ACTION-AWARE SPATIAL SUPERVISION FOR WEAKLY SUPERVISED AFFORDANCE GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Weakly supervised affordance grounding (WSAG) aims to localize action-relevant object regions using only image-level supervision. Existing methods exploit exocentric human–object interaction images to support affordance localization in egocentric images, with recent approaches further leveraging vision–language foundation models to generate dense pseudo labels for the egocentric view. However, exocentric cues are still aggregated over object regions without explicit guidance from the queried action, while egocentric pseudo labels and predictions may remain spatially imprecise. We therefore propose a unified framework that improves spatial reliability across exocentric transfer, pseudo-label learning, and prediction optimization. First, we introduce an Action-Aware Exocentric Pooling (AAEP) module that combines queried-action-guided feature aggregation with fine-grained exocentric interaction pseudo-mask supervision, emphasizing action-relevant interaction regions. Second, we develop a Reliability-Calibrated Pseudo-Label Refinement strategy that performs action-aware refinement of egocentric pseudo labels and adaptively fuses the refined labels with the originals. Finally, we introduce a Peak-Anchored Support Expansion (PASE) Loss that encourages compact, appearance-consistent support around reliable peaks, reducing overly concentrated responses. Experiments on AGD20K demonstrate that our model achieves state-of-the-art performance, with a 4.2% reduction in KLD and gains of 1.6% in SIM and 4.1% in NSS on the Unseen split.

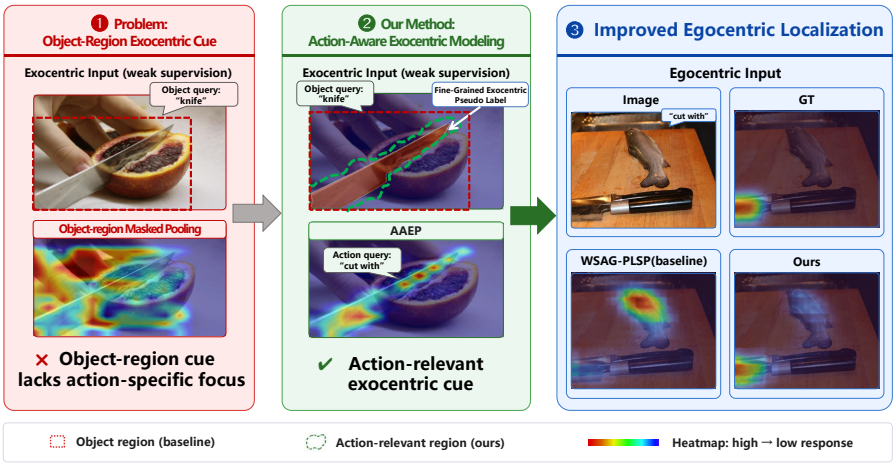

Figure 1: Action-aware exocentric modeling. Object-region pooling may mix action-relevant and irrelevant patches. Under the same object constraint, AAEP uses the action query and fine-grained interaction pseudo-mask supervision to emphasize action-relevant regions.

1 INTRODUCTION

Affordance grounding localizes object regions that support a queried action. Densely supervised methods require costly pixel-level annotations (Myers et al., 2015; Do et al., 2018), motivating

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.

sparse supervision (Sawatzky et al., 2017) and image-level weakly supervised affordance grounding (WSAG). Cross-view methods transfer interaction cues from exocentric images to egocentric objects through interaction evidence, part localization, and multimodal reasoning (Luo et al., 2022; Li et al., 2023; Xu et al., 2024), but localization quality depends on the spatial precision of transferred cues.

Vision–language foundation models further enable dense pseudo supervision from object and part semantics (Rai et al., 2024). WSAG-PLSP generates egocentric pseudo labels using semantic part queries and off-the-shelf segmentation models (Xu & Mu, 2025). However, these labels may overcover objects or include action-irrelevant regions.

Action relevance is central to affordance grounding, since the functional role of an object region varies with the queried action. For example, a knife is held by its handle but cuts with its blade. Recent methods have improved egocentric supervision by generating finer-grained pseudo labels. However, the exocentric cues used for cross-view transfer remain comparatively coarse: interaction features are pooled over the detected object region without explicit conditioning on the queried action. As illustrated in Fig. 1, such pooling may mix patches with different relevance to the queried action, yielding coarse cross-view cues for localizing the corresponding functional region.

To address these limitations, we introduce action-aware spatial supervision across exocentric transfer, egocentric pseudo-label refinement, and prediction optimization. The framework combines action-conditioned interaction evidence with reliability-aware supervision and peak-anchored support expansion.

Our contributions are summarized as follows:

- We introduce Action-Aware Exocentric Pooling (AAEP), which leverages the queried action to guide exocentric feature aggregation and aligns the resulting spatial attention with fine-grained exocentric interaction pseudo masks, thereby emphasizing action-relevant interaction regions.
- We develop a Reliability-Calibrated Pseudo-Label Refinement strategy for egocentric supervision. It constructs action-aware refinement candidates according to the quality of the original pseudo labels and adaptively balances the original and refined supervision based on candidate reliability, limiting unreliable corrections while preserving useful original responses.
- We propose a Peak-Anchored Support Expansion (PASE) Loss to alleviate overly concentrated predictions by conservatively extending responses from reliable pseudo-label peaks toward compact, appearance-consistent support regions. Experiments on AGD20K demonstrate consistent improvements, particularly on unseen compositions.

2 Related Work

2.1 Visual Affordance Learning

Affordance describes the actions enabled by an object (Gibson, 1979; Şahin et al., 2007). Early visual methods localized functional regions using RGB-D geometry (Myers et al., 2015), pixel-wise RGB prediction (Roy & Todorovic, 2016), or joint object detection and affordance segmentation (Do et al., 2018). These approaches require costly dense annotations. Sparse supervision reduces this burden but still requires spatial labels (Sawatzky et al., 2017), whereas image-level supervision uses only object and action labels. Although affordance grounding also extends to 3D point clouds (Deng et al., 2021; Yang et al., 2023), we focus on 2D images.

2.2 Weakly Supervised Affordance Grounding

Weakly supervised affordance grounding (WSAG) localizes action-relevant object regions from image-level labels without manually annotated affordance masks. Existing methods mainly follow two directions: CAM-based cross-view learning and foundation-model-assisted pseudo supervision. Cross-View-AG established the cross-view setting by transferring interaction cues from exocentric images and producing localization maps with CAM (Luo et al., 2022; Zhou et al., 2016), with subsequent methods improving cross-view transfer through part prototypes (Li et al., 2023), multimodal or relational modeling (Xu et al., 2024; Jang et al., 2024; Wang et al., 2025), and selective

contrastive or closed-loop learning (Moon et al., 2025; Tang et al., 2025). More recently, vision–language foundation models have enabled dense pseudo supervision from semantic object and part cues (Rai et al., 2024). WSAG-PLSP follows this direction by translating object–action pairs into semantic part queries to generate egocentric pseudo masks, while using exocentric interaction images for cross-view regularization (Xu & Mu, 2025). However, its exocentric evidence remains comparatively coarse, as interaction features are aggregated over detected object regions without explicit action-conditioned spatial selection. Such aggregation may mix patches with different relevance to the queried action. We therefore exploit exocentric interactions to provide finer-grained, action-aware spatial guidance for egocentric affordance localization.

