

Closed-Loop Transfer for Weakly-supervised Affordance Grounding

Jiajin Tang^{1*}, Zhengxuan Wei^{1*}, Ge Zheng¹, Sibe Yang^{2†}

¹ShanghaiTech University

²School of Computer Science and Engineering, Sun Yat-sen University

{tangjj, weizhx2022}@shanghaitech.edu.cn yangsb3@mail.sysu.edu.cn

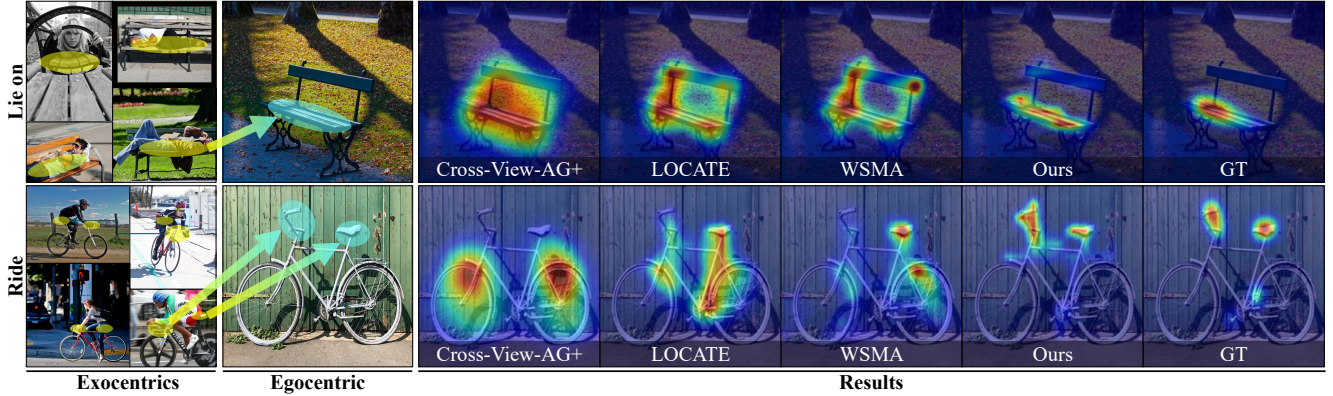


Figure 1. Visualization samples for lie on and ride affordances on egocentric images, comparing state-of-the-art methods with ours.

Abstract

Humans can perform previously unexperienced interactions with novel objects simply by observing others engage with them. Weakly-supervised affordance grounding mimics this process by learning to locate object regions that enable actions on egocentric images, using exocentric interaction images with image-level annotations. However, extracting affordance knowledge solely from exocentric images and transferring it one-way to egocentric images limits the applicability of previous works in complex interaction scenarios. Instead, this study introduces LoopTrans, a novel closed-loop framework that not only transfers knowledge from exocentric to egocentric but also transfers back to enhance exocentric knowledge extraction. Within LoopTrans, several innovative mechanisms are introduced, including unified cross-modal localization and denoising knowledge distillation, to bridge domain gaps between object-centered egocentric and interaction-centered exocentric images while enhancing knowledge transfer. Experiments show that LoopTrans achieves consistent improvements across all metrics on image and video benchmarks, even handling challenging scenarios where object interaction regions are fully occluded by the human body. Code is available at <https://github.com/nagara214/LoopTrans>.

1. Introduction

The term “affordance,” first introduced by J.Gibson [16], is later formalized in computer vision and robotics to typically describe the “action possibilities” offered by objects [2, 34, 38, 56], such as a knife affords cutting or a bicycle affords riding. Affordance grounding [8, 12, 13, 35, 37, 49] further refines this by not only predicting the actions an object can afford but also pinpointing the specific regions that enable those actions, e.g., a bicycle’s handlebars for pushing and both its handlebars and seat for riding (see Fig 1). In contrast to most vision perception systems [6, 10, 20, 47, 57] that primarily focus on how objects appear [46, 48, 54], such as instance and part segmentation, affordance grounding emphasizes how objects function. This is essential for embodied intelligent agents to actively interact with and use objects in the real world [1, 14, 18, 44, 50], while also facilitating downstream tasks such as object manipulation [3, 19] and human-object interaction [5, 17].

Humans can infer precise affordance grounding on objects across diverse actions and environments, even performing unfamiliar interactions with novel objects, simply by observing others interact with them. To mimic this learning process, we follow a practical weakly-supervised affordance grounding setting [27, 29, 31, 36], learning object affordances from human-object interaction images or videos without using any pixel-level affordance annotations. As shown in Fig 1, given exocentric human-object interaction

* Equal contribution. † Corresponding author is Sibe Yang.

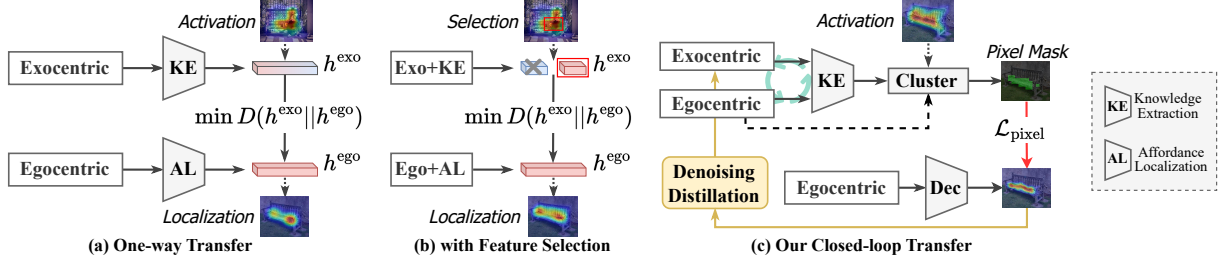


Figure 2. Comparison of (a) one-way exo-to-ego transfer framework [29, 52], (b) one-way transfer with feature selection [27], and (c) our closed-loop transfer framework, LoopTrans. KE refers to extracting human-object interaction knowledge from exocentric images, while AL refers to localizing affordance regions of objects in egocentric images.

images and object images with corresponding interaction labels (e.g., lie on) during training, the goal in inference is to ground the affordances of each interaction label on the target egocentric object image.

Two essential yet challenging cores of affordance grounding are (1) *extracting affordance knowledge from exocentric interactions* and (2) *transferring it to egocentric localization*. **On one hand, for affordance knowledge extraction**, current methods [22, 27, 52] primarily *rely solely on exocentric interaction images*, using CAM [55] to generate activation maps. CAM is a classic approach for localizing category-relevant regions using image-level labels by highlighting the most discriminative areas for classification. These methods extract and represent interaction knowledge using image features from the generated activation maps, as shown in the “exocentric+KE” branch in Fig 2a-b. However, the diversity and complexity of interaction scenes make exocentric images alone insufficient for precise affordance activation, leading to vague, broad regions that may include background or human body parts in simple scenarios, and scattered attention in more complex interactions (see Appendix). **On the other hand, transferring knowledge to egocentric localization** presents notable challenges due to *the significant domain gap between exocentric and egocentric images, as well as the increased difficulty of transferring knowledge from occluded objects in interaction regions*. First, exocentric images are often cluttered with small and potentially occluded interaction regions, while egocentric images are clear and object-centered. Most methods [29, 52] that constrain appearance similarity of affordance regions between views fail to address knowledge transfer in complex interaction scenes with large domain gaps, often resulting in mislocalization of non-affordance object parts, as illustrated in the “egocentric+AL” branch in Fig 2a. Second, while recent advances [27] propose part selection (Fig 2b) on egocentric images to mitigate partial occlusion issues, they still rely on exocentric interaction appearance for guidance, limiting applicability in fully occluded interactions, such as “lie on” and “ride,” as shown in Fig 1.

In this paper, we propose a novel closed-loop knowledge transfer framework, LoopTrans, to address these chal-

lenges. Unlike the conventional one-way, non-closed-loop pipeline from exocentric interaction to activation and then to egocentric localization (see the arrow flow in Fig 2a and 2b), LoopTrans introduces an innovative closed-loop mechanism that enables egocentric localization to feed back into knowledge attention (Fig 2c). LoopTrans’s dual design naturally addresses key challenges: (1) Egocentric localization on simple, object-centered egocentric images provides clear localization without interference from the background or human body. It can naturally be used to refine knowledge activation in complex exocentric images (see yellow arrow flow in Fig 2c), focusing more precisely on potential affordance regions and resolving coarse and scattered activation issues. (2) In turn, the shared and unified affordance knowledge activation learning across exocentric and egocentric modalities (see green arrow flow in Fig 2c) not only reduces the domain gap but also enables direct use of egocentric activation to transfer affordance knowledge (see red arrow flow in Fig 2c), bypassing exocentric-to-egocentric appearance transfer and naturally overcoming transfer difficulties caused by occlusions in egocentric interactions.

