Selective Contrastive Learning for Weakly Supervised Affordance Grounding

Appendix

A. Datasets and Implementation Details

Datasets. To benchmark weakly supervised affordance grounding (WSAG) methods, we use two datasets, *i.e.*, AGD20K [4], and HICO-IIF [7]. AGD20K is composed of 3,755 egocentric images with 20,061 exocentric images that belong to 36 affordance classes with 50 object classes. Dense annotations are labeled according to the probability of interaction between the human and object regions where Gaussian blur is applied afterwards to generate the heatmaps. HICO-IIF [7] comprises 1,088 egocentric images and 4,793 exocentric images. HICO-IIF is collected from HICO-DET [1] and IIT-AFF [5] where both datasets are equipped with object and affordance categories.

Implementation Details. Following previous works [3, 7], we employ DINO ViT-S/16 for all experiments and set E, the number of exocentric images per egocentric image to 3. In addition, we set K, the number of clusters used to segment the objects in exocentric images for part-level prototypical contrastive learning, to 3. The model is optimized using the SGD optimizer with a learning rate of 1e-3, weight decay of 5e-4, and batch size of 8. Additionally, while maintaining consistent parameters across datasets, we vary the number of training epochs between ADE20k and HICO-IIF. Specifically, we train the ADE20k dataset for 15 epochs in both seen and unseen scenarios, whereas HICO-IIF is trained for 50 epochs. The extended training duration (3-4x) for HICO-IIF accounts for its dataset size, which is approximately 3–4 times smaller than ADE20k, requiring additional iterations to achieve performance saturation. The MLP is defined with a feed-forward network and each projection layer contains two convolution layers, followed by a classifier to generate CAMs. Projection layers for each contrastive loss are designed with a linear layer with a normalization layer.

Furthermore, as mentioned in the paper, we employ the strategy of ClearCLIP [2] to enhance local discriminability in the visual features of CLIP ViT-B/16. ClearCLIP introduces three key modifications to the original CLIP architecture in its final layer: (1) removal of the residual connection, (2) reorganization of spatial information through self-self attention (i.e., query-to-query attention [6]), and (3) elimination of the feed-forward network. These modifications are applied without the fine-tuning phase so that it uses the pretrained weights of the original CLIP. The impact of ClearCLIP over naïve CLIP is shown in Tab. A1.

Table A1. Affordance grounding results using CLIP-B/16 and ClearCLIP-B/16 in the AGD20k-Seen scenario.

Method	ZeroShot	KLD	SIM	NSS
CLIP	O	1.774	0.250	0.640
	X	1.160	0.412	1.267
ClearCLIP	O	1.574	0.294	0.945
	X	1.124	0.433	1.280

Table A2. CLIP prompt comparison in the AGD20k-Seen scenario. {action} represents the action labels.

Method	Prompt	KLD	SIM	NSS
CLIP	{action}	1.826	0.242	0.522
	"an item to" {action} "with"	1.774	0.250	0.640
ClearCLIP	{action}	1.672	0.277	0.795
	"an item to" {action} "with"	1.574	0.294	0.945

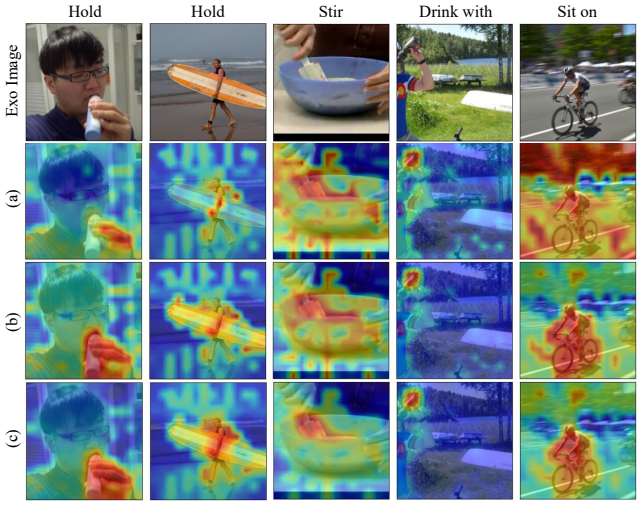


Figure A1. Visualization of object affinity map for exocentric image, with various kinds of prompt. (a): {action}, (b): "an item to" {action} "with", (c): multiplication of "an item to" {action} "with" and "a person" {action} "an item".