3 METHOD

3.1 PROBLEM SETTING

We study weakly supervised cross-view affordance grounding under the Cross-View-AG protocol (Luo et al., 2022). During training, for a given affordance query, the model jointly exploits egocentric object images and exocentric interaction images, with the latter providing complementary cross-view evidence for egocentric affordance localization. At inference, given an egocentric image I^{ego} and query a, it predicts a dense affordance heatmap H for the queried action.

3.2 METHOD OVERVIEW

As illustrated in Fig. 2, we propose an action-aware spatial supervision framework for weakly supervised affordance grounding. Our key observation is that different actions may involve distinct functional regions of the same object, motivating more explicit action guidance during learning. Accordingly, Action-Aware Exocentric Pooling (AAEP) uses the queried action to guide exocentric feature aggregation, reliability-calibrated refinement improves egocentric pseudo labels through adaptive fusion, and Peak-Anchored Support Expansion (PASE) encourages predictions to cover coherent affordance regions around reliable peaks. The original affordance reasoning and fusion module of WSAG-PLSP (Xu & Mu, 2025) is retained, while the proposed components enhance spatial supervision at the exocentric, egocentric, and prediction stages.

3.3 ACTION-AWARE EXOCENTRIC POOLING

Exocentric images provide rich human–object interaction cues, but not all visual regions are equally relevant to the queried affordance, as illustrated in Fig. 1. In particular, only a subset of the object region may correspond to the functional part involved in the queried action. We therefore introduce AAEP, which leverages the action query as a semantic selector over exocentric object patches and aggregates action-relevant interaction evidence before cross-view transfer. Rather than directly predicting the exact functional region, AAEP provides an action-conditioned preference for selecting more relevant exocentric cues.

Action-aware exocentric pooling. As illustrated in Fig. 2, AAEP first projects exocentric tokens into keys and values, uses the action query to obtain object-masked attention weights, and then pools the values into an action-aware exocentric feature. Given exocentric patch features $\mathbf{F}_i^{\mathrm{exo}} = \{\mathbf{f}_{ij}^{\mathrm{exo}}\}_{j=1}^{N}$ and action embedding $\mathbf{t}_i^a$, AAEP projects them into a shared matching space:

$$\mathbf{q}_i = W_q \mathbf{t}_i^a, \qquad \mathbf{k}_{ij} = W_k \mathbf{f}_{ij}^{\mathrm{exo}}. \tag{1}$$

Here W_q and W_k are learnable projections, $\mathbf{q}_i$ is the query representing the action, and $\mathbf{k}_{ij}$ is the key representing exocentric patch j. Because exocentric patches also contain the actor and background, we use the detected object box as a patch-level constraint to restrict action matching to the manipulated object. Let $\mathbf{B}_i^{\mathrm{exo}}$ be the detected object box and $\mathbf{R}_{ij}$ the image region covered by patch j. We define the object-patch set as

$$\Omega_i^{\mathrm{obj}} = \{j \mid \mathbf{R}_{ij} \cap \mathbf{B}_i^{\mathrm{exo}} \neq \varnothing\}. \tag{2}$$

The object constraint removes most actor and background responses and defines where action matching is performed. Within this region, the action query determines the relative importance of different

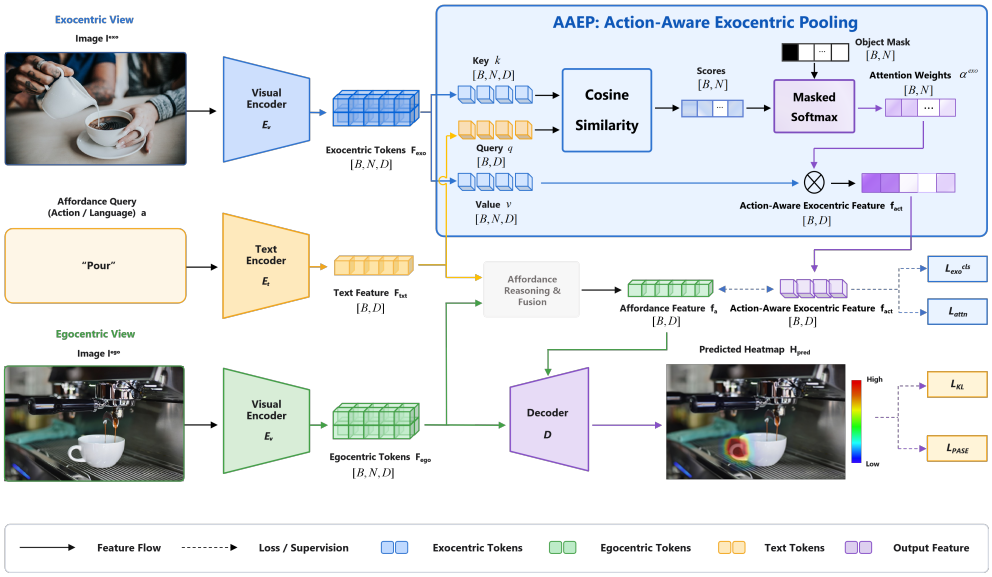

Figure 2: Overall pipeline of the proposed framework. Given exocentric and egocentric images together with an affordance query, the encoders extract view-specific visual tokens and the action representation. AAEP leverages the queried action to guide attention over exocentric tokens and aggregates them into an action-aware exocentric representation for cross-view supervision. The egocentric representation is fused with the affordance query and decoded into a dense affordance heatmap, which is optimized with the overall training objective augmented by the proposed PASE Loss. Solid and dashed arrows denote feature flow and training supervision, respectively.

object patches. For each candidate patch $j \in \Omega_i^{\text{obj}}$, we compute an action–patch compatibility score:

$$s_{ij} = \frac{\text{sim}(\mathbf{q}_i, \mathbf{k}_{ij})}{\tau}, \qquad j \in \Omega_i^{\text{obj}}, \tag{3}$$

where $\text{sim}(\mathbf{q}, \mathbf{k}) = \mathbf{q}^\top \mathbf{k} / (\|\mathbf{q}\|_2 \|\mathbf{k}\|_2)$ and τ controls the sharpness of the matching scores. A larger s_{ij} indicates that patch j is more compatible with the queried action. We apply an object-masked softmax:

$$\alpha_{ij}^{\text{exo}} = \begin{cases} \dfrac{\exp(s_{ij})}{\sum_{k \in \Omega_i^{\text{obj}}} \exp(s_{ik}) + \epsilon}, & j \in \Omega_i^{\text{obj}}, \\ 0, & j \notin \Omega_i^{\text{obj}}. \end{cases} \tag{4}$$

This is equivalent to assigning $-\infty$ logits to patches outside the detected object before softmax, ensuring that the attention distribution is normalized only over object patches.

The attention weights are then used to pool value-projected exocentric patches:

$$\mathbf{v}_{ij} = W_v \mathbf{f}_{ij}^{\text{exo}}, \qquad \mathbf{f}_i^{\text{act}} = \sum_j \alpha_{ij}^{\text{exo}} \mathbf{v}_{ij}. \tag{5}$$

The resulting feature $\mathbf{f}_i^{act}$ summarizes the exocentric object under the queried action, while $\boldsymbol{\alpha}_i^{exo}$ provides its corresponding spatial weighting. Since action-conditioned similarity alone does not guarantee accurate functional localization, we further supervise this attention using fine-grained interaction pseudo labels.