Specifically, LoopTrans works as follows in a closed-loop process. First, interaction $\rightarrow$ activation: Based on image-level affordance labels, LoopTrans learns a unified classifier for both egocentric and exocentric images, using a shared CAM to highlight both their affordance knowledge activation. Exocentric images provide interaction knowledge, while egocentric images help CAM activation focus more on the object, reducing background and human body interference. More importantly, this shared CAM can effectively identify affordance regions in egocentric objects even when interactions in exocentric images are fully occluded. Second, activation $\rightarrow$ localization: Thanks to the shared knowledge attention, LoopTrans refines egocentric localization directly through egocentric activation, rather than relying on exocentric-to-egocentric appearance transfer as in previous methods, thereby reducing the challenges of cross-domain knowledge transfer. LoopTrans directly selects clustered object parts in egocentric images based on egocentric activation, refining coarse activation regions into precise affordance localization. Third, localization $\rightarrow$ activation: LoopTrans leverages more precise localization to

improve knowledge activation for both egocentric and exocentric images. Given the noisier background of an exocentric image compared to an egocentric one, we introduce a novel denoising distillation method. Using exocentric activation as an anchor, we align the egocentric activation with this anchor while distancing its noise, thereby sharpening the focus on more precise and complete affordance regions and effectively segregating background noise. These three components form the end-to-end LoopTrans, establishing a closed-loop knowledge transfer from activation to localization and back, enhancing the precision and transfer of affordance knowledge.

In summary, our main contributions are multi-fold:

- We are the first to propose a closed-loop knowledge transfer mechanism for affordance grounding, where exocentric knowledge activation and egocentric localization mutually enhance each other in a closed loop.
- We propose a shared CAM that enables unified knowledge activation using both exocentric and egocentric images. It not only leverages object-centered egocentric images for clearer activation but also addresses challenges in cross-domain transfer.
- We introduce a novel denoising distillation mechanism that transfers egocentric localization back into the shared CAM, reducing the impact of background noise from exocentric images and focusing activation on object regions.
- Our LoopTrans achieves significant and consistent improvements in weakly-supervised affordance grounding across all metrics on both image and video benchmarks, demonstrating its effectiveness and robustness.

2. Related Work

Learning to Localize from Weakly Supervision. Weakly supervised localization tasks rely on image-level labels or keypoint annotations to guide object localization and segmentation. Class Activation Map (CAM) [55], a foundational approach, highlights discriminative regions linked to image-level labels but often misses less salient areas. To address this, methods have incorporated augmented training [33, 51], semantic [15, 23, 53] and spatial priors [41] to improve completeness and accuracy of highlighted regions. Recently, weak supervision has been extended to affordance localization, where models learn from exocentric images or videos of human-object interactions and transfer activation maps to egocentric images for localization [27, 29]. Despite using interaction data, the task remains challenging due to the diversity of exocentric scenes. Some studies decompose interactions into shared affordance features and individual biases [29, 31], extract hand actions to support affordance inference [30], or leverage knowledge priors like CLIP [43] to refine affordance grounding with text-based cues [52].

Visual Affordance Grounding focuses on identifying image or video regions likely corresponding to specific hu-

man interactions. Early methods relied on small datasets with pixel-level annotations [8, 12, 24, 35], using object geometry [35] or appearance [24] to infer affordances. More recent approaches favor weakly supervised methods, which are more feasible in real-world scenarios. Studies [45] demonstrate that effective affordance localization can be achieved using minimal keypoint annotations. Later works [29, 36] incorporate human-object interaction (HOI) priors, relying solely on affordance category labels to reduce annotation costs. However, the diversity and complexity of HOI scenes introduce new challenges for model learning. To address these, [30] uses hand cues to reduce action ambiguity, while [29] decomposes interactions to capture shared affordances across diverse contexts. Approaches like [52] and [27] employ localized knowledge transfer to filter out irrelevant backgrounds, with [27] further adding a part selection to isolate specific object-part features. Recognizing the limitations of traditional activation maps in complex scenes, [52] and [22, 42] incorporate auxiliary activations from CLIP and large language models for improved support.

3. Preliminary

Problem Definition. Given pairs of exocentric and egocentric images with image-level affordance labels, weakly supervised affordance grounding [29, 31, 36] aims to extract interaction knowledge from exocentric images and accurately locate object parts corresponding to affordance labels in egocentric images. Specifically, given a pair of exocentric and egocentric images, $\{I^{\text{exo}}, I^{\text{ego}}\}$ along with N affordance categories $\{c_1, \dots, c_N\}$, the objective is to generate egocentric affordance activation maps $\mathcal{G}^{\text{ego}} \in \mathbb{R}^{H \times W \times N}$, corresponding to the N categories. Here, H and W are the height and width of feature maps, respectively.

Visual Feature Extraction. Self-supervised ViT DINO [4] is employed to extract image patch features, denoted as $\mathcal{F}^{\text{exo}} \in \mathbb{R}^{H \times W \times C}$ and $\mathcal{F}^{\text{ego}} \in \mathbb{R}^{H \times W \times C}$, where C is the feature dimension. The interaction-focused exocentric feature $\mathcal{F}^{\text{exo}}$ encodes affordance knowledge by highlighting interaction regions and types, while the object-centered egocentric feature $\mathcal{F}^{\text{ego}}$ concentrates on object-specific information without background and human interference.

Class Activation Mapping (CAM) [55] provides a mechanism for localizing affordance activation regions through image-level weakly-supervised learning. Given an input feature map $\mathcal{F} \in \mathbb{R}^{H \times W \times C}$, it is first transformed via an MLP followed by a two-layer convolutional block. A subsequent 1×1 convolutional layer with N category-specific kernels generates activation maps $\mathcal{G} \in \mathbb{R}^{H \times W \times N}$, where each kernel’s output corresponds to the activation region for a specific affordance category. Activation maps undergo global pooling to generate activation scores for final class probability prediction. The CAM process is defined as:

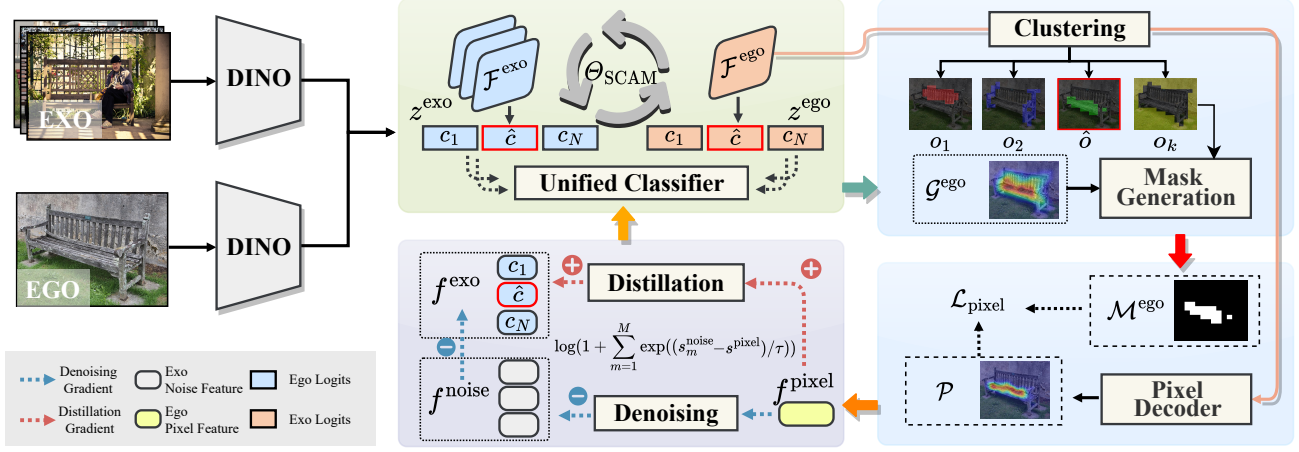


Figure 3. Overall framework of our proposed LoopTrans enables closed affordance knowledge transfer, with the process interaction $\rightarrow$ activation $\rightarrow$ localization $\rightarrow$ activation, spanning both exocentric and egocentric domains.

$$\begin{aligned} \mathcal{G} &= \Theta_{\text{CAM}}(\mathcal{F}; \theta), z = \text{GAP}(\mathcal{G}), \\ \mathcal{L}_{\text{cls}} &= -\sum_{i=1}^N \mathbb{I}(c_i = \hat{c}) \log \sigma(z_i), \end{aligned} \quad (1)$$

where Θ_{CAM} represents the CAM module, θ is the trainable parameters of CAM, $\text{GAP}(\cdot)$ denotes global average pooling, $z \in \mathbb{R}^N$ represents activation scores, with class probabilities from the sigmoid function σ . $\mathbb{I}$ is the indicator function, and $\mathcal{L}_{\text{cls}}$ is the classification loss designed to align the predicted probability with the ground-truth label $\hat{c}$. By directly minimizing this classification loss based on the activation scores, its corresponding activation maps effectively highlight the regions of the affordance category.

