more noticeable on HICO-IIF and AGD20K-Seen. Both these datasets have affordances with single representing object, indicating a sensitivity of refining heads to the exploration rate of each affordance.

• *Shift on smaller dataset* : The results for HICO-IIF are drastically different than for the other datasets. The two worst settings on AGD20K splits are the best there. It shows an instability in the model at the queries adaptation.

Because of the variety in results, we proposed our results with $n_{refine} = 1$ for AGD20K splits and $n_{refine} = 0$ for HICO-IIF. This choice was made to offset the catastrophic failure in the case of HICO-IIF with $n_{refine} = 1$.

Ablative study - Losses : For L_{clip} and L_{multi}, the related ablation study is provided in Table 3. These experiments help conclude the necessity of both loss. Results are worse in all metrics with only L_{multi}, but can be competitive to a certain degree with only L_{clip}. These results are analyzed in three dimensions :

• *Necessity of* L_{clip} : Only lacking L_{clip} leads to the model underperforming on every metric.

• *Broad or Narrow detection* : A notion on each metric is the narrow or broadness promoted in the grounding. KLD and SIM promote broader groundings, while NSS promotes narrower metrics. Hence, the imbalance between losses may lead to narrower metrics as they perform better on NSS and worse on KLD and SIM.

• *Stability* : A trade off is visible between performance and stability. The better performing settings on AGD20K are also those that fail on HICO-IIF. Thus L_{multi} seems to create variance in the training, leading to better results on certain cases and catastrophic failure on others.

5 Conclusion

This paper presents a novel weakly supervised affordance grounding model, EEMAG. Its main idea is to learn affordance knowledge from exocentric and semantic inputs. This is possible thanks to the new Multimodal Branch that extracts relevant information and guides the Egocentric Branch towards useful visual features. During testing, it uses all this information to determine affordance regions with a derivative to the CAM method. EEMAG is proven to have better results than other state of the art models. However, it still is faced by other limitations. Most notably, it is unstable in its training. It is limiting much of the framework, and needs a new approach to stabilise. Also it lacks very fine detection in most of the results, with blurry groundings. Moreover, some information could be reused. For example, the multimodal score could be used to perform better or ground without prior knowledge of the affordance searched. Thus, to address these challenges, we plan to improve the model with finer spatial fusion and work on a stronger relationship between the multimodal and egocentric branches.

REFERENCES

Tommaso Apicella, Alessio Xompero, Edoardo Ragusa, Riccardo Berta, Andrea Cavallaro, and Paolo Gastaldo. Affordance segmentation of hand-occluded containers from exocentric images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 1898–1907, October 2023.

Paola Ardón, Èric Pairet, Katrin S. Lohan, Subramanian Ramamoorthy, and Ronald P. A. Petrick. Building Affordance Relations for Robotic Agents - A Review, May 2021. URL http://arxiv.org/abs/2105.06706. arXiv:2105.06706 [cs.AI].

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 (ICCV)*, pp. 9650–9660, October 2021.

Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5882–5889, 2018. doi: 10.1109/ICRA.2018.8460902.

James J. Gibson. *The Ecological Approach to Visual Perception: Classic Edition*. Psychology Press, New York, November 2014. ISBN 978-1-315-74021-8. doi: 10.4324/9781315740218.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General Perception with Iterative Attention. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 4651–4664. PMLR, July 2021. URL https://proceedings.mlr.press/v139/jaegle21a.html. ISSN: 2640-3498.

Ji Ha Jang, Hoigi Seo, and Se Young Chun. INTRA: Interaction Relationship-Aware Weakly Supervised Affordance Grounding. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 18–34, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73039-9. doi: 10.1007/978-3-031-73039-9_2.

Kento Kawaharazuka, Naoaki Kanazawa, Yoshiki Obinata, Kei Okada, and Masayuki Inaba. Continuous Object State Recognition for Cooking Robots Using Pre-Trained Vision-Language Models and Black-Box Optimization. *IEEE Robotics and Automation Letters*, 9(5):4059–4066, May 2024. ISSN 2377-3766. doi: 10.1109/LRA.2024.3375257. URL https://ieeexplore.ieee.org/abstract/document/10465607.

Daniel Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2000.

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 (CVPR)*, pp. 10922–10931, June 2023.

Lei Li and Angela Dai. Hoi-page: Zero-shot human-object interaction generation with part affordance guidance, 2025. URL https://arxiv.org/abs/2506.07209.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, et al. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026.

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 (CVPR)*, pp. 2252–2261, June 2022.

Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Learning Visual Affordance Grounding From Demonstration Videos. *IEEE Transactions on Neural Networks and Learning Systems*, 35(11):16857–16871, November 2024. ISSN 2162-2388. doi: 10.1109/TNNLS.2023.3298638. URL https://ieeexplore.ieee.org/abstract/document/10246333.

Jonathan R. A. Maier and Georges M. Fadel. Affordance based design: a relational theory for design. *Res Eng Design*, 20(1):13–27, March 2009. ISSN 1435-6066. doi: 10.1007/s00163-008-0060-3. URL https://doi.org/10.1007/s00163-008-0060-3.

Anh Pham Thi Minh. Overview of class activation maps for visualization explainability, 2023. URL https://arxiv.org/abs/2309.14304.

WonJun Moon, Hyun Seok Seong, and Jae-Pil Heo. Selective contrastive learning for weakly supervised affordance grounding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5210–5220. IEEE, 2025.

Amir Moslemi, Anna Briskina, Zubeka Dang, and Jason Li. A survey on knowledge distillation: Recent advancements. *Machine Learning with Applications*, 18:100605, December 2024. ISSN 2666-8270. doi: 10.1016/j.mlwa.2024.100605. URL https://www.sciencedirect.com/science/article/pii/S2666827024000811.

Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. Affordancellm: Grounding affordance from vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 7587–7597, June 2024.

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 Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Johann Sawatzky and Jurgen Gall. Adaptive binarization for weakly supervised affordance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

Lingjing Xu, Yang Gao, Wenfeng Song, and Aimin Hao. Weakly Supervised Multimodal Affordance Grounding for Egocentric Images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(6):6324–6332, March 2024. ISSN 2374-3468. doi: 10.1609/aaai.v38i6.28451. URL https://ojs.aaai.org/index.php/AAAI/article/view/28451.

Congcong Yin, Qiuju Zhang, and Wenqiang Ren. A New Semantic Edge Aware Network for Object Affordance Detection. *J Intell Robot Syst*, 104(1):2, December 2021. ISSN 1573-0409. doi: 10.1007/s10846-021-01525-9. URL https://doi.org/10.1007/s10846-021-01525-9.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-Language Models for Vision Tasks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8): 5625–5644, August 2024a. ISSN 1939-3539. doi: 10.1109/TPAMI.2024.3369699. URL https://ieeexplore.ieee.org/abstract/document/10445007.

Yang Zhang, Huiyong Li, Tao Ren, Yuanbo Dou, and Qingfeng Li. Multi-scale Fusion and Global Semantic Encoding for Affordance Detection. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, July 2022. doi: 10.1109/IJCNN55064.2022.9892363. URL https://ieeexplore.ieee.org/abstract/document/9892363. ISSN: 2161-4407.

Zhipeng Zhang, Zhimin Wei, Guolei Sun, Peng Wang, and Luc Van Gool. Self-Explainable Affordance Learning with Embodied Caption, April 2024b. URL http://arxiv.org/abs/2404.05603. arXiv:2404.05603 [cs.CV].

Xue Zhao, Yang Cao, and Yu Kang. Object affordance detection with relationship-aware network. *Neural Comput & Applic*, 32(18):14321–14333, September 2020. ISSN 1433-3058. doi: 10.1007/s00521-019-04336-0. URL https://doi.org/10.1007/s00521-019-04336-0.

A Appendix

A.1 AI use statement

In this work, we **used generative AI** tools for assisting with translation. We have **not used generative AI** tools for implementing methods, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, supporting qualitative and thematic data analysis and interpreting results. Generating synthetic data sets and cleaning and reformating dataset are **not applicable** to this work. Additionally, we **used generative AI** tools for editing and debugging software code, editing a research paper to improve readability, drafting parts of a research paper and formating references. We have reviewed all AI-assisted work. LLM-generated code was mainly used for debugging purposes. The resulting code was adapted and tested for correctness on multiple occasions. Translations and editings were made to create different drafting styles and improve overall readability. Formated references have been reviewed to avoid inconsistencies. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

A.2 Ethics statement

While inspired by the work of Xu et al. Xu et al. (2024), this work aims to differ on the integration of a new input at inference time. This work aims to extend the already exisiting model with a new approach and not copy their work.

A.3 Reproductibility statement and code release

The code has been thought to be reproductible. All seeds are fixed at the begining. All datasets used are available online and were not changed during any test. To improve furthermore the reproductibility of the results, the initialisation of the different datasets sort the dataset before a random shuffle. A file is provided as a reference for all library versions used. The reproductability has been tested though the use of two computers running the code achieving similar results.

The code will be provided for camera ready files. This is an effort to keep anonymity during submission. A repository has already been created and the code inside will no longer be changed.