Action-conditioned exocentric spatial supervision. Following prior action–object-to-part semantics (Xu & Mu, 2025), we construct exocentric spatial supervision by adapting the existing prompt mapping to interaction images. For the i-th training sample, let v_i denote the queried action (verb) and o_i denote the manipulated object (noun). Formally, we map the action–object pair to a set of candidate part queries:

$$\mathcal{Q}_i = \mathcal{P}(v_i, o_i). \tag{6}$$

For example, the underspecified pair (*drink with*, *cup*) is mapped to {*cup rim*, *cup opening*}, converting a generic object query into candidate interaction parts. To ground each part query in the exocentric image, VLPart (Sun et al., 2023) first localizes $p \in \mathcal{Q}_i$ with a part box $\mathbf{B}_{i,p}^{\text{part}}$, which SAM (Kirillov et al., 2023) subsequently converts into a candidate mask constrained to the detected exocentric object region. After validity filtering, the highest-ranked candidate is retained and denoted as $\mathbf{M}_i^{\text{int}}$. If no valid candidate is available, the corresponding spatial supervision is skipped. These masks serve as soft spatial supervision rather than hard part annotations. Further details of the prompt policy and mask-selection procedure are provided in Supplementary Sec. B.

For each valid interaction mask $\mathbf{M}_i^{\text{int}}$, we resize it to match the exocentric patch grid, normalize the resized mask into a spatial distribution $\overline{\mathbf{M}}_i^{\text{int}}$, and use this distribution to supervise the action-conditioned attention produced by AAEP:

$$\mathcal{L}_{\text{attn}} = \frac{1}{\sum_i g_i + \epsilon} \sum_i g_i \ \text{JS}\Big(\overline{\mathbf{M}}_i^{\text{int}} \ \Big\| \ \boldsymbol{\alpha}_i^{\text{exo}}\Big), \tag{7}$$

where $\boldsymbol{\alpha}_i^{\text{exo}}$ is the object-masked action-conditioned attention distribution defined in Eq. (4), and g_i indicates whether valid spatial supervision is available. The JS divergence encourages consistency between the two soft spatial distributions without treating the pseudo mask as an exact target.

3.4 Reliability-Calibrated Pseudo-Label Refinement

Part-level semantic prompting provides finer egocentric pseudo labels, but these masks are not uniformly reliable. Some masks remain close to the whole-object extent, while others already provide localized functional evidence. Applying the same correction to all samples can therefore damage otherwise reliable supervision. We address this by separating refinement into two stages: the quality of the original pseudo label determines how strongly a candidate should be corrected, while the reliability of the refined candidate determines how much it should contribute to the final supervision.

Original pseudo-label quality. Unlike object segmentation, affordance grounding does not aim to recover the complete object extent. A pseudo label covering most of the object can therefore contain substantial regions that are weakly related to the queried action. We use spatial coverage as a simple indicator of such whole-object-like supervision.

Given an egocentric image I_i, an action-object pair (v_i, o_i), an initial pseudo label P_i produced by the underlying part-level pseudo-labeling method (Xu & Mu, 2025), and the detected object region O_i, let $B(P_i)$ denote the binarized pseudo label. Specifically, we define two spatial coverage ratios:

$$a_i^{img} = \frac{|B(P_i)|}{|I_i|}, \qquad a_i^{obj} = \frac{|B(P_i) \odot O_i|}{|O_i|}.$$

The image-level coverage a_i^{img} measures whether the pseudo label occupies a large image region, while the object-level coverage a_i^{obj} measures whether it covers most of the target object. A large a_i^{obj} indicates that the pseudo label may behave as whole-object supervision.

Reliability-calibrated candidate refinement. Motivated by the fact that different actions may involve different functional regions of the same object, we construct an action-aware spatial prior $A_i = A(v_i, O_i)$ by applying a verb-specific functional-region template within the detected object region. The template assigns higher weights to regions plausibly supporting the queried action, such as support surfaces for *sit on*, graspable boundaries for *hold*, and functional ends for tool-use actions. Thus, A_i provides a coarse action-dependent spatial preference rather than a semantic part mask. It is used only as a correction cue, whose contribution is controlled by reliability-calibrated fusion with the original pseudo label. The template definitions are provided in Supplementary Sec. B.

We denote labels with $a_i^{obj} > \tau_{obj}$ or $a_i^{img} > \tau_{img}$ as the whole-object-like subset $\mathcal{H}$. Among the remaining labels, those satisfying $a_i^{obj} > \tau_{soft}$ form the moderately broad subset $\mathcal{M}$. Since uniformly applying the prior may distort already localized labels, the correction is selected by the coverage group:

$$\hat{P}_i = \begin{cases} B(P_i) \odot A_i, & i \in \mathcal{H}, \\ P_i \odot \big(\lambda + (1-\lambda) A_i\big), & i \in \mathcal{M}, \\ P_i \odot C_i, & \text{otherwise.} \end{cases}$$

Labels in $\mathcal{H}$ receive strong action-guided correction, while labels in $\mathcal{M}$ are softly reweighted. For already localized labels, we only suppress responses outside a loose object region using

$$C_i = \mathrm{Norm}(G_{\sigma_c} * \mathrm{Dilate}(O_i; k_c)) .$$

We set $\tau_{obj} = 0.62$, $\tau_{img} = 0.35$, $\tau_{soft} = 0.45$, and $\lambda = 0.35$. If the refined candidate becomes excessively small, we revert to the original pseudo label to avoid over-correction.

Reliability-guided fusion. The refined pseudo label $\hat{P}_i$ is not always more reliable than the original pseudo label P_i. We therefore determine its contribution to the final supervision using a reliability score $r_i \in [0, 1]$. The score evaluates the spatial extent of the refined pseudo label: overly broad refinements are down-weighted, while extremely broad predictions are rejected. The detailed calibration is provided in Supplementary Sec. B.

The final fusion is defined as

$$c_i = \mathrm{clip}(\gamma r_i, 0, 1) , \qquad Y_i = c_i \hat{P}_i + (1 - c_i) P_i, \tag{8}$$

where γ controls the maximum contribution of the refined pseudo label. A low c_i retains more of the original pseudo label, whereas a high c_i allows the refined label to contribute more strongly. For clarity, we additionally provide pseudocode for the complete pseudo-label refinement and fusion procedure in Supplementary Algorithm S1.

3.5 Peak-Anchored Support Expansion Loss

Refining pseudo labels improves the supervision target but does not guarantee that the learned prediction covers the corresponding functional region. During optimization, the model can still concentrate probability on a few highly discriminative locations. We therefore introduce PASE loss, which regularizes prediction support around reliable pseudo-label peaks.

Importantly, PASE does not enlarge or modify the pseudo label. Instead, it identifies a small set of candidate pixels that are supported by the pseudo label and locally consistent with its high-confidence peak, and requires only a bounded amount of prediction mass in this region. This avoids treating uncertain neighboring pixels as hard supervision while discouraging the prediction from collapsing onto isolated peaks.