One-Way Exo-Ego Transfer. As affordance-based knowledge sources are exocentric and localization targets are egocentric, spanning different image domains, existing methods have predominantly employed a one-way framework that aligns exocentric features to egocentric ones for grounding [22, 27, 29–31, 36, 52]. They use two independent CAM modules to activate affordance regions in exocentric and egocentric images separately, then align the features corresponding to the two activation regions. Specifically, for paired images I^{exo} and I^{ego} sharing the same category $\hat{c}$, paired CAM modules with distinct parameters θ^{exo} and θ^{ego} are employed to produce the corresponding activation maps $\mathcal{G}^{\text{exo}}$ and $\mathcal{G}^{\text{ego}}$. The activation maps are then combined with feature maps through weighted averaging to obtain corresponding features h^{exo} and h^{ego} , which are finally aligned to achieve one-way transfer. The entire computational process is formulated as follows:

$$\begin{aligned} h^{\text{exo}} &= \text{GAP}(\mathcal{R}(\mathcal{G}_{\hat{c}}^{\text{exo}}) \circ \mathcal{F}^{\text{exo}}), \\ h^{\text{ego}} &= \text{GAP}(\mathcal{R}(\mathcal{G}_{\hat{c}}^{\text{ego}}) \circ \mathcal{F}^{\text{ego}}), \mathcal{L}_{\text{align}} = \|h^{\text{exo}} - h^{\text{ego}}\|_2^2, \end{aligned} \quad (2)$$

where $\mathcal{R}(\cdot)$ and $\circ$ denote min-max normalization and the Hadamard product, respectively. $\mathcal{G}_{\hat{c}}^{\text{exo}}$ denotes the activation map of class $\hat{c}$ in $\mathcal{G}^{\text{exo}}$, and $\|\cdot\|_2$ means L2 norm.

4. Method

Overview. Existing weakly supervised affordance grounding methods face two critical limitations: (1) one-way feature alignment heavily depends on exocentric features. However, exocentric activation regions often include human hands or body parts, blending object information with background elements. This prevents egocentric activation features and maps from focusing solely on the object, while isolated CAM modules further amplify view discrepancies. (2) one-way exocentric-to-egocentric transfer underutilizes the object-centric nature of egocentric images, which could also aid exocentric activation. A bidirectional transfer enhances activation consistency and mitigates the domain gap between context-rich exocentric and object-centric egocentric images. Therefore, we propose **LoopTrans**, a closed-loop framework (Fig. 3), which facilitates the knowledge transfer in a loop, including three key stages:

- **Interaction $\rightarrow$ Activation.** We achieve shared interaction knowledge activation and transfer through a unified CAM module jointly trained on both views (Sec 4.1). Our shared activation allows exocentric interaction patterns to directly activate object-centric affordance regions in egocentric images, effectively eliminating background interference caused by explicit feature alignment.
- **Activation $\rightarrow$ Localization.** Leveraging egocentric images’ object-centric nature, we use DINO feature clustering to extract object parts and train a pixel decoder with part-level pseudo-masks to refine coarse activation maps from the previous stage into precise, egocentric affordance localization (Sec 4.2).
- **Localization $\rightarrow$ Activation.** Refined egocentric localization from the second stage is fed back to enhance shared knowledge activation in the first stage, eliminating irrelevant context and directing activation toward affordance-relevant objects via our denoising distillation (Sec 4.3).

4.1. Unified Exo-to-Ego Activation

In this section, we aim to extract interaction knowledge to jointly highlight activation maps from exocentric and egocentric images. Instead of previous one-way grounding methods that process exocentric $\{I^{\text{exo}}\}$ and egocentric $\{I^{\text{ego}}\}$ images separately using distinct CAM modules, we propose Θ_{SCAM} , a Shared CAM module that unifies exo-ego activation through parameter sharing and co-training.

Specifically, while preserving the vanilla CAM architecture [55] (as illustrated in Fig 4a green), Θ_{SCAM} processes both exocentric features $\mathcal{F}^{\text{exo}}$ and egocentric features $\mathcal{F}^{\text{ego}}$ using identical learnable parameters θ , as follows:

$$\begin{aligned} \mathcal{G}^{\text{exo}}, \mathcal{G}^{\text{ego}} &= \Theta_{\text{SCAM}}(\{\mathcal{F}^{\text{exo}}, \mathcal{F}^{\text{ego}}\}; \theta), \\ z^{\text{exo}}, z^{\text{ego}} &= \text{GAP}(\mathcal{G}^{\text{exo}}), \text{GAP}(\mathcal{G}^{\text{ego}}), \\ \mathcal{L}_{\text{cls}} &= -\sum_{i=1}^N \mathbb{I}(c_i = \hat{c}) \log(\sigma(z_i^{\text{exo}}) \cdot \sigma(z_i^{\text{ego}})), \end{aligned} \quad (3)$$

where shared parameters θ enforce cross-view consistency. The classification loss $\mathcal{L}_{\text{cls}}$ maximizes joint confidence $\sigma(z_i^{\text{exo}}) \cdot \sigma(z_i^{\text{ego}})$ for class $\hat{c}$, driving Θ_{SCAM} to align affordance predictions across views. This bidirectional synergy allows exocentric activations to suppress human-body interference using egocentric object cues, while egocentric maps implicitly acquire affordance interaction knowledge from exocentric interaction context.

Notably, to further mitigate background and human noise in exocentric images during shared activation, we introduce multiple noise-absorbing heads into the final activation convolutional layer, as illustrated in Fig 4a purple. These heads isolate non-affordance context and background noise for exocentric images, effectively reducing domain gap-induced interference in cross-view shared activations. The detailed implementation and functionality of these heads are elaborated in Section 4.3.

4.2. Region Activation to Pixel Localization

Leveraging our unified knowledge extraction through shared CAM, LoopTrans can directly generate the egocentric activation map $\mathcal{G}^{\text{ego}}$, thereby alleviating the challenges associated with cross-domain knowledge transfer. However, both weakly-supervised object localization and affordance grounding tasks have long faced a key challenge [15, 26, 29, 53]: CAM highlights only the most salient regions, resulting in activation maps that inadequately cover the entire interaction part. To address this, we transform the coarse and even ambiguous knowledge activation maps into clearly defined object part pseudo-masks (Sec 4.2.1). Furthermore, we train a pixel-level affordance decoder Θ_{pixel} that utilizes the pseudo-masks to generate accurate affordance localization. (Sec 4.2.2).

4.2.1. Activation to Object Part

We leverage the properties of the self-supervised ViT DINO [4] to partition egocentric images into semantically

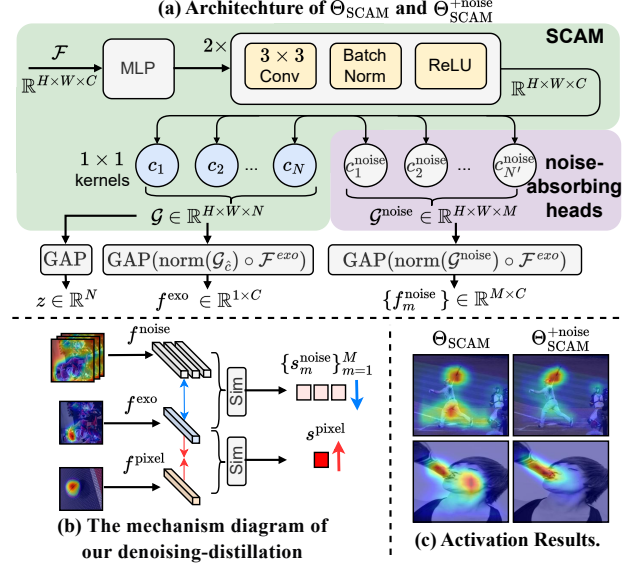


Figure 4. (a) Network architectures of Θ_{SCAM} and our proposed $\Theta_{\text{SCAM}}^{+\text{noise}}$; (b) The mechanism of our denoising-distillation; (c) Exocentric activation results of Θ_{SCAM} and $\Theta_{\text{SCAM}}^{+\text{noise}}$.

distinct parts through unsupervised clustering, subsequently generating pseudo-labels that encompass complete object parts based on the activation map $\mathcal{G}^{\text{ego}}$. Specifically, given the egocentric image feature $\mathcal{F}^{\text{ego}}$ that mainly contains objects, we first apply unsupervised clustering to divide them into K parts $\{o_1, \dots, o_K\}$, where $o_k \in \{0, 1\}^{H \times W}$. Each part o_k is assigned a clear and distinct semantic region, such as bench $\rightarrow$ backrest, armrest, seat, and background, as illustrated in the top right corner of Fig 3. Next, we select the part with the highest Intersection over Union (IoU) score relative to the activation map $\mathcal{G}^{\text{exo}}$ as the accurate localization result, which serves as the pseudo mask $\mathcal{M}^{\text{exo}}$. The detailed computational process is as follows:

$$\mathcal{M}^{\text{exo}} = \arg \max_{o_k \in \{o_1, \dots, o_K\}} \text{IoU}(o_k, \mathbb{I}(\mathcal{R}(\mathcal{G}_c^{\text{exo}}) \geq \mu)), \quad (4)$$

where $\mathcal{G}_c^{\text{exo}}$ means the activation map for class $\hat{c}$, and μ is the threshold used to filter foreground, and $\mathbb{I}(\cdot \geq \mu)$ is an indicator function that returns 1 when the value is greater than or equal to the threshold μ , and 0 otherwise.