B. Object Affinity Map

In this section, we provide a detailed explanation of how the object affinity map A is obtained. Using ClearCLIP [2], we apply different strategies to infer object affinity maps for egocentric and exocentric images.

For the egocentric affinity map, we calculate the similarity between the egocentric image and action-prompted queries. The action-prompted queries are created by augmenting the action label with a fixed prefix, "an item to", and a postfix, "with". For example, the action label "catch" is augmented as "an item to catch with". However, when the

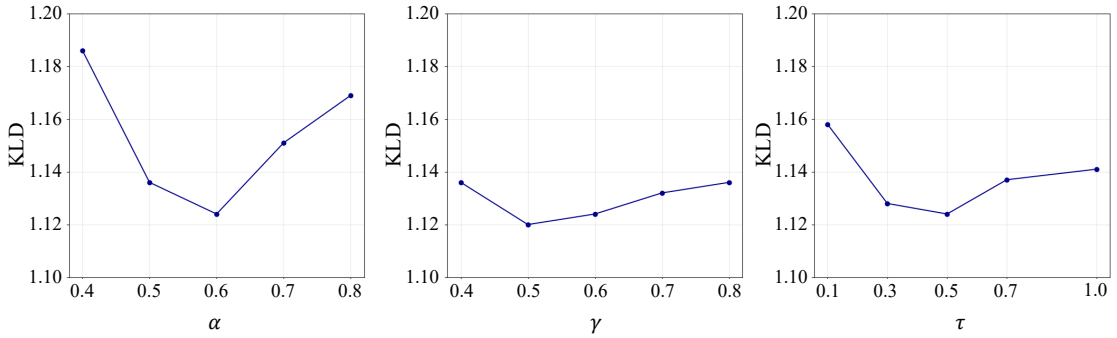


Figure A2. Ablation studies of various hyperparameters. The X-axis denotes the value of each hyperparameter, the Y-axis shows the KLD performance.

action label already ends with "with", such as "brush with" or "cut with", the postfix "with" is not added. The impact of action-prompted queries is shown in Tab. A2.

On the other hand, the object affinity map for exocentric images is generated using two prompting methods to focus primarily on the object parts involved in the interaction within the exocentric image, as shown in Fig. A1. To identify objects in exocentric images, we first use the same action-prompted queries as those applied to egocentric images, as shown in row (b) of Fig. A1. However, we observe that the activation is widely distributed across the foreground objects. To address this, we additionally utilize entity-prompted queries to localize the entity interacting with the objects. We hypothesize that the intersection of the action-prompted and entity-prompted queries will yield a more accurate localization map compared to a simple similarity map derived solely from action labels. The entity-prompted query is structured with the prefix "a person" and the postfix "an item". For example, the action label "catch" is augmented as "a person catch an item". Yet, the similarity map obtained using the entity-prompted query may not fully capture the object parts, as the focus is on the entity in the sentence. To address this, we apply local average pooling, which smooths the activation of each patch by averaging it with nearby patches. Finally, we combine the affinity maps generated from the action- and entity-prompted queries by multiplying them to produce the object affinity map for exocentric images in row (c).

C. Hyperparameter Ablation

We study the impact of thresholds α and γ which control the reliability of selected affordable parts. The threshold α determines whether the part segment within objects in exocentric images corresponds to the desired object part, while γ is used to binarize object affinity map of both egocentric and exocentric images into the foreground targets and the background. Performance comparisons for varying α and γ are illustrated in Fig. A2. Our results indicate that α, used for selecting reliable clusters (groups of pixels), is more sensitive than γ. However, both thresholds consistently achieve optimal performance within the range of 0.5 to 0.6. In this work, we set α and γ to 0.6.

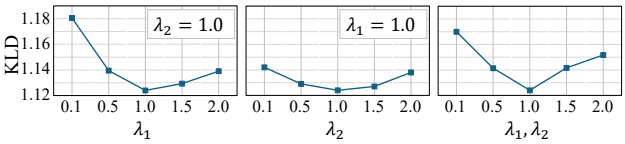


Figure A3. Study on loss coefficients. λ_1 and λ_2 are coefficients for prototypical and pixel contrastive learning, respectively. We vary each coefficient while keeping the others fixed at their default value of 1 and also examine their impact when adjusted simultaneously.