Let Y denote the fused pseudo-supervision target defined in Eq. (8), and let $\bar{Y}$ be its spatially normalized distribution. Given the prediction logits Z, we obtain the normalized prediction $P = \mathrm{softmax}_{\mathrm{spatial}}(Z)$. PASE first identifies a high-response peak in Y and then selects a compact non-peak support region Ω_{sel}, consisting of pixels that are both supported by the pseudo-supervision target and appearance-consistent with the peak. The selected region only defines where the probability-mass constraint is evaluated; it neither modifies Y nor treats the neighboring pixels as additional hard supervision. Further details are provided in Supplementary Sec. C, and the complete support-selection procedure is summarized in Supplementary Algorithm S3.

PASE then compares the pseudo-label and prediction probability mass within Ω_{sel}. The pseudo-label mass is computed from the spatially normalized label, $m_{\mathrm{tar}} = \sum_{i \in \Omega_{\mathrm{sel}}} \bar{\mathbf{Y}}_i$, while $m_{\mathrm{pred}} = \sum_{i \in \Omega_{\mathrm{sel}}} \mathbf{P}_i$. We convert m_{tar} into a bounded requirement m_{req} and penalize the prediction only when the assigned mass is insufficient:

$$\mathcal{L}_{\mathrm{PASE}} = \frac{1}{|\mathcal{B}_v|} \sum_{n \in \mathcal{B}_v} \left[m_{\mathrm{req}}^{(n)} - m_{\mathrm{pred}}^{(n)} \right]_+ . \tag{9}$$

where $\mathcal{B}_v$ contains samples with valid peaks and non-empty selected candidate regions. Thus, PASE is inactive when the selected region already receives sufficient prediction mass, keeping the expansion weak. Implementation settings are provided in the supplementary material (Supplementary Sec. C).

Our final training objective combines the base affordance grounding loss, the exocentric attention alignment loss, and the proposed PASE regularizer:

$$\mathcal{L} = \mathcal{L}_{\mathrm{base}} + \lambda_{\mathrm{attn}} r_{\mathrm{attn}}(e) \mathcal{L}_{\mathrm{attn}} + \lambda_{\mathrm{PASE}} r_{\mathrm{PASE}}(e) \mathcal{L}_{\mathrm{PASE}}. \tag{10}$$

Table 1: Main comparison on AGD20K under the standard Seen/Unseen protocol. Baseline results are reported from the corresponding original papers. The best result for each metric is marked in bold.

Method	Setting	Seen			Unseen		
		KLD ↓	SIM ↑	NSS ↑	KLD ↓	SIM ↑	NSS ↑
Cross-View-AG (Luo et al., 2022)	Weak sup.	1.5380	0.3340	0.9270	1.7870	0.2850	0.8290
Cross-View-AG+ (Luo et al., 2024)	Weak sup.	1.4890	0.3420	0.9810	1.7650	0.2790	0.8820
LOCATE (Li et al., 2023)	Weak sup.	1.2260	0.4010	1.1770	1.4050	0.3720	1.1570
WSMA (Xu et al., 2024)	Weak sup.	1.1760	0.4160	1.2470	1.3350	0.3820	1.2200
Strategies (Rai et al., 2024)	FM-guided	1.1940	0.4000	1.2230	1.4070	0.3620	1.1700
WorldAfford (Chen et al., 2024)	Instruction	1.2010	0.4060	1.2550	1.3930	0.3800	1.2250
INTRA (Jang et al., 2024)	Weak sup.	1.1990	0.4070	1.2390	1.3650	0.3750	1.2090
WSAG-PLSP baseline (Xu & Mu, 2025)	Weak sup.	0.9380	0.5030	1.4770	1.2560	0.4280	1.3460
WSAG-PLSP full (Xu & Mu, 2025)	Weak sup.	0.8900	0.5100	1.5470	1.1530	0.4370	1.4180
SelectiveCL (Moon et al., 2025)	Weak sup.	1.1240	0.4330	1.2800	1.2430	0.4050	1.3680
LoopTrans (Tang et al., 2025)	Weak sup.	1.0880	0.4450	1.3220	1.2470	0.4030	1.3150
Ours	Weak sup.	**0.8675**	**0.5218**	**1.5740**	**1.1044**	**0.4438**	**1.4760**

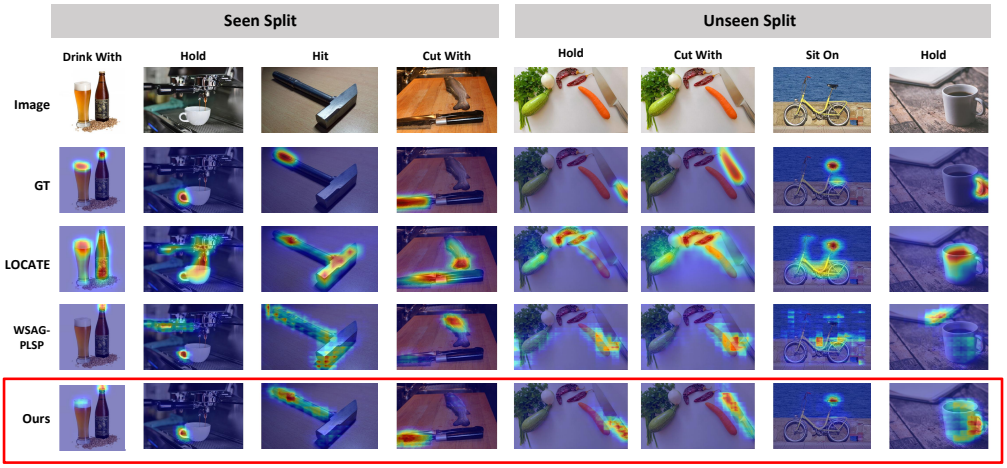

Figure 3: Qualitative comparison on AGD20K. Each column corresponds to an action–object query. From top to bottom, the rows show the input image, ground-truth heatmap, LOCATE (Li et al., 2023), WSAG-PLSP (Xu & Mu, 2025), and our method.

Table 2: Module-level ablation on AGD20K under the final controlled setting. The best result for each metric is marked in bold.

Method	AAEP	Label refine	PASE loss	Seen			Unseen		
				KLD ↓	SIM ↑	NSS ↑	KLD ↓	SIM ↑	NSS ↑
Ours	✓	✓	✓	**0.8675**	**0.5218**	**1.5740**	**1.1044**	**0.4438**	**1.4760**
w/o AAEP	×	✓	✓	0.8905	0.5012	1.4803	1.2184	0.4278	1.3400
w/o label refine	✓	×	✓	0.8815	0.5047	1.5689	1.1665	0.4325	1.4100
w/o PASE loss	✓	✓	×	0.8842	0.5102	1.5537	1.1422	0.4354	1.4166

Here, λ_{attn} and λ_{PASE} are loss weights, while $r_{\text{attn}}(e)$ and $r_{\text{PASE}}(e)$ are epoch-dependent ramp-up coefficients. $\mathcal{L}_{\text{base}}$ collects the original egocentric localization and exocentric training objectives, including the exocentric classification loss $\mathcal{L}_{\text{exo-cls}}$ (Xu & Mu, 2025).

4 Experiments

4.1 Experimental Setup

Datasets and evaluation metrics. We evaluate on AGD20K (Luo et al., 2022), which contains 20,061 exocentric interaction images and 3,755 egocentric images across 36 affordance categories.