4.2.2. Object Part to Localization

The objective of this section is to train a pixel-level affordance decoder, Θ_{pixel} , which employs region-complete pseudo mask, $\mathcal{M}^{\text{exo}}$, as supervision to learn the final affordance localization in exocentric images. Θ_{pixel} has the same architecture with Θ_{CAM} but has its own learnable parameters. Given the feature $\mathcal{F}^{\text{exo}}$ of the input exocentric image and the corresponding pseudo mask $\mathcal{M}^{\text{exo}}$, we feed $\mathcal{F}^{\text{exo}}$ into Θ_{pixel} to obtain the per-pixel localization map $\mathcal{P} \in \mathbb{R}^{H \times W \times N}$. Here, N represents the number of affordance categories, $0 \leq \mathcal{P}_{i,j,c} \leq 1$ represents the probability that the pixel at position (i, j) of the image corresponds to

the localization of the c -th affordance class, and we denote $\mathcal{P}_{\hat{c}} \in \mathbb{R}^{H \times W}$ as the probability map corresponding to affordance category $\hat{c}$. We supervise the pixel-level affordance decoder using $\mathcal{L}_{\text{pixel}}$, which combines dice [28] with MSE losses, as follows:

$$\begin{aligned} \mathcal{P} &= \Theta_{\text{pixel}}(\mathcal{F}^{\text{ego}}; \theta), \mathcal{L}_{\text{mse}} = \frac{1}{HW} \|\mathcal{P}_{\hat{c}} - \mathcal{M}^{\text{ego}}\|^2, \\ \mathcal{L}_{\text{dice}} &= 1 - \frac{2 \sum_{i,j} \mathcal{P}_{i,j,\hat{c}} \cdot \mathcal{M}_{i,j}^{\text{ego}}}{\sum_{i,j} \mathcal{P}_{i,j,\hat{c}} + \sum_{i,j} \mathcal{M}_{i,j}^{\text{ego}}}. \end{aligned} \quad (5)$$

With per-pixel supervision, LoopTrans achieves precise and complete localization region for affordances. More importantly, comprehensive supervision at the object-region level ensures consistency in knowledge transfer across images, fundamentally addressing the information asymmetry during transferring caused by domain difference between exocentric and egocentric.

4.3. Ego-to-Exo Denoising Distillation

Precise affordance activation in exocentric images is hindered by scene complexity and small object scales, where conventional CAMs tend to over-activate human-centric regions while overlooking subtle interaction cues (Fig 4c). To address this limitation, we introduce the denoising distillation that reverse-propagates pixel-level affordance priors from egocentric to exocentric views, thereby refining interaction knowledge by effectively suppressing background and human interference.

As shown in Fig 4a-b, our denoising distillation mechanism enhances the SCAM module by integrating parallel noise-absorbing heads, forming $\Theta_{\text{SCAM}}^{+\text{noise}}$. This extension aims to explicitly isolate non-affordance patterns (e.g., human limbs, cluttered backgrounds) from exocentric features. Given an exocentric feature map $\mathcal{F}^{\text{exo}}$, we simultaneously generate the primary affordance activation $\mathcal{G}^{\text{exo}}$ and M noise-specific activations $\mathcal{G}^{\text{noise}} \in \mathbb{R}^{H \times W \times M}$ through additional M convolution kernels:

$$\begin{aligned} \mathcal{G}^{\text{exo}}, \mathcal{G}^{\text{noise}} &= \Theta_{\text{SCAM}}^{+\text{noise}}(\mathcal{F}^{\text{exo}}; \theta^{+\text{noise}}) \\ f^{\text{exo}} &= \text{GAP}(\mathcal{R}(\mathcal{G}_{\hat{c}}^{\text{exo}}) \circ \mathcal{F}^{\text{exo}}) \\ f^{\text{pixel}} &= \text{GAP}(\mathcal{R}(\mathcal{P}_{\hat{c}}) \circ \mathcal{F}^{\text{ego}}), \\ \{f_m^{\text{noise}}\} &= \text{GAP}(\mathcal{R}(\{\mathcal{G}_m^{\text{noise}}\}) \circ \mathcal{F}^{\text{exo}}), \end{aligned} \quad (6)$$

where $\theta^{+\text{noise}}$ denotes the parameters with additional noise-absorbing heads, $\mathcal{P}_{\hat{c}}$ represents the egocentric localization result corresponding to the target class $\hat{c}$. The Hadamard product $\circ$, followed by global average pooling (GAP), is used to extract exocentric affordance activation features f^{exo} , noise features f_m^{noise} , and egocentric localization feature f^{pixel} .

To ensure f_m^{noise} captures noise effectively while directing affordance activation $\mathcal{G}_{\hat{c}}^{\text{exo}}$ toward affordance regions, we introduce a denoising distillation mechanism. The core idea is to align exocentric affordance-related features f^{exo}

with clean egocentric object features f^{pixel} , while enforcing noise features f^{noise} to diverge from affordance-related features f^{exo} . This naturally pushes noise activation toward object-irrelevant contexts and background regions. The denoising distillation is as follows:

$$\begin{aligned} s_m^{\text{noise}}, s^{\text{pixel}} &= \text{sim}(f_m^{\text{noise}}, f^{\text{exo}}), \text{sim}(f^{\text{pixel}}, f^{\text{exo}}) \\ \mathcal{L}_{\text{dill}} &= \log(1 + \sum_{m=1}^M \exp((s_m^{\text{noise}} - s^{\text{pixel}})/\tau)), \end{aligned} \quad (7)$$

where $\text{sim}(a, b)$ denotes the cosine similarity calculation, $s_m^{\text{noise}} \in \mathbb{R}^1$ represents the similarity between the noise features of the m -th noise-absorbing head and the exocentric activation features, and $s^{\text{pixel}} \in \mathbb{R}^1$ denotes the similarity between the egocentric localization features and the exocentric activation features. The loss function encourages the exocentric activation features to align with the precise localization features from the pixel decoder while penalizing the similarity between the affordance of exocentric images and irrelevant background features.

Finally, the overall loss for LoopTrans is defined as:

$$\mathcal{L} = \lambda_{\text{cls}} \mathcal{L}_{\text{cls}} + \lambda_{\text{dill}} \mathcal{L}_{\text{dill}} + \lambda_{\text{pixel}} \mathcal{L}_{\text{pixel}} + \lambda_{\text{corr}} \mathcal{L}_{\text{corr}}, \quad (8)$$

where $\lambda_{\text{cls}}, \lambda_{\text{dill}}, \lambda_{\text{pixel}}, \lambda_{\text{corr}}$ represents the weights assigned to the different loss components. $\mathcal{L}_{\text{corr}}$ is used to align the affordance correlations between exocentric and egocentric images following [29, 52]. The entire model is trained in an end-to-end manner.

5. Experiments

5.1. Datasets and Implementation Details

Image Benchmarks and Metrics. Following previous works [22, 27, 29, 52], we conduct experiments on AGD20K [29], which is a large-scale dataset comprising both exocentric and egocentric images and includes the splits ‘‘Seen’’ and ‘‘Unseen’’, as well as HICO-IFF derived from HICO-DET [5] and IIT-AFF [37]. For a fair comparison, we adopt the same metrics as in previous works for performance evaluation: Kullback-Leibler Divergence (KLD), Similarity (SIM), and Normalized Scanpath Saliency (NSS).

Video Benchmarks and Metrics. Following works [30, 36] in video affordance localization, we evaluate using the OPRA [13] and EPIC-Kitchens [11] datasets. Building on our image framework, we integrate an LSTM (after DINO) to represent the exo video using features from its last frame. Notably, videos of the OPRA dataset are collected from YouTube. However, since some resources are no longer available, results in Table 2 marked with the $\dagger$ represent experiments conducted with the full dataset, yielding outcomes that are currently unattainable with the accessible subset. For evaluation, we use the most common metrics in video affordance localization—KLD, SIM, and the Area

Method	Pub.	AGD20K-Seen			AGD20K-Unseen			HICO-IFF		
		KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
Weakly Supervised Object Localization										
SPA [40]	CVPR21	5.528	0.221	0.357	7.425	0.169	0.262	—	—	—
EIL [32]	CVPR20	1.931	0.285	0.522	2.167	0.277	0.330	—	—	—
TS-CAM [15]	ICCV21	1.842	0.260	0.336	2.104	0.201	0.151	—	—	—
Affordance Grounding										
Hotspots [36]	ICCV19	1.773	0.278	0.615	1.994	0.237	0.577	—	—	—
Cross-view-AG [29]	CVPR22	1.538	0.334	0.927	1.787	0.285	0.829	1.779	0.263	0.946
Cross-view-AG+ [31]	—	1.489	0.342	0.981	1.765	0.279	0.882	1.836	0.256	0.883
LOCATE [27]	CVPR23	1.226	0.401	1.177	1.405	0.372	1.157	1.593	0.327	0.966
WSMA [52]	AAAI24	<u>1.176</u>	<u>0.416</u>	<u>1.247</u>	<u>1.335</u>	<u>0.382</u>	<u>1.220</u>	<u>1.465</u>	<u>0.358</u>	<u>1.012</u>
INTRA [22]	ECCV24	1.199	0.407	1.239	1.365	0.375	1.209	—	—	—
Ours	—	1.088	0.445	1.322	1.247	0.403	1.315	1.399	0.379	1.226

Table 1. Comparison results on AGD20K-Seen, AGD20K-Unseen, and HICO-IFF benchmarks, where interaction knowledge is derived from exocentric images. The highest performance is **bolded**, and the second-highest is underlined.