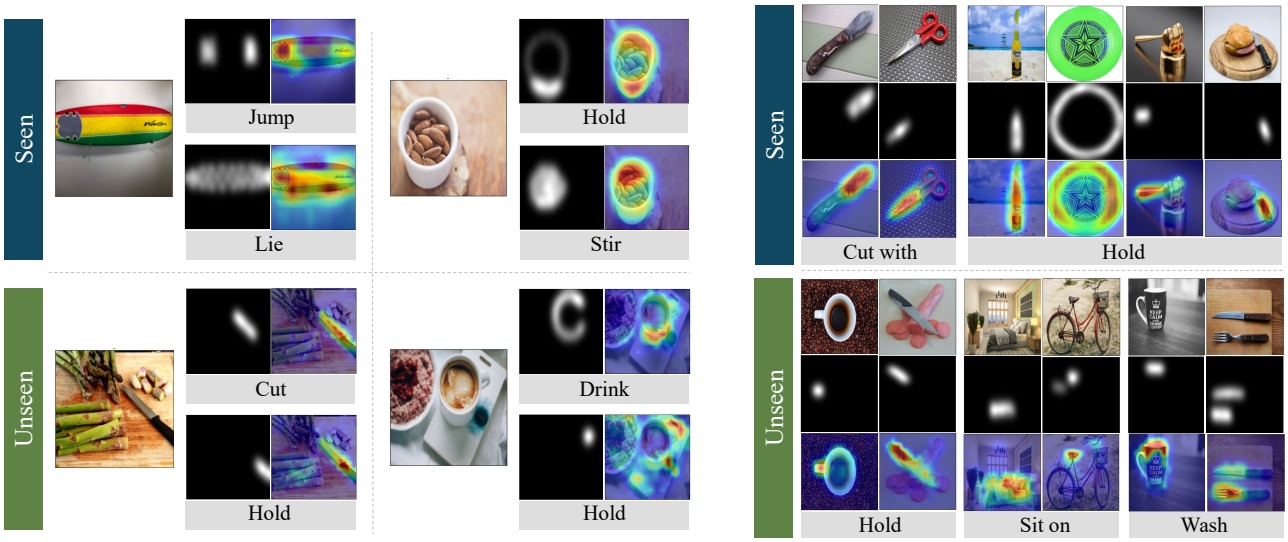
Additionally, we examine the effects of varying τ, the scaling parameter used in both prototypical and pixel contrastive losses. Results are shown on the right side of Fig. A2. In this work, we set τ to 0.5 as it outcomes the best result.

Although the performance slightly decreases when adjusting our hyperparameters, our results demonstrate the robustness of the framework. In particular, our model consistently achieves state-of-the-result performances regardless of hyperparameters α, γ, and τ.

Study on loss coefficients are in Fig. A3. As shown, our default value of 1 yields its best result. Nevertheless, our proposed approach consistently outperforms baselines by a significant margin, demonstrating its robustness and insensitivity to extensive parameter tuning.

D. Bias on Object and Affordance Classes

Objects can be involved in various actions, and likewise, different affordance classes may occur across diverse objects. This presents a particular challenge in weakly supervised affordance grounding, where the distinctions between classes are not explicitly provided. In Fig. A4, we examine how our proposed approach performs under such scenarios. First, Fig. A4 (a) illustrates the prediction results when different affordance classes are queried for the same object class. While



(a) Different affordance regions within the same object class

(b) Same affordance regions across different object classes

Figure A4. Visualization of the test image, ground-truth label, and our prediction on AGD20K dataset.

Table A3. Comparison results between DINO attention map and CLIP affinity map to measure pIoU.

Dataset-Scenario	Method	KLD↓	SIM↑	NSS↑
AGD20K-Seen	DINO-attn	1.124	0.433	1.280
	CLIP-obj.	1.126	0.435	1.273
AGD20K-Unseen	DINO-attn	1.243	0.405	1.368
	CLIP-obj.	1.257	0.398	1.360

the predictions are not perfectly accurate, the model still exhibits meaningful distinctions between affordance classes despite the absence of explicit class-level cues. Fig. A4 (b) further visualizes how well the model generalizes affordance understanding across diverse object classes, demonstrating notably consistent performance. These results support that our strategy effectively minimizes biases toward specific object–affordance pairings, promoting robust affordance predictions.

E. DINO Attention Map for Prototype Selection

In prototype generation for prototypical contrastive learning, we utilize the self-attention map from DINO to measure pIoU, which allows us to select the most suitable prototype among three candidates and perform part-level learning. We emphasize that the DINO attention map can be replaced by any alternative capable of identifying the main object within egocentric images. To validate this flexibility, we conduct experiments using the CLIP affinity map as an alternative, applying a specific threshold (0.75) to distinguish foreground from background regions. Table A3 compares the results obtained using DINO attention maps and CLIP affinity maps, demonstrating the robustness and versatility of our method.