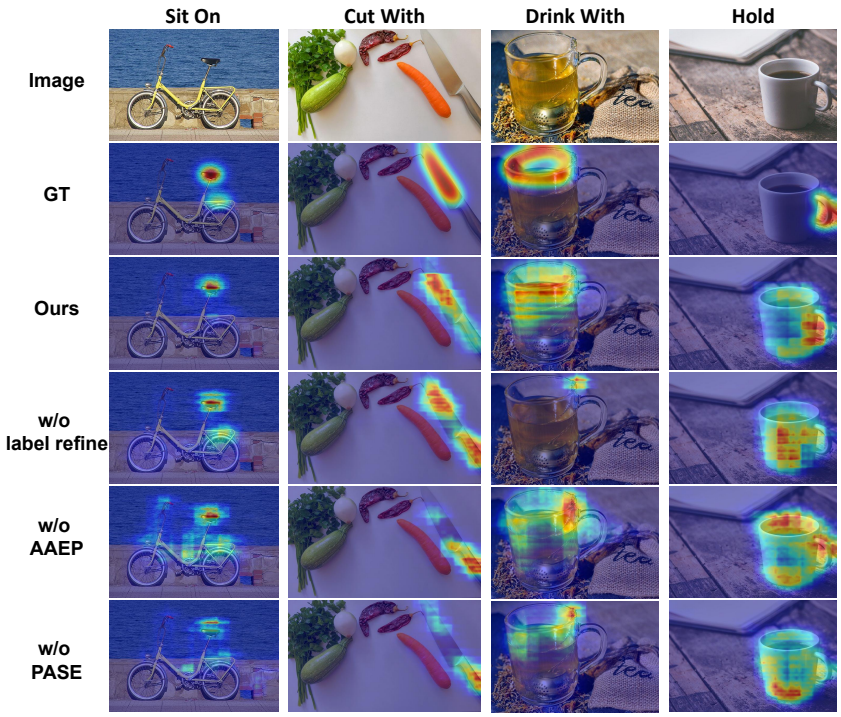

Figure 4: Qualitative comparison on AGD20K. Each column represents an action–object query, and rows show predictions of our method and ablated variants.

Following the standard protocol, we report results on the Seen and Unseen splits. The Seen split contains object categories observed during training, whereas the Unseen split evaluates transfer to novel object categories. We use Kullback–Leibler divergence (KLD), similarity (SIM), and normalized scanpath saliency (NSS). Lower KLD and higher SIM/NSS indicate better alignment with ground-truth affordance maps.

Implementation details. We build on the WSAG-PLSP framework (Xu & Mu, 2025). The CLIP ViT-B/16 image encoder (Radford et al., 2021) is initialized from pretrained weights and fine-tuned with a $0.1\times$ learning-rate multiplier, while the CLIP text encoder remains frozen with precomputed verb, noun, and part embeddings. The original training objective is retained, and the proposed modules are optimized with their corresponding losses. Specifically, AAEP uses object-constrained cosine attention with temperature $\tau = 0.07$ and JS divergence supervision with weight 0.01, activated from epoch 6 with a six-epoch linear warm-up. The pseudo-label refinement employs confidence-weighted fusion with $\gamma = 0.90$. PASE is weighted by 0.012 and activated from epoch 6 with a five-epoch linear warm-up. We apply standard data augmentation and train the model using AdamW for 30 epochs with a learning rate of 10^{-4}, weight decay of 0.01, and batch size of 20. Training requires approximately 4.8 hours on a single NVIDIA RTX 3090 GPU.

4.2 Ablation Study

We examine two design questions: (i) does action-aware exocentric aggregation remain useful when egocentric pseudo labels are refined, and (ii) do refined targets suffice, or is explicit support expansion also needed? Table 2 compares the full model with variants that remove one component while retaining the other two. The consistent degradation across both splits supports retaining all three components.

Action-aware transfer remains useful after label refinement. Removing AAEP causes the largest degradation among the tested variants across all metrics on both splits, despite retaining refinement and PASE. Thus, improving egocentric supervision does not remove the benefit of selecting

action-relevant exocentric evidence. In the *sit on* example of Fig. 4, the ablated response spreads across the bicycle, whereas the full model concentrates near the annotated contact region. This comparison supports retaining action-conditioned aggregation alongside object-level constraints.

Target precision and support coverage require separate treatment. Removing either label refinement or PASE degrades every reported metric, but their qualitative effects differ. Without refinement, the *hold* response covers more of the cup, consistent with less precise supervision. Without PASE, the *drink with* response is more fragmented despite using refined targets. These examples illustrate the design tension: responses should exclude irrelevant regions while covering coherent support around the affordance peak. The results support combining target refinement with constrained expansion rather than relying on either alone.

Configuration choice. We retain AAEP, reliability-calibrated refinement, and PASE because the full configuration performs best across the evaluated metrics and splits, with no observed metric trade-off among these variants. This supports the module combination; the ablations do not separately establish the optimal fusion strength or expansion extent. Additional qualitative results appear in Supplementary Fig. S3.

4.3 Comparison to State-of-the-Art

We compare with eleven published baselines: cross-view methods Cross-View-AG and Cross-View-AG+ (Luo et al., 2022; 2024); weakly supervised affordance grounding methods LOCATE, WSMA, INTRA, SelectiveCL, and LoopTrans (Li et al., 2023; Xu et al., 2024; Jang et al., 2024; Moon et al., 2025; Tang et al., 2025); foundation-model-guided Strategies (Rai et al., 2024); instruction-based WorldAfford (Chen et al., 2024); and the WSAG-PLSP variants (Xu & Mu, 2025).

Quantitative Comparison

As shown in Table 1, our method achieves the best performance across all six metrics. Compared with the previous state of the art, it reduces KLD by 2.5% on the Seen split and 4.2% on the more challenging Unseen split. The larger gain on Unseen compositions indicates that action-aware spatial supervision improves the transfer of functional-region evidence beyond familiar object–action associations.

Qualitative Comparison

Figure 3 compares our method with LOCATE and WSAG-PLSP on Seen and Unseen examples. Our predictions more closely follow action-relevant regions, including the blade for *cut with knife*, the grasp region for *hold cup*, and the contact region for *sit on bicycle*, while competing responses are often broader or less action-specific. Additional comparisons on two Seen panels and two Unseen panels are provided in Supplementary Figs. S1 and S2.

5 Conclusion

In this work, we present an action-aware spatial supervision framework for weakly supervised affordance grounding that improves spatial reliability across exocentric transfer, pseudo-label learning, and prediction optimization. AAEP combines action-conditioned feature aggregation with fine-grained interaction pseudo-mask supervision to emphasize action-relevant exocentric evidence. Reliability-calibrated refinement improves egocentric pseudo labels while preserving reliable original supervision, and PASE encourages compact, appearance-consistent support around reliable peaks. Experiments on AGD20K demonstrate state-of-the-art performance, particularly on the Unseen split. These results highlight the value of aligning spatial supervision with the queried action throughout the learning pipeline for accurate and transferable affordance localization.

AI USE STATEMENT

AI tools were used as auxiliary support for experimental coding, manuscript editing, LaTeX preparation, and figure refinement.