	Method	EPIC			OPRA		
		KLD ↓	SIM ↑	AUC-J ↑	KLD ↓	SIM ↑	AUC-J ↑
Supervised	Img2heatmap [36]	1.400	0.359	0.794	1.473	0.355	0.821
	Demo2Vec [13]	—	—	—	1.197	0.482	0.847
	Afformer [7]	0.97	0.56	0.88	1.05	0.53	0.89
Weakly Supervised	Hotspot [†] [36]	—	—	—	1.427	0.362	0.806
	HAG-Net [†] [30]	—	—	—	1.409	0.365	0.812
	MLNET [9]	6.116	0.318	0.746	4.022	0.284	0.763
	EGOGAZE [21]	2.241	0.273	0.614	2.428	0.245	0.646
	SALGAN [39]	1.508	0.395	0.774	2.116	0.309	0.769
	DEEPGAZEII [25]	1.352	0.394	0.751	1.897	0.296	0.720
	Hotspot [36]	1.258	0.404	0.785	<u>1.537</u>	<u>0.342</u>	<u>0.754</u>
	HAG-Net [30]	1.209	0.414	0.801	—	—	—
	Ours	1.130	0.431	0.827	1.429	0.358	0.804
Image-to-Video generalization	LOCATE [27]	<u>1.382</u>	0.394	0.668	1.620	0.342	0.682
	WSMA [52]	1.425	0.371	0.720	<u>1.536</u>	0.344	0.748
	Ours	1.244	0.405	0.785	1.457	0.355	0.789

Table 2. Comparison results on EPIC and OPRA benchmarks, where interaction knowledge is derived from exocentric videos.

Under the Curve for the Jaccard index (AUC-J). For more details on the datasets, please refer to the Appendix.

Implementation Details. Following [27, 29, 31, 52], we set the input image resolution to 224×224 . For video inputs, we sample 8 frames at equal intervals and similarly resize them to 224×224 . The cluster number k is set to 4. All experiments are conducted on a single NVIDIA TITAN, using SGD as the optimizer with a learning rate of 1×10^{-3} .

5.2. Comparison with State-of-the-Art Methods

Comparison on Image Benchmarks. As shown in Table 1, we compare LoopTrans with state-of-the-art methods across three benchmarks (including both splits of AGD20K). Our approach consistently outperforms all existing methods across all evaluation metrics and settings. On AGD20K, we achieve average improvements of 6.7% over the the previous best-performing model WSMA [52] in KLD, SIM, and NSS, representing a relative increase of 236% compared to the gap between prior state-of-the-art methods. Unlike WSMA, which uses additional text domain prompts

as a medium for knowledge transfer, our LoopTrans fundamentally addresses localization challenges and inaccuracies in knowledge extraction caused by the domain gap through a unified shared CAM and a reverse egocentric-to-exocentric denoising distillation. Additionally, compared to LOCATE [27], which refines exocentric-to-egocentric transfer through feature selection, our shared CAM demonstrates more accurate knowledge transfer under occlusion and in complex scenes, yielding improvements of 11.3% across all three metrics. For HICO-IFF, our consistent improvement of 10.5% over WSMA further underscores the effectiveness and adaptability of LoopTrans.

Comparison on Video Benchmarks. As shown in Table 2, we comprehensively evaluate the proposed LoopTrans on video datasets from two perspectives: weakly supervised learning and image-to-video generalization. In both settings, LoopTrans consistently outperforms other methods, demonstrating substantial improvements in cross-domain knowledge transfer between exocentric and egocentric domains and significantly enhancing affordance localization

	Unified CAM	Pixel Alignment	Denoising Distillation	KLD ↓	SIM ↑	NSS ↑
Seen	✓			1.318	0.384	1.135
			✓	1.259	0.409	1.179
	✓	✓		1.251	0.392	1.196
	✓		✓	1.149	0.425	1.266
	✓	✓	✓	1.222	0.405	1.183
				1.088	0.443	1.322
Unseen	✓			1.635	0.332	0.853
			✓	1.508	0.341	1.122
	✓	✓		1.468	0.344	1.157
	✓		✓	1.335	0.394	1.258
	✓	✓	✓	1.431	0.368	1.189
				1.247	0.403	1.315

Table 3. Ablation study on AGD20K benchmark.

accuracy. (1) Weakly Supervised Setting: We compare LoopTrans with approaches that leverage temporal interaction knowledge through dedicated temporal modules. On the EPIC dataset, LoopTrans achieves an average improvement of 4.6% over HAG-Net [30]; on the OPRA dataset, it improves by 6.1% over Hotspot [36]. These gains highlight how our unified shared CAM not only bridges the domain gap between exocentric and egocentric perspectives but also supports robust knowledge transfer across video and image modalities, even with substantial modality gaps. (2) Image-to-Video Generalization: Trained on the AGD20K image dataset, LoopTrans is evaluated exclusively on video datasets. Against the current state-of-the-art open-source method, WSMA, LoopTrans achieves a further improvement of about 7.5%, underscoring its adaptability and robustness in bridging domain gaps. Notably, our framework performs consistently across exocentric benchmarks without requiring specialized temporal modules.

5.3. Ablation Study and Discussions

Beginning with our baseline, we develop six model variants across two subsets of AGD20K, conducting a total of 12 ablation experiments to rigorously evaluate the contributions of our modules as shown in Table 3. **(1) Baseline Model:** We use the one-way affordance grounding pipeline, LOCATE [27], as our baseline without applying feature selection, employing $\mathcal{L}_{\text{corr}}$ [29, 52] for knowledge transfer. **(2) Shared CAM:** Adding shared CAM yields consistent performance improvements (+4.5%) on seen split over KLD. This improvement demonstrates that our shared CAM effectively facilitates knowledge extraction and activation between exocentric and egocentric images, enabling each domain to leverage the other’s strengths. **(3) Denoising Distillation:** This mechanism enhances baseline performance by 5.1% by reinforcing egocentric information back to exocentric data, establishing a closed-loop knowledge cycle that enables CAM to filter out background noise and human-related artifacts while directing attention to interactive objects. **(4) Pixel Alignment:** Since pixel align-

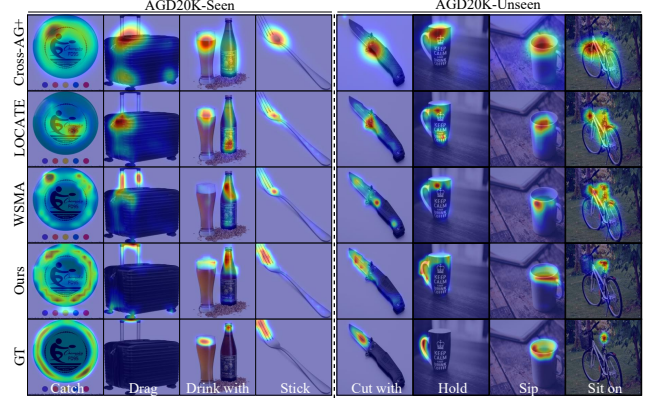


Figure 5. Visualization results compared with existing methods.

ment requires precise activation maps for egocentric images unavailable in the baseline’s one-way framework, we combine it with Shared CAM. Compared to using Shared CAM alone, Pixel Alignment further improves performance by 8.7%, demonstrating its ability to refine initial knowledge activation maps into more regionally complete object segments. **(5) Shared CAM with Denoising Distillation:** This combination achieves an average performance increase of 7.5% over using either component independently, suggesting that refined exocentric information better bridges the domain gap while shared knowledge extraction enhances final affordance localization. **(6) Complete Model:** Our full model achieves 1.088, 0.443, and 1.322 on KLD, SIM, and NSS metrics respectively on the seen split, demonstrating the synergistic effectiveness of all proposed components.

5.4. Visualization

As shown in Figure 5, we present additional qualitative affordance grounding results. Compared to Cross-View-AG+ [31], LOCATE [27], and WSMA [52], our approach significantly outperforms them in both localization accuracy and completeness. Notably, for occlusion-prone affordances like “sit on” and “catch”, our method achieves precise localization, while previous methods struggled due to occlusion. Moreover, our pixel-level decoding ensures that localized regions are precise, unlike the vague and broad results of prior methods. Additionally, our closed-loop knowledge transfer enables significantly better performance on challenging unseen splits compared to feature-based one-way approaches.

6. Conclusion

This paper proposes a closed-loop knowledge transfer framework for weak affordance localization. By incorporating shared knowledge extraction, pixel-level alignment, and positive knowledge feedback with denoising distillation, our model achieves consistent and significant improvements across all benchmarks.

Acknowledgment: This work is supported by the National Natural Science Foundation of China (No.62206174).