F. Additional Qualitative Results

Additional qualitative results in comparison to baseline methods are depicted in Fig. A5 and Fig. A6. Particularly, Fig. A5 illustrates the results in the seen domain, while Fig. A6 focuses on the unseen domain. As observed, we find that our proposed approach consistently demonstrates more accurate results than previous works.

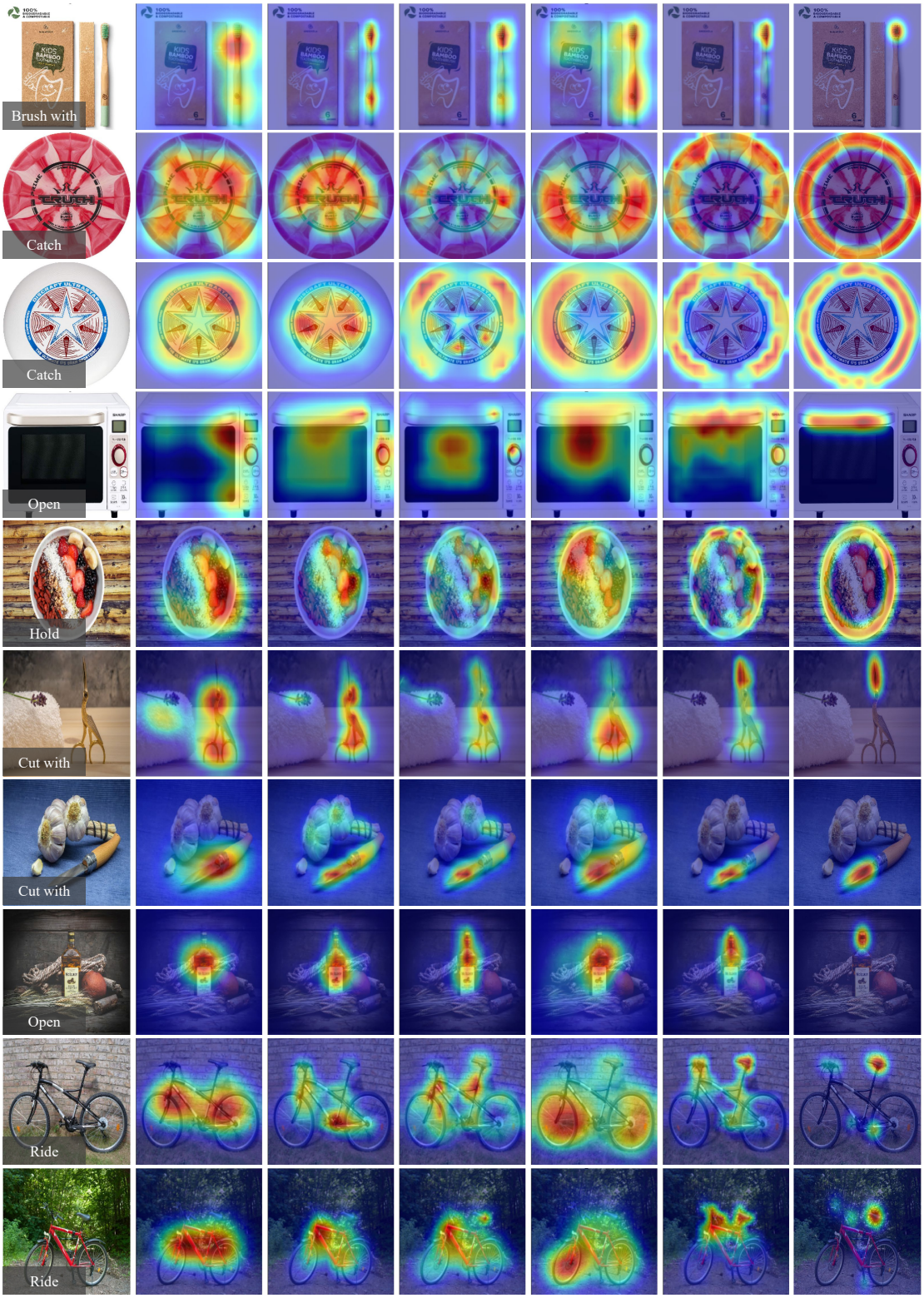


Figure A5. Affordance grounding results of our approach and other methods in the seen domain.