REPRODUCIBILITY STATEMENT

We provide sufficient methodological, implementation, and experimental details throughout the paper and appendix to support reproducibility of the proposed approach and reported results.

REFERENCES

Changmao Chen, Yuren Cong, and Zhen Kan. Worldafford: Affordance grounding based on natural language instructions. In *Proceedings of the IEEE 36th International Conference on Tools with Artificial Intelligence*, pp. 822–828, 2024. doi: 10.1109/ICTAI62512.2024.00120.

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

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

James J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, 1979. ISBN 9780395270493.

Ji Ha Jang, Hoigi Seo, and Se Young Chun. INTRA: Interaction relationship-aware weakly supervised affordance grounding. In *Proceedings of the European Conference on Computer Vision*, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

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*, pp. 10922–10931, 2023.

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*, pp. 2252–2261, 2022.

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. doi: 10.1007/s11263-023-01962-z.

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

Austin Myers, Ching L. Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pp. 1374–1381, 2015. doi: 10.1109/ICRA.2015.7139369.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Arushi Rai, Kyle Buettner, and Adriana Kovashka. Strategies to leverage foundational model knowledge in object affordance grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1714–1723, 2024.

Anirban Roy and Sinisa Todorovic. A multi-scale CNN for affordance segmentation in RGB images. In *Proceedings of the European Conference on Computer Vision*, pp. 186–201, 2016. doi: 10.1007/978-3-319-46493-0_12.

Erol Şahin, Maya Çakmak, Mehmet R. Doğar, Emre Uğur, and Göktürk Üçoluk. To afford or not to afford: A new formalization of affordances toward affordance-based robot control. *Adaptive Behavior*, 15(4):447–472, 2007. doi: 10.1177/1059712307084689.

Johann Sawatzky, Abhilash Srikantha, and Juergen Gall. Weakly supervised affordance detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2795–2804, 2017.

Peize Sun, Shoufa Chen, Chenchen Zhu, Fanyi Xiao, Ping Luo, Saining Xie, and Zhicheng Yan. Going denser with open-vocabulary part segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15453–15465, 2023.

Jiajin Tang, Zhengxuan Wei, Ge Zheng, and Sibei Yang. Closed-loop transfer for weakly-supervised affordance grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9530–9539, 2025.

Yuxuan Wang, Aming Wu, Muli Yang, Yukuan Min, Yihang Zhu, and Cheng Deng. Reasoning mamba: Hypergraph-guided region relation calculating for weakly supervised affordance grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27618–27627, 2025.

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, 2024. doi: 10.1609/aaai.v38i6.28451.

Peiran Xu and Yadong Mu. Weakly-supervised affordance grounding guided by part-level semantic priors. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/13b28ea4105bf409ab6b8ac8184af812-Abstract-Conference.html.

Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Grounding 3D object affordance from 2D interactions in images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10905–10915, 2023.

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*, pp. 2921–2929, 2016.

SUPPLEMENTARY MATERIAL

OVERVIEW

A Supplementary overview 12

METHOD DETAILS

B Method details 12
B.1 Action-Aware Exocentric Pooling 12
B.2 Reliability-Calibrated Pseudo-Label Refinement 13
C Peak-Anchored Support Expansion Loss 16

QUALITATIVE RESULTS

D Additional qualitative results 17

A SUPPLEMENTARY OVERVIEW

This supplementary material provides implementation details and additional qualitative results that complement the main paper. We first describe the construction of Action-Aware Exocentric Pooling (AAEP), including how action–object part prompts are organized into interaction roles and converted into exocentric spatial supervision. We then detail the Reliability-Calibrated Pseudo-Label Refinement module, which combines the original WSAG-PLSP pseudo label with action-aware spatial priors under simple reliability checks. Next, we provide the full construction of the Peak-Anchored Support Expansion Loss (PASE), including peak extraction, candidate-region selection, the mass constraint, and the main hyperparameters used in our experiments. Finally, we include additional qualitative comparisons on the Seen and Unseen splits of AGD20K, together with visual ablations that illustrate the contribution of each proposed component.

B METHOD DETAILS

B.1 ACTION-AWARE EXOCENTRIC POOLING

EXOCENTRIC SPATIAL SUPERVISION FOR AAEP

We provide additional details on the construction of the fine-grained exocentric interaction mask $\mathbf{M}_i^{\text{int}}$ used to supervise AAEP. The construction includes two steps: adapting semantic part queries to exocentric interactions and generating the corresponding interaction masks.

Interaction-role-based prompt adaptation. A single part prompt may be insufficient for robust localization in exocentric images due to viewpoint variation, occlusion, and the small size of interaction-relevant parts. We therefore expand the original semantic prompt with role-consistent alternatives while preserving its semantic prior. Specifically, based on the action–object-to-part mapping of prior work (Xu & Mu, 2025), each action is associated with one of four interaction roles: effect, grasp, support, and interface. When the original prompt is insufficiently specific, additional candidate queries consistent with the corresponding role are introduced:

$$\mathcal{Q}(v, o) = \{\Phi_{\text{PLSP}}(v, o)\} \cup \Delta_{\text{exo}}(v, o), \tag{S1}$$

where $\Phi_{\text{PLSP}}(v, o)$ denotes the original part prompt and $\Delta_{\text{exo}}(v, o)$ contains additional role-consistent candidates when needed. Tool-use actions emphasize effect regions, grasp-related actions emphasize graspable regions, support actions emphasize supporting regions, and device-operation actions emphasize interaction interfaces. This provides multiple semantically consistent descriptions of the interaction region and improves the robustness of exocentric part localization. Table S1 shows representative examples.

Table S1: Representative examples of adapting action–object prompts into interaction-role-based candidate queries for exocentric spatial supervision.

Action	Object	Source prompt	Prioritized role	Candidate queries
cut with	knife	knife blade	Effect	knife blade; cutting edge; knife tip
hold	knife	knife handle	Grasp	knife handle; knife grip
drink with	cup	cup	Effect	cup rim; cup opening
brush with	toothbrush	toothbrush	Effect	toothbrush bristles; toothbrush head
carry	skateboard	skateboard	Grasp	skateboard side; skateboard edge; skateboard deck

Exocentric interaction mask generation. For each candidate query $p \in \mathcal{Q}(v_i, o_i)$, VLPart (Sun et al., 2023) localizes the corresponding part in the exocentric image $\mathbf{I}_i^{\text{exo}}$, and SAM (Kirillov et al., 2023) converts the predicted box into a segmentation mask:

$$\begin{aligned} \mathbf{B}_{i,p}^{\text{part}} &= G_{\text{VLPart}}\left(\mathbf{I}_i^{\text{exo}}, p\right), \\ \mathbf{M}_{i,p}^{\text{int}} &= G_{\text{SAM}}\left(\mathbf{I}_i^{\text{exo}}, \mathbf{B}_{i,p}^{\text{part}}\right) \odot \mathbf{O}_i^{\text{exo}}. \end{aligned} \tag{S2}$$

The object mask $\mathbf{O}_i^{\text{exo}}$ removes responses outside the manipulated object. When multiple candidate masks are available, they are evaluated according to detection confidence, spatial extent, object overlap, and interaction-role consistency. The highest-ranked valid candidate is selected as the final interaction mask $\mathbf{M}_i^{\text{int}}$.