References

- [1] Jeremy N Bailenson, Kim Swinth, Crystal Hoyt, Susan Persky, Alex Dimov, and Jim Blascovich. The independent and interactive effects of embodied-agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments. *Presence*, 14(4): 379–393, 2005. 1
- [2] Christopher J Burke, Philippe N Tobler, Michelle Baddeley, and Wolfram Schultz. Neural mechanisms of observational learning. *Proceedings of the National Academy of Sciences*, 107(32):14431–14436, 2010. 1
- [3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1
- [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, 5
- [5] 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, 6, 12
- [6] Chaoqi Chen, Yushuang Wu, Qiyuan Dai, Hong-Yu Zhou, Mutian Xu, Sibe Yang, Xiaoguang Han, and Yizhou Yu. A survey on graph neural networks and graph transformers in computer vision: A task-oriented perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1
- [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. 7
- [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. 1, 3
- [9] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. A deep multi-level network for saliency prediction. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3488–3493. IEEE, 2016. 7
- [10] Qiyuan Dai and Sibe Yang. Curriculum point prompting for weakly-supervised referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13711–13722, 2024. 1
- [11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018. 6, 12
- [12] 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)*, pages 5882–5889. IEEE, 2018. 1, 3
- [13] 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. 1, 6, 7, 12
- [14] Stan Franklin. Autonomous agents as embodied ai. *Cybernetics & Systems*, 28(6):499–520, 1997. 1
- [15] Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, and Qixiang Ye. Ts-cam: Token semantic coupled attention map for weakly supervised object localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2886–2895, 2021. 3, 5, 7, 13
- [16] James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology press, 1979. 1
- [17] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8359–8367, 2018. 1
- [18] Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei. Embodied intelligence via learning and evolution. *Nature communications*, 12(1):5721, 2021. 1
- [19] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1
- [20] Hanzhuo Huang, Yufan Feng, Cheng Shi, Lan Xu, Jingyi Yu, and Sibe Yang. Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. *Advances in Neural Information Processing Systems*, 36:26135–26158, 2023. 1
- [21] Yifei Huang, Minjie Cai, Zhenqiang Li, and Yoichi Sato. Predicting gaze in egocentric video by learning task-dependent attention transition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 754–769, 2018. 7
- [22] Ji Ha Jang, Hoigi Seo, and Se Young Chun. Intra: Interaction relationship-aware weakly supervised affordance grounding. *arXiv preprint arXiv:2409.06210*, 2024. 2, 3, 4, 6, 7
- [23] Eunji Kim, Siwon Kim, Jungbeom Lee, Hyunwoo Kim, and Sungroh Yoon. Bridging the gap between classification and localization for weakly supervised object localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14258–14267, 2022. 3
- [24] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International journal of robotics research*, 32(8):951–970, 2013. 3
- [25] Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge. Deepgaze ii: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*, 2016. 7

- [26] Seungho Lee, Minhyun Lee, Jongwuk Lee, and Hyunjung Shim. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5495–5505, 2021. 5
- [27] 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, 6, 7, 8, 12, 13, 14
- [28] Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced nlp tasks. *arXiv preprint arXiv:1911.02855*, 2019. 6
- [29] 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, 3, 4, 5, 6, 7, 8, 12, 13, 14
- [30] 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*, 2023. 3, 6, 7, 8
- [31] Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, and Dacheng Tao. Grounded affordance from exocentric view. *International Journal of Computer Vision*, 132(6):1945–1969, 2024. 1, 3, 4, 7, 8, 12, 13, 14
- [32] Jinjie Mai, Meng Yang, and Wenfeng Luo. Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 7, 13
- [33] Jinjie Mai, Meng Yang, and Wenfeng Luo. Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8766–8775, 2020. 3
- [34] Luis Montesano, Manuel Lopes, Alexandre Bernardino, and José Santos-Victor. Learning object affordances: from sensory–motor coordination to imitation. *IEEE Transactions on Robotics*, 24(1):15–26, 2008. 1
- [35] 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. 1, 3
- [36] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8688–8697, 2019. 1, 3, 4, 6, 7, 8, 13
- [37] 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, 12
- [38] François Osiurak, Yves Rossetti, and Arnaud Badets. What is an affordance? 40 years later. *Neuroscience & Biobehavioral Reviews*, 77:403–417, 2017. 1
- [39] Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E O’Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017. 7
- [40] Xingjia Pan, Yingguo Gao, Zhiwen Lin, Fan Tang, Weiming Dong, Haolei Yuan, Feiyue Huang, and Changsheng Xu. Unveiling the potential of structure preserving for weakly supervised object localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11642–11651, 2021. 7, 13
- [41] Xingjia Pan, Yingguo Gao, Zhiwen Lin, Fan Tang, Weiming Dong, Haolei Yuan, Feiyue Huang, and Changsheng Xu. Unveiling the potential of structure preserving for weakly supervised object localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11642–11651, 2021. 3
- [42] 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*, pages 7587–7597, 2024. 3
- [43] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3
- [44] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9339–9347, 2019. 1
- [45] Johann Sawatzky and Jurgen Gall. Adaptive binarization for weakly supervised affordance segmentation. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 1383–1391, 2017. 3
- [46] Cheng Shi, Yulin Zhang, Bin Yang, Jiajin Tang, Yuexin Ma, and Sibe Yang. Part2object: Hierarchical unsupervised 3d instance segmentation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 1
- [47] Jiajin Tang, Ge Zheng, Cheng Shi, and Sibe Yang. Contrastive grouping with transformer for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23570–23580, 2023. 1
- [48] Jiajin Tang, Ge Zheng, and Sibe Yang. Temporal collection and distribution for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15466–15476, 2023. 1
- [49] Jiajin Tang, Ge Zheng, Jingyi Yu, and Sibe 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. 1

- [50] Fei Xia, Amir R Zamir, Zhiyang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: Real-world perception for embodied agents. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9068–9079, 2018. [1](#)
- [51] Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Rui-Wei Zhao, Tao Zhang, Xuequan Lu, and Shang Gao. Cream: Weakly supervised object localization via class re-activation mapping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9437–9446, 2022. [3](#)
- [52] 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. [2](#), [3](#), [4](#), [6](#), [7](#), [8](#), [12](#), [14](#)
- [53] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1325–1334, 2018. [3](#), [5](#)
- [54] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibe Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191, 2023. [1](#)
- [55] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. [2](#), [3](#), [5](#)
- [56] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part II 13*, pages 408–424. Springer, 2014. [1](#)
- [57] Yuchen Zhu, Cheng Shi, Dingyou Wang, Jiajin Tang, Zhengxuan Wei, Yu Wu, Guanbin Li, and Sibe Yang. Re-thinking query-based transformer for continual image segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4595–4606, 2025. [1](#)

7. Challenges of the Current One-Way Transfer Framework

As discussed in the main text, two essential yet challenging aspects of weakly supervised affordance grounding are: (1) extracting affordance knowledge from exocentric interactions and (2) transferring it to egocentric localization. **Challenge of Vague and Broad Activation.** For knowledge extraction, previous one-way pipelines [27, 29, 31, 52] often struggle to generate accurate exocentric activation maps due to interference from complex and diverse interaction scenarios, such as background clutter and the presence of human bodies. This makes it difficult to capture the correct affordance regions. As shown in Figure 6, the previous method [27] produces coarse or even erroneous knowledge activation maps. In contrast, our approach leverages a shared CAM mechanism that integrates object-centric and interaction-centric knowledge learning, enabling precise activation focused on affordance-relevant object regions.

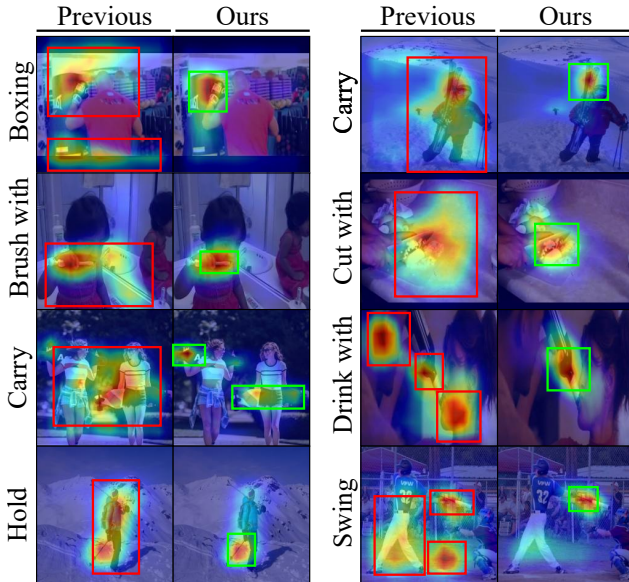


Figure 6. Comparison of knowledge activation on exocentric images between our LoopTrans and the previous one-way transfer framework.

