Experimental Results

Datasets and Evaluation Metrics

To assess the efficacy of our methodology, we employed two distinct datasets: ADE-20K (Luo et al. 2022) and HICO-IIF. Comprehensive details regarding the ADE-20K dataset have been expounded upon in the main body of the text. In this segment, we will delve into the finer details of the HICO-IIF dataset.

HICO-IIF comprises two datasets, namely HICO-DET (Chao et al. 2018) and IIT-AFF (Nguyen et al. 2017). HICO-DET is designed explicitly for detecting Human-Object Interactions (HOI) within images, whereas IIT-AFF encompasses ten object categories and nine affordance categories. Given that our approach necessitates both egocentric and exocentric images, we selected exocentric images from the HICO-DET dataset and egocentric images from the IIT-AFF dataset. Subsequently, these selections formed the composite dataset, HICO-IIF, encompassing ten affordance classes (cut_with, drink_with, hold, open, pour, sip, stick, stir, swing, type_on) and seven object categories (knife, bottle, cup, bowl, tennis_racket, keyboard, laptop). HICO-IIF comprises a training set of 4383 images and a test set of 1498 images. The creation of this dataset is aimed at evaluating the performance of models when trained with a relatively limited dataset size.

Regarding the evaluation metrics, we have opted not to utilize commonly employed segmentation evaluation metrics such as Intersection over Union (IOU). This decision arises from the distinction between the dataset's ground truth and conventional binary masks, as the ground truth in the dataset represents probability distributions of affordance regions. Following precedent studies, we have thus employed the following three metrics: Kullback-Leibler Divergence (KLD) (Bylinskii et al. 2018), Similarity (SIM) (Swain and Ballard 1991), and Normalized Scanpath Saliency (NSS) (Peters et al. 2005).

Beginning with KLD, its computation is as follows:

$$KLD\left(H^{'}, G^{'}\right) = \sum_i G_i^{'} log\left(\frac{G_i^{'}}{H_i^{'}}\right). \tag{A1}$$

H represents the heatmap generated by the model, while $H^{'} = H/\sum_i H_i$. Correspondingly, G represents the ground truth, with $G^{'} = G/\sum_i G_i$. The primary objective of Kullback-Leibler Divergence (KLD) is to quantify the disparity between the distributions of these two entities.

Subsequently, we turn our attention to Similarity (SIM), which serves to gauge the likeness between the two distributions.

$$SIM\left(H^{'}, G^{'}\right) = \sum_i min(H_i^{'}, G_i^{'}). \tag{A2}$$

Lastly, we examine Normalized Scanpath Saliency (NSS), which is employed to quantify the correspondence between the predicted heatmap and the ground truth. Given the heatmap (H) and the binary ground truth (GT), the initial step involves computing intermediary values.

$$N = \sum_i GT_i, \bar{H} = \frac{H - \mu}{\sigma}, \tag{A3}$$

where μ represents the mean of H and σ denotes the standard deviation of H. Building upon the formulation above, we have:

$$NSS(H, GT) = \frac{1}{H}\sum_i \bar{H} \times GT_i \tag{A4}$$

Implementation Details

We employ the pre-trained DINO-ViT-S as the backbone network for both the egocentric and exocentric branches, and we freeze the weights during training. The backbone network for the text branch is the pre-trained text encoder from CLIP. In the experiments, we resize the images to 256×256 and then crop them to 224×224 as the final input size for the network. Lastly, we train for 13 epochs, and the hyperparameters λ_{cls}, λ_{clip}, λ_d, λ_{l_rela} are set to 1, 1, 0.5, and 0.5, respectively. In the paper, all quantitative results presented are averaged over three repetitions under the same configuration.

Furthermore, the exocentric branch requires the simultaneous input of n exocentric images. In our experiments, we set n to 3. The threshold used during the inference phase is set to 0.2. We conducted extensive experiments for these two hyperparameters, as shown in Figure A1. First, analyzing the parameter n, we tried setting n to 1, 2, 3, 4, and 5, respectively. From the observations in Figure A1, we can deduce that on both datasets of ADE20K, there is minimal difference between n being 2 or 3. However, it is evident that for the HICO-IIF dataset, the performance with n=3 surpasses that with n=2. Additionally, while n=4 achieved the best results in the two datasets(ADE20K-Seen, HICO-IIF), it exhibited subpar performance in the ADE20K-Unseen dataset. Consequently, we opted to set n as 3. This phenomenon can be attributed to the fact that when n is less than 3, there is a lack of exocentric images to capture common features. Conversely, when n exceeds 3, an excessive introduction of individual variations leads to performance degradation.

Regarding the threshold parameter, we conducted multiple experiments. Upon observing the experimental results, it becomes apparent that for KLD, the differences among the results of various threshold settings are negligible. However, in terms of SIM and NSS, one tends to increase while the other decreases with increasing threshold values. In the end, a balanced compromise was chosen, leading to the adoption of an intermediate value of 0.2.