The selected mask is used as soft spatial supervision rather than a hard affordance annotation. If no valid candidate is available, the corresponding attention supervision is skipped. As described in the main paper, $\mathbf{M}_i^{\text{int}}$ is resized to the exocentric patch grid and normalized to supervise the action-conditioned attention of AAEP.

B.2 Reliability-Calibrated Pseudo-Label Refinement

The original pseudo labels may remain spatially broad or include regions that are weakly related to the queried action. However, applying the same correction to all pseudo labels may also distort those that are already well localized. We therefore refine the pseudo supervision in two stages. First, an action-aware spatial prior provides a coarse correction cue for constructing a more action-relevant candidate pseudo label. Second, reliability calibration controls how strongly the refined candidate contributes to the final supervision, reducing the influence of unreliable corrections.

Action-Aware Spatial Prior

A broad pseudo label may behave more like object-level supervision than affordance-specific supervision. To provide a coarse action-dependent correction, we construct a spatial prior within the detected object region. The prior captures general interaction tendencies rather than exact semantic parts and is used only as a correction cue instead of a hard affordance annotation.

Construction of the action-aware prior. Given the detected object mask O_i, we first obtain its tight bounding box $\mathcal{B}_i$. A verb-specific template $T_{v_i}(\mathcal{B}_i)$ is defined in the normalized coordinates of the box and resized to the detected object. Different action groups are associated with different coarse spatial patterns, as summarized in Table S2. The resulting action-aware spatial prior is defined as

$$A_i = A(v_i, O_i) = \text{Norm}(G_\sigma(T_{v_i}(\mathcal{B}_i) \odot O_i)), \tag{S3}$$

where G_σ denotes Gaussian smoothing and $\text{Norm}(\cdot)$ rescales the resulting map to $[0, 1]$.

The templates encode coarse interaction tendencies rather than exact semantic parts. Support actions favor supporting surfaces, grasp-related actions favor graspable boundaries, tool-use actions favor functional ends or edges, and rim-related actions favor openings. Thus, A_i provides an action-dependent spatial preference within the detected object without requiring ground-truth affordance masks.

Table S2: Action-aware spatial prior templates used for candidate pseudo-label refinement. Regions are defined relative to the detected object box.

Prior type	Representative actions	Object-region template	Purpose
Support-top	*sit on*, *lie on*, *ride*, *jump*	Upper object region	Emphasize supporting or contact surfaces while suppressing whole-object activation.
Grasp-edge	*hold*, *carry*, *lift*, *pick up*, *throw*, *catch*, *swing*, *drag*, *push*, *pack*	Left and right object edges, with weak top/bottom support	Bias refinement toward graspable or manipulable object regions.
Tool-edge	*cut with*, *brush with*, *write*, *stir*, *hit*, *stick*	Edge- or end-like regions	Reduce whole-object supervision for actions involving functional ends or edges.
Rim-top	*drink with*, *sip*, *pour*, *open*	Upper rim or opening region, with weak side support	Emphasize openings or upper contact regions.
Interface-center	*type on*, *talk on*, *text on*, *take photo*, *look out*	Central object region	Preserve interface-like regions for devices or viewing objects.
Contact-surface	*beat*, *kick*, *boxing*, *cut*, *eat*, *peel*, *wash*	Conservative object-region constraint	Avoid aggressive refinement for broad or ambiguous contact regions.

Quality-aware candidate refinement. The action-aware prior should not be applied equally to all pseudo labels. We therefore use the spatial coverage of the original pseudo label to determine the strength of refinement. Given the original pseudo label P_i and object region O_i, we compute its image-level and object-level coverage as

$$a_i^{\text{img}} = \frac{|B(P_i)|}{|I_i|}, \qquad a_i^{\text{obj}} = \frac{|B(P_i) \odot O_i|}{|O_i|}, \tag{S4}$$

where $B(P_i)$ denotes the binarized pseudo label.

We define the whole-object-like subset as

$$\mathcal{H} = \left\{ i \mid a_i^{\text{obj}} > \tau_{\text{obj}} \ \vee \ a_i^{\text{img}} > \tau_{\text{img}} \right\}, \tag{S5}$$

and among the remaining samples, the moderately broad subset as

$$\mathcal{M} = \left\{ i \notin \mathcal{H} \mid a_i^{\text{obj}} > \tau_{\text{soft}} \right\}. \tag{S6}$$

Accordingly, broad pseudo labels receive stronger action-aware correction, moderately broad labels are softly reweighted, and already localized labels receive only conservative object-region cleanup:

$$\hat{P}_i = \begin{cases} B(P_i) \odot A_i, & i \in \mathcal{H}, \\ P_i \odot (\lambda + (1-\lambda) A_i), & i \in \mathcal{M}, \\ P_i \odot C_i, & \text{otherwise}, \end{cases} \tag{S7}$$

where

$$C_i = \text{Norm}(G_{\sigma_c} * \text{Dilate}(O_i; k_c)) \tag{S8}$$

is a loose object constraint used to suppress responses outside the target object without aggressively modifying an already localized pseudo label.

We use $\tau_{\text{obj}} = 0.62$, $\tau_{\text{img}} = 0.35$, $\tau_{\text{soft}} = 0.45$, and $\lambda = 0.35$. In this way, the action-aware prior is applied more strongly when the original pseudo label exhibits whole-object-like activation, while localized responses are largely preserved.

RELIABILITY CALIBRATION

Although the action-aware prior can improve spatial localization, the refined candidate $\hat{P}_i$ is not necessarily more reliable than the original pseudo label P_i. In particular, a candidate that still covers a large portion of the image is unlikely to provide useful part-level supervision. We therefore use its spatial extent as a simple reliability cue to control its contribution to the final target.

For each refined candidate $\hat{P}_i$, we first obtain its binary support:

$$\mathbf{B}_i^{\mathrm{ref}} = \mathbb{I}\Big(\hat{P}_i \geq 0.5\Big), \tag{S9}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. Let Ω_i denote the image domain. The spatial extent of the refined candidate is measured as

$$a_i = \frac{1}{|\Omega_i|} \sum_{x \in \Omega_i} \mathbf{B}_i^{\mathrm{ref}}(x) = \frac{|\mathbf{B}_i^{\mathrm{ref}}|}{|\Omega_i|}. \tag{S10}$$

The reliability score is then defined as

$$r_i = \begin{cases} 1, & a_i \leq 0.45, \\ 0.25, & 0.45 < a_i \leq 0.80, \\ 0, & a_i > 0.80. \end{cases} \tag{S11}$$

Candidates with moderately large spatial extent are down-weighted rather than discarded, since area alone provides only a coarse reliability estimate. Candidates covering more than 80% of the image are treated as failed refinements and rejected.

The reliability score determines how strongly the refined candidate contributes to the final supervision:

$$c_i = \mathrm{clip}(\gamma r_i, 0, 1), \qquad Y_i = c_i \hat{P}_i + (1 - c_i) P_i, \tag{S12}$$

where γ controls the maximum contribution of the refined candidate. A high c_i gives more weight to $\hat{P}_i$, whereas a low c_i preserves more information from the original pseudo label P_i. We use $\gamma = 0.90$ in all experiments.

Algorithm S1 summarizes the complete quality-aware refinement and reliability-calibrated fusion procedure.