Invalid Transfer Caused by Occlusion Challenge. For the exocentric-to-egocentric transfer, one-way pipelines aim to achieve feature-based transfer by leveraging interaction knowledge in the exocentric view and aligning features between the exocentric and egocentric perspectives. However, in most scenarios involving actions like riding or holding, the interaction areas on the objects are often occluded by the human body due to the nature of these actions. As illustrated in Figure 7, even if the vague activation regions indicated by previous methods are roughly correct, occlusion by the human body prevents the accurate extraction of

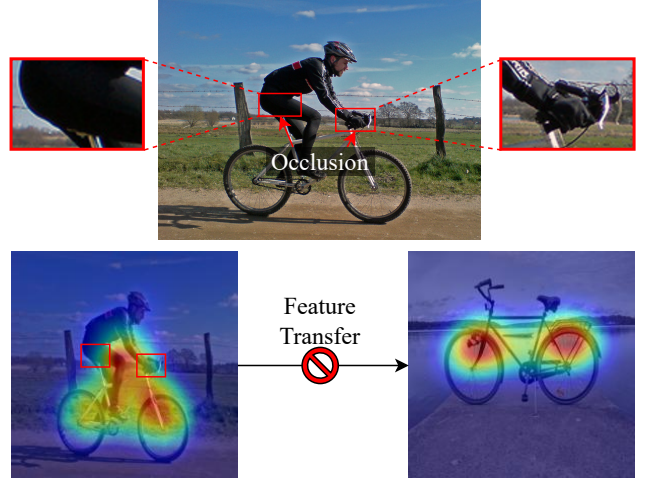


Figure 7. Ineffective knowledge feature extraction and transfer caused by occlusion in one-way transfer frameworks.

affordance-related object features. This renders the knowledge (features) transferred by one-way pipelines ineffective. In contrast, our approach jointly leverages shared exocentric interaction knowledge and egocentric object knowledge, rather than relying on explicit feature transfer. This allows interaction knowledge to be directly activated on egocentric images, fundamentally addressing the issue of ineffective feature-based transfer caused by occlusion.

8. Details of Datasets and Metrics

Image Datasets. AGD20K [29] is a large-scale dataset specifically designed for affordance grounding, comprising 20,061 exocentric images and 3,755 egocentric images annotated with 36 distinct affordance categories. This dataset facilitates evaluations in both Seen and Unseen settings, allowing us to assess the model’s ability to generalize across different object categories. HICO-IIF [52], a composite dataset derived from HICO-DET [5] and IIT-AFF [37], includes exocentric images from HICO-DET and egocentric images from IIT-AFF, covering ten affordance classes and seven object categories. This combination enables performance evaluation even with a relatively limited dataset size, consisting of 4,383 training images and 1,498 test images.

Video Datasets. OPRA [13] comprises approximately 16,000 product review videos sourced from YouTube, showcasing interactions with household appliances. Each video is paired with a static product image, an action label, and an affordance heatmap that highlights relevant interaction regions. EPIC-Kitchens [11] features unscripted egocentric videos depicting various kitchen activities, annotated with action and object labels. Target images are selected from specific frames, and crowd-sourced annotations provide ground-truth heatmaps for relevant affordance regions. Together, these datasets enhance our ability to

	Method	Big			Middle			Small		
		KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
Seen	EIL [32]	1.047	0.461	0.389	1.794	0.284	0.710	3.057	0.123	0.231
	SPA [40]	5.745	0.317	0.222	4.990	0.228	0.440	6.076	0.118	0.297
	TS-CAM [15]	1.039	0.424	0.166	1.814	0.248	0.401	2.652	0.132	0.352
	Hotspots [36]	0.986	0.448	0.408	1.738	0.265	0.672	2.587	0.149	0.683
	Cross-View-AG [29]	0.766	0.533	0.652	1.485	0.322	1.040	2.373	0.175	0.927
	Cross-View-AG+ [31]	0.787	0.521	0.660	1.481	0.314	1.089	2.381	0.167	0.959
	LOCATE [27]	0.676	0.580	0.706	1.178	0.390	1.316	2.029	0.216	1.349
	Ours	0.661	0.607	0.737	1.069	0.435	1.411	1.657	0.276	1.774
Unseen	EIL [32]	1.199	0.393	0.271	1.906	0.246	0.482	3.082	0.113	0.116
	SPA [40]	8.299	0.259	0.254	6.938	0.186	0.333	7.784	0.095	0.144
	TS-CAM [15]	1.238	0.351	0.072	1.970	0.208	0.236	2.766	0.113	0.124
	Hotspots [36]	1.015	0.425	0.548	1.872	0.242	0.605	2.693	0.134	0.544
	Cross-View-AG [29]	0.884	0.500	0.728	1.595	0.303	0.945	2.558	0.147	0.692
	Cross-View-AG+ [31]	0.867	0.485	0.776	1.658	0.279	0.988	2.630	0.133	0.754
	LOCATE [31]	0.571	0.629	0.956	1.302	0.373	1.257	2.223	0.189	1.071
	Ours	0.568	0.619	1.021	1.140	0.417	1.422	1.965	0.223	1.355

Table 4. Comparison results on AGD20K with different affordance region scales.

learn affordance grounding from both images and videos of human-object interactions.

Details of Metrics

- **Kullback-Leibler Divergence (KLD):** The Kullback-Leibler Divergence quantifies the divergence between two probability distributions. The formula for KLD is:

$$\text{KLD}(P, Q) = \sum_i Q_i \log \left(\frac{Q_i}{P_i} \right). \quad (9)$$

Here, $P \in \mathbb{R}^{HW}$ is the predicted heatmap, and $Q \in \mathbb{R}^{HW}$ is the true distribution. H and W represent the height and width of the distributions, respectively.

- **Similarity (SIM):** The Similarity metric assesses the degree of overlap between the predicted and true distributions. The formula for SIM is:

$$\text{SIM}(P, Q) = \sum_i \min(P_i, Q_i). \quad (10)$$

Here, $P \in \mathbb{R}^{HW}$ is the predicted distribution, and $Q \in \mathbb{R}^{HW}$ is the true distribution.

- **Normalized Scanpath Saliency (NSS):** Normalized Scanpath Saliency evaluates how well the predicted heatmap aligns with the ground truth binary map. Specifically, we first normalize the predicted distribution and then calculate the NSS:

$$\bar{P} = \frac{P - \mu(P)}{\sigma(P)}, \quad (11)$$

$$\text{NSS}(\bar{P}, M) = \frac{1}{HW} \sum_i \bar{P}_i \times M_i.$$

Here, $P \in \mathbb{R}^{HW}$ is the predicted heatmap, $M \in \{0, 1\}^{HW}$ is the ground truth binary map, and $\mu(P)$ and $\sigma(P)$ are the mean and standard deviation of P , respectively.

Details of Image-to-Video Generalization Setting. Since the affordance categories in OPRA and EPIC differ from those in AGD20K, models trained on AGD20K cannot be directly tested on these video datasets. However, we observe that due to the richness of affordance categories in AGD20K, most affordances in other datasets have similar counterparts in AGD20K. Thus, we align the affordance categories by mapping each category in the video dataset to its most relevant counterpart in the image dataset for testing.

9. More Experimental Results

Comparison on Different Scales. Following [27, 29], we divide the test set into three subsets—large, medium, and small—based on the size of the affordance object regions. The subsets are defined as follows: affordance regions occupying more than 10% of the image area are categorized as large, those between 3% and 10% as medium, and those smaller than 3% as small. As shown in Table 4, we conduct experiments on AGD20K [29] dataset to assess the robustness of our approach across different affordance region scales. The results demonstrate that our method consistently outperforms previous approaches [27, 29, 31] across all size categories.

Ablation on Image Radio. The ablation study evaluates the effect of varying exocentric-to-egocentric sample ratios (1:1 to 5:1) on AGD20K-Seen and AGD20K-Unseen

Exo:Ego	AGD20K-Seen			AGD20K-Unseen		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
1:1	1.077	0.449	1.335	1.250	0.411	1.312
2:1	1.097	0.445	1.321	1.265	0.405	1.306
3:1	1.088	0.445	1.322	1.247	0.403	1.315
4:1	1.068	0.452	1.341	1.285	0.396	1.301
5:1	1.103	0.443	1.309	1.265	0.402	1.307

Table 5. Ablation study on the ratio of exocentric images to ego images

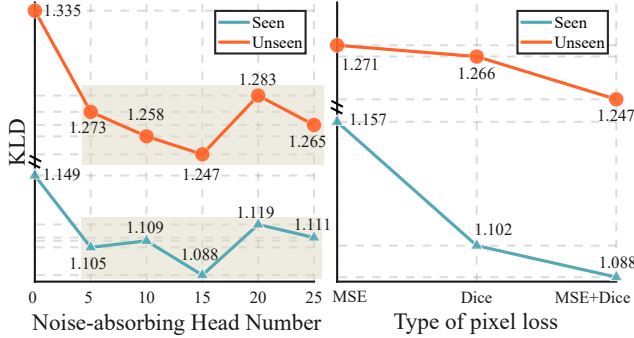


Figure 8. Ablation study on the number of noise-absorbing heads (left) and type of pixel decoder loss (right).

datasets. On AGD20K-Seen, the 4:1 ratio achieves the best performance across all metrics (KLD: 1.068, SIM: 0.452, NSS: 1.341), indicating that emphasizing exocentric data enhances affordance feature extraction while maintaining egocentric localization. However, performance declines at 5:1, suggesting diminishing returns with excessive exocentric emphasis. For AGD20K-Unseen, the 3:1 ratio performs best (NSS: 1.315, KLD: 1.247, SIM: 0.403), demonstrating robust generalization to unseen affordance objects. Further increases in the ratio reduce SIM and NSS, underscoring the need for sufficient egocentric samples to support generalization. For a fair comparison with prior methods [27, 29], we adopt the 3:1 ratio as the final setting.