Ablation Study

In the main body of the paper, we mentioned the ablation experiments concerning L_{g_rela} and L_{l_rela}. Here, we further elaborate on this experimentation. As depicted in Figure 2(a), L_{g_rela} is detailedly described, revealing that L_{g_rela} is directly computed from classification scores. In contrast, L_{l_rela}, as illustrated in Figure 2(b), involves an initial computation of the heatmaps using CAM, followed by the subsequent calculation of action correlations.

To further validate the roles of L_{g_rela} and L_{l_rela}, we present visualizations of experimental results in Figure A3. By observing the visual outcomes with the inclusion of L_{g_rela} or L_{l_rela}, we can deduce that the experimental performance of L_{l_rela} is notably superior to that of L_{g_rela}.

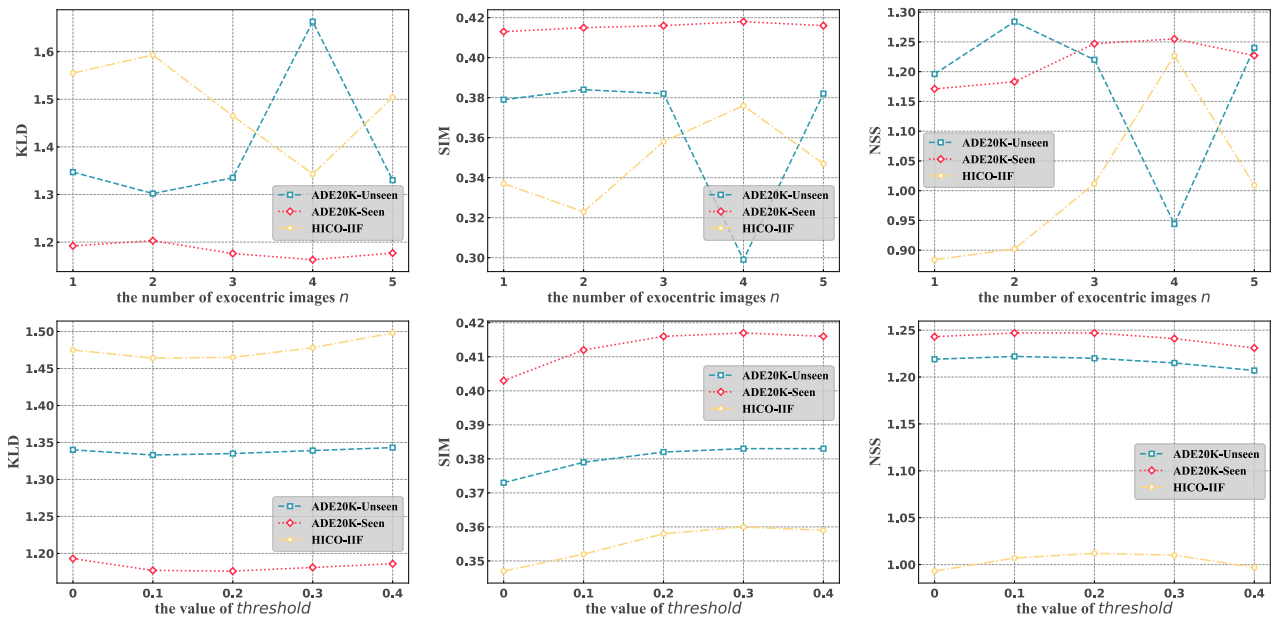
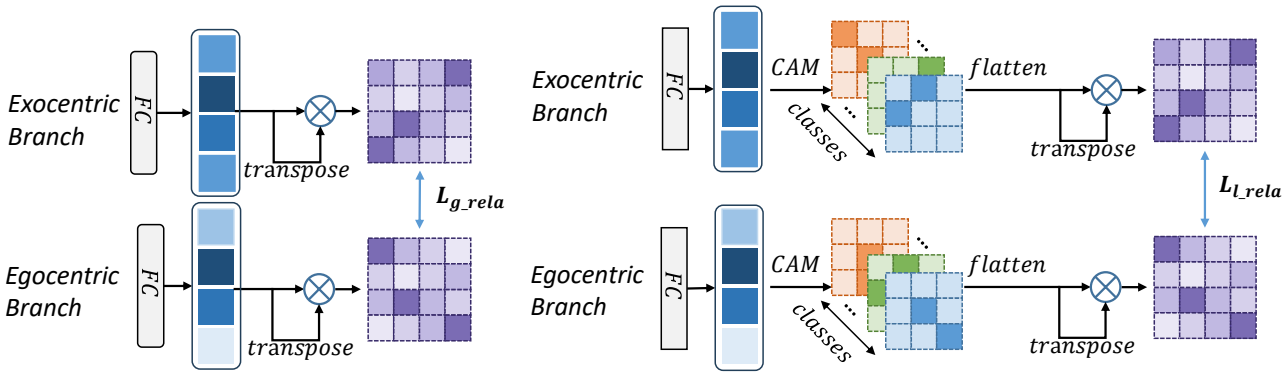


Figure A1: Experimental Investigation of Hyperparameters. A smaller value of the Kullback-Leibler Divergence (**KLD** ↓) indicates superior performance, while larger numerical values of Similarity (**SIM** ↑) and Normalized Scanpath Saliency (**NSS** ↑) correspond to more optimal results.