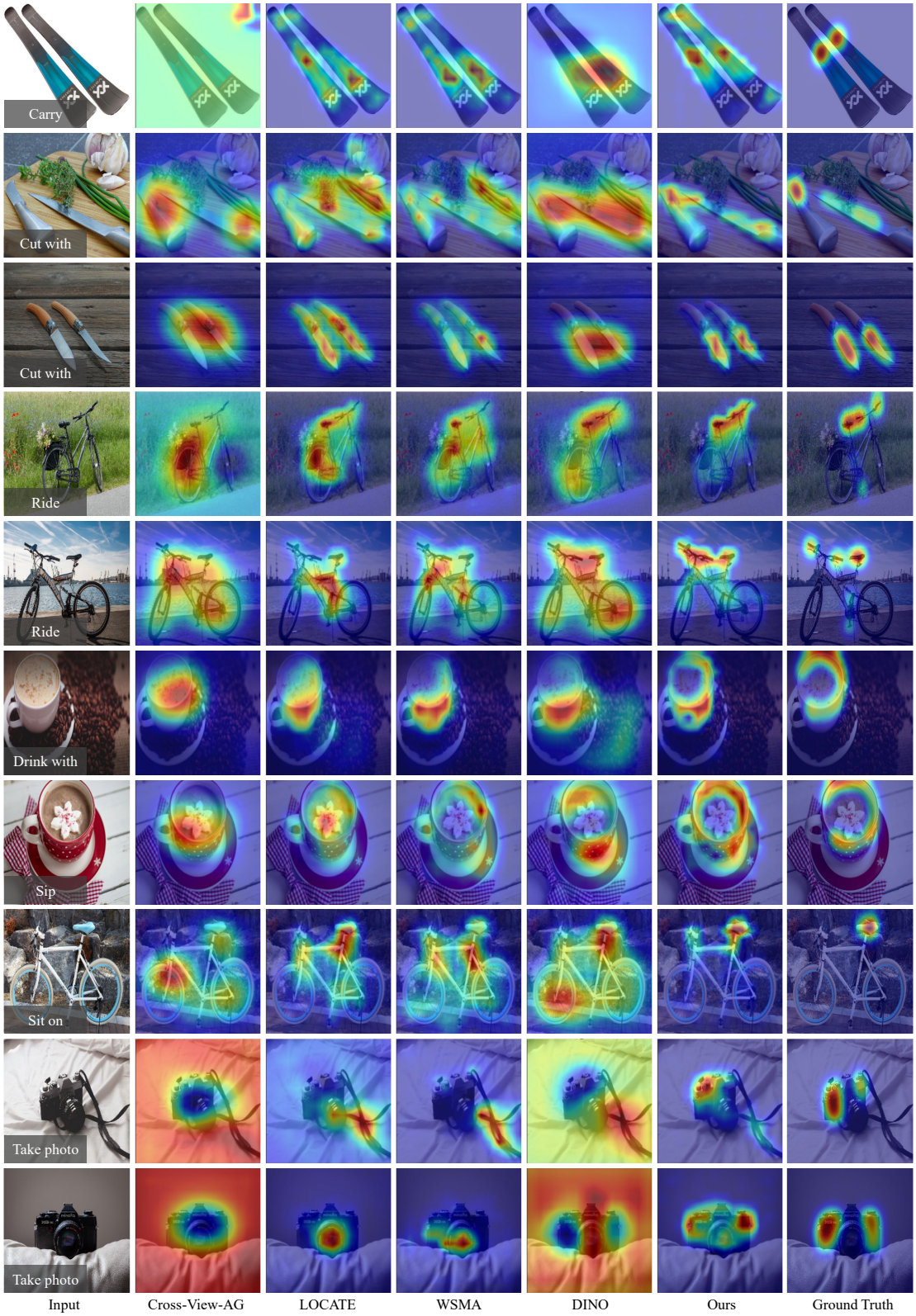


Figure A6. Affordance grounding results of our approach and other methods in the unseen domain.

References

[1] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, pages 381–389. IEEE, 2018. 1

[2] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *ECCV*, 2024. 1

[3] 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

[4] 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

[5] Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5908–5915. IEEE, 2017. 1

[6] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *European Conference on Computer Vision*, pages 315–332. Springer, 2025. 1

[7] Lingjing Xu, Yang Gao, Wenfeng Song, and Aimin Hao. Weakly supervised multimodal affordance grounding for egocentric images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6324–6332, 2024. 1