Ablation on the Number of Noise-absorbing Heads. As shown in Figure 8 (left), we conduct experiments to assess the impact of the number of noise-absorbing heads. Setting the number of noise-absorbing heads to zero—indicating a purely distilled, non-denoising approach—leads to significantly reduced performance, as noisy features in exocentric images disrupt knowledge extraction and hinder exocentric-to-egocentric transfer. Increasing the number of heads improves performance, particularly on the seen split, with optimal results at 15.

Ablation on the Type of Pixel Decoder Loss. As shown in Figure 8 (right), we conduct experiments to assess the impact of pixel decoder loss type. The results shows that combining MSE and Dice losses markedly outperforms either loss alone, especially on the unseen split, where the pixel-level Dice loss enhances cross-category localization

alignment.

10. More Visualizations

As illustrated in Figure 9 and Figure 10, we present visualization results across multiple affordance categories. Our method consistently surpasses all existing approaches [27, 29, 31, 52] across all categories, demonstrating notable advantages in scenarios where affordance regions are occluded by human interactions. For example, in the *hold* and *ride* categories, our approach effectively localizes occluded regions such as handles and saddle areas in egocentric images—achievements that remain unattainable by prior methods.

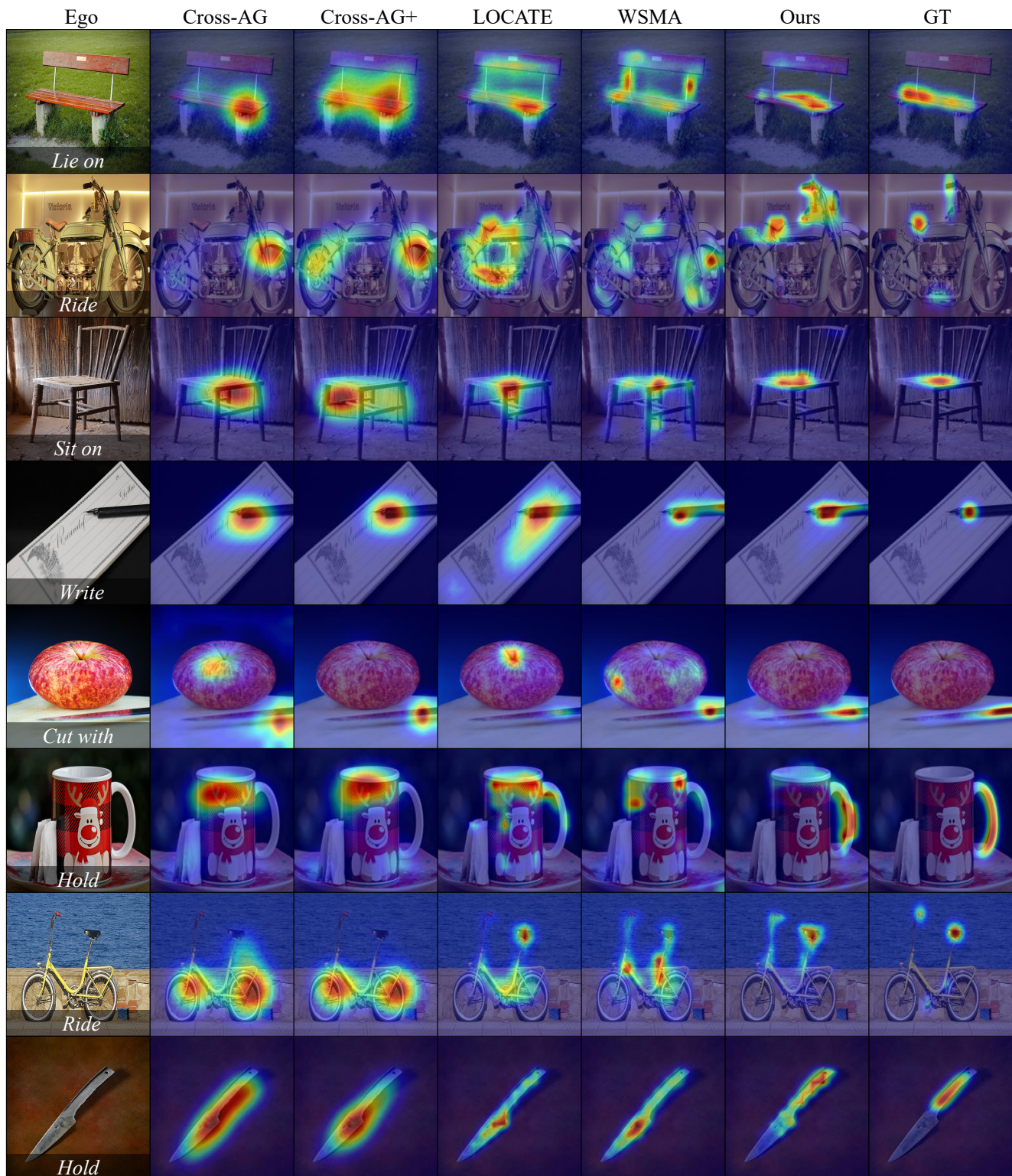


Figure 9. More visualization for affordance grounding results on egocentric images.

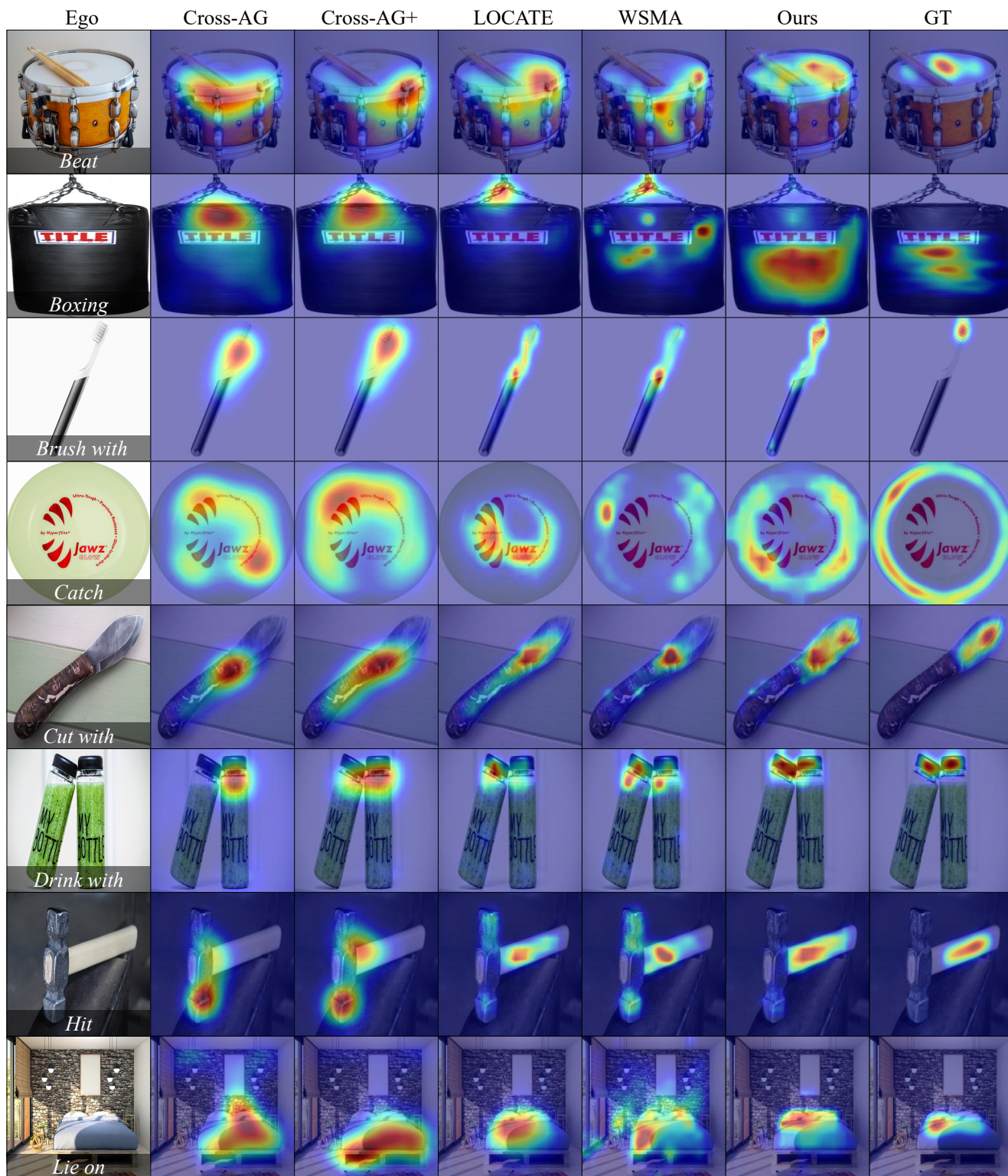


Figure 10. More visualization for affordance grounding results on egocentric images.