(a) L_{g_rela} (b) L_{l_rela}

Figure A2: Description of L_{g_rela} and L_{l_rela}. L_{g_rela} solely utilizes classification scores for computation. L_{l_rela} first computes heatmaps before calculating correlations.

For instance, under the affordance label "catch," the model augmented with L_{g_rela} manages to capture the position of the soccer ball, yet the results exhibit a certain lack of precision. In contrast, L_{l_rela} demonstrates a more favorable outcome. We infer that this is due to the loss of significant information when using classification scores directly for computation, as heatmaps encapsulate more detailed information than what is contained within classification scores.

Limitations

Here, we aim to provide a more comprehensive delineation of the limitations inherent in our approach. Firstly, due to dataset biases, there is a paucity of images depicting complex interactions, such as scenarios where a single image contains objects of different categories yet shares the same affordance label. Secondly, we have observed that substantial variations in individual behaviors can yield less satisfactory experimental results. For instance, within our experimental outcomes, the affordance region associated with the action "hold" for the object "book" nearly encompasses the entire book. We surmise that this phenomenon arises from substantial variability in how individuals hold a book, with some grasping only the lower portion while others opt to grip the book's side. Consequently, under such circumstances, the model fails to capture the common features of this action, leading to the affordance region encompassing the en-

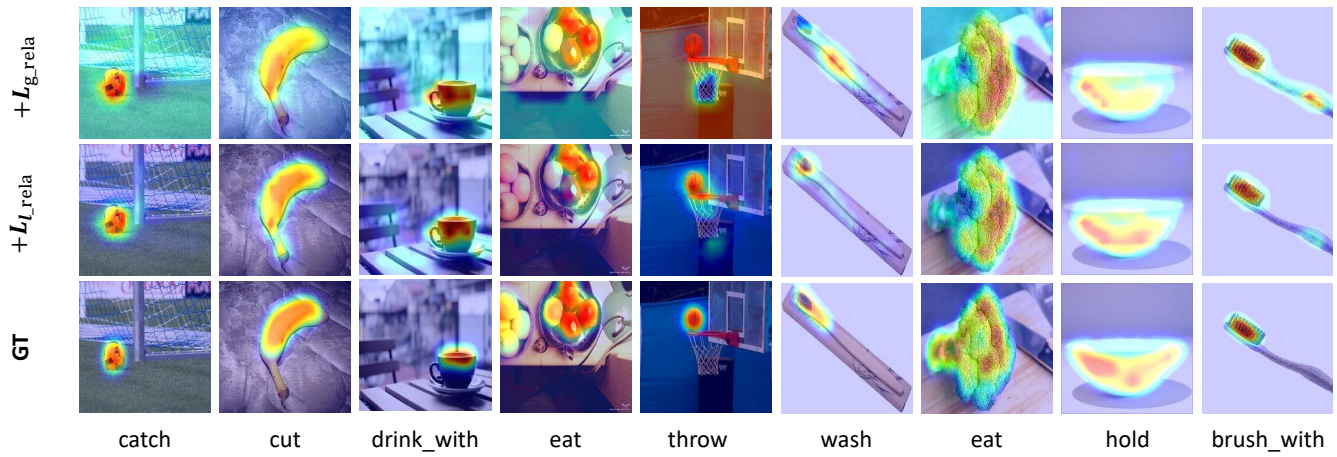


Figure A3: Comparison of Visualization Results with the Inclusion of L_{g_rela} or L_{l_rela}

tire book. To alleviate the aforementioned constraints, we intend to further expand the dataset and conduct finer-grained categorization of exocentric images that exhibit highly diverse variations in affordance regions.

References

Bylinskii, Z.; Judd, T.; Oliva, A.; Torralba, A.; and Durand, F. 2018. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3): 740–757.

Chao, Y.-W.; Liu, Y.; Liu, X.; Zeng, H.; and Deng, J. 2018. Learning to detect human-object interactions. In *2018 ieee winter conference on applications of computer vision (wacv)*, 381–389. IEEE.

Luo, H.; Zhai, W.; Zhang, J.; Cao, Y.; and Tao, D. 2022. Learning affordance grounding from exocentric images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2252–2261.

Nguyen, A.; Kanoulas, D.; Caldwell, D. G.; and Tsagarakis, N. G. 2017. 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)*, 5908–5915. IEEE.

Peters, R. J.; Iyer, A.; Itti, L.; and Koch, C. 2005. Components of bottom-up gaze allocation in natural images. *Vision research*, 45(18): 2397–2416.

Swain, M. J.; and Ballard, D. H. 1991. Color indexing. *International journal of computer vision*, 7(1): 11–32.