Algorithm S1 Reliability-Calibrated Pseudo-Label Refinement

Input: image I_i, action v_i, object mask O_i, original pseudo label P_i
Build action-aware spatial prior $A_i = A(v_i, O_i)$
Build loose object constraint $C_i = \mathrm{Norm}(G_{\sigma_c} * \mathrm{Dilate}(O_i; k_c))$
Compute original-label coverage a_i^{img} and a_i^{obj}
Determine quality group $\mathcal{H}$ or $\mathcal{M}$
if $i \in \mathcal{H}$ **then**
    $\hat{P}_i \leftarrow B(P_i) \odot A_i$
else if $i \in \mathcal{M}$ **then**
    $\hat{P}_i \leftarrow P_i \odot (\lambda + (1 - \lambda) A_i)$
else
    $\hat{P}_i \leftarrow P_i \odot C_i$
end if
Compute refined-candidate extent a_i
Determine reliability score r_i
$c_i \leftarrow \mathrm{clip}(\gamma r_i, 0, 1)$
$Y_i \leftarrow c_i \hat{P}_i + (1 - c_i) P_i$
Output: fused pseudo-supervision target Y_i

Algorithm S2 Label Construction with Reliability Calibration

Input: image I_i, action v_i, object mask O_i, original pseudo label P_i
Build action-aware spatial prior $A_i = A(v_i, O_i)$
Build loose object constraint $C_i = \mathrm{Norm}(G_{\sigma_c} * \mathrm{Dilate}(O_i; k_c))$
Compute image and object coverage a_i^{img} and a_i^{obj}
if $i \in \mathcal{H}$ **then**
    $\hat{P}_i \leftarrow B(P_i) \odot A_i$
else if $i \in \mathcal{M}$ **then**
    $\hat{P}_i \leftarrow P_i \odot (\lambda + (1-\lambda)A_i)$
else
    $\hat{P}_i \leftarrow P_i \odot C_i$
end if
Compute candidate reliability r_i from its spatial extent
Set $c_i \leftarrow \mathrm{clip}(\gamma r_i, 0, 1)$
$Y_i \leftarrow c_i \hat{P}_i + (1 - c_i) P_i$
Output: fused pseudo-supervision target Y_i

C Peak-Anchored Support Expansion Loss

Although pseudo-label refinement improves the spatial quality of supervision, the prediction may still concentrate on a few highly discriminative locations. PASE addresses this issue without modifying the pseudo label itself. Instead, it identifies a compact and appearance-consistent support region around reliable pseudo-label peaks and requires the prediction to assign sufficient probability mass to this region. The procedure consists of three steps: peak-based candidate construction, appearance-consistent support selection, and probability-mass regularization.

Let $\mathbf{Y}$ denote the fused pseudo-supervision target defined in Eq. (8) of the main paper. We first spatially normalize it as

$$\bar{\mathbf{Y}}_i = \frac{\mathbf{Y}_i}{\sum_j \mathbf{Y}_j + \epsilon}, \tag{S13}$$

where ϵ ensures numerical stability.

Peak-based candidate construction. We first identify reliable high-response locations in the pseudo-supervision target. Specifically, we form the eligible peak set

$$\mathcal{P} = \{i \mid \mathbf{Y}_i > 0.12\}, \tag{S14}$$

and retain the top 8% responses as Ω_{peak}. If $\mathcal{P}$ is empty, the maximum-response location is used instead.

To identify possible support around these peaks, we construct a broader positive region using the threshold $\theta_s = 0.035$:

$$\Omega_{\mathrm{sup}} = \{i \mid \mathbf{Y}_i > \theta_s\}, \qquad \Omega_{\mathrm{res}} = \Omega_{\mathrm{sup}} \setminus \Omega_{\mathrm{peak}}. \tag{S15}$$

Here, Ω_{res} only defines candidate locations for support expansion. These locations are not directly treated as additional supervision, since they may still contain noisy or visually unrelated responses.

Appearance-consistent support selection. To further select reliable candidates, we measure their appearance consistency with the high-confidence peak region. The input image is first locally averaged to reduce pixel-level noise. Let $\widetilde{\mathbf{I}}$ denote the averaged image and $\mathbf{c}_{\mathrm{peak}}$ the mean appearance over Ω_{peak}. For each candidate $i \in \Omega_{\mathrm{res}}$, we compute

$$a_i = \exp\left(-\frac{\mathrm{MSE}(\widetilde{\mathbf{I}}_i, \mathbf{c}_{\mathrm{peak}})}{\tau_a}\right), \qquad q_i = \mathbf{Y}_i a_i, \tag{S16}$$

where MSE denotes the channel-averaged squared RGB distance. The score q_i favors locations that are supported by the pseudo label and have appearance similar to the peak region.

We then retain only a compact subset of the highest-scoring candidates:

$$\Omega_{\mathrm{sel}} = \mathrm{TopK}_{K_i}(\Omega_{\mathrm{res}}; q). \tag{S17}$$

The support size K_i is dynamically constrained by both the image size and the pseudo-label support size, preventing the selected region from expanding over the entire object. In our implementation, the selected support is capped at 1.8% of the image and 20% of the pseudo-label support. Importantly, q_i is used only for candidate ranking and does not modify the pseudo-supervision values after selection.

Probability-mass regularization. Once Ω_{sel} is determined, PASE compares the probability mass assigned to this region by the pseudo-supervision target and the model prediction. Specifically,

$$\begin{aligned} m_{\mathrm{tar}} &= \sum_{i \in \Omega_{\mathrm{sel}}} \bar{\mathbf{Y}}_i, \qquad & m_{\mathrm{pred}} &= \sum_{i \in \Omega_{\mathrm{sel}}} \mathbf{P}_i, \\ m_{\mathrm{req}} &= \mathrm{clip}\left(\eta m_{\mathrm{tar}}, m_{\min}, m_{\max}\right), \end{aligned} \tag{S18}$$

where $\mathbf{P}$ denotes the spatially normalized prediction. The PASE objective penalizes the prediction only when $m_{\mathrm{pred}} < m_{\mathrm{req}}$. Therefore, PASE does not force the prediction to reproduce all selected pixels; it only prevents the prediction from collapsing onto a few isolated peaks.

For all experiments, we use a peak fraction of 0.08, peak and support thresholds of 0.12 and 0.035, respectively, and $\tau_a = 0.45$. The PASE loss is weighted by 0.012.

Algorithm S3 summarizes the complete support-selection and probability-mass regularization procedure.

Algorithm S3 Peak-Anchored Support Expansion

1: **Input:** prediction logits Z, fused pseudo-supervision target Y, and input image I
2: Extract high-confidence peak region Ω_{peak} from Y
3: Construct residual candidate region Ω_{res}
4: Compute appearance-consistency scores for residual candidates
5: Select compact support region Ω_{sel}
6: Compute target mass m_{tar} and predicted mass m_{pred} on Ω_{sel}
7: Compute the bounded requirement m_{req}
8: Apply the PASE loss if $m_{\mathrm{pred}} < m_{\mathrm{req}}$

D ADDITIONAL QUALITATIVE RESULTS

We provide additional qualitative results to complement the examples shown in the main paper. These visualizations use the same AGD20K Seen/Unseen protocol and compare the predicted affordance maps under identical action–object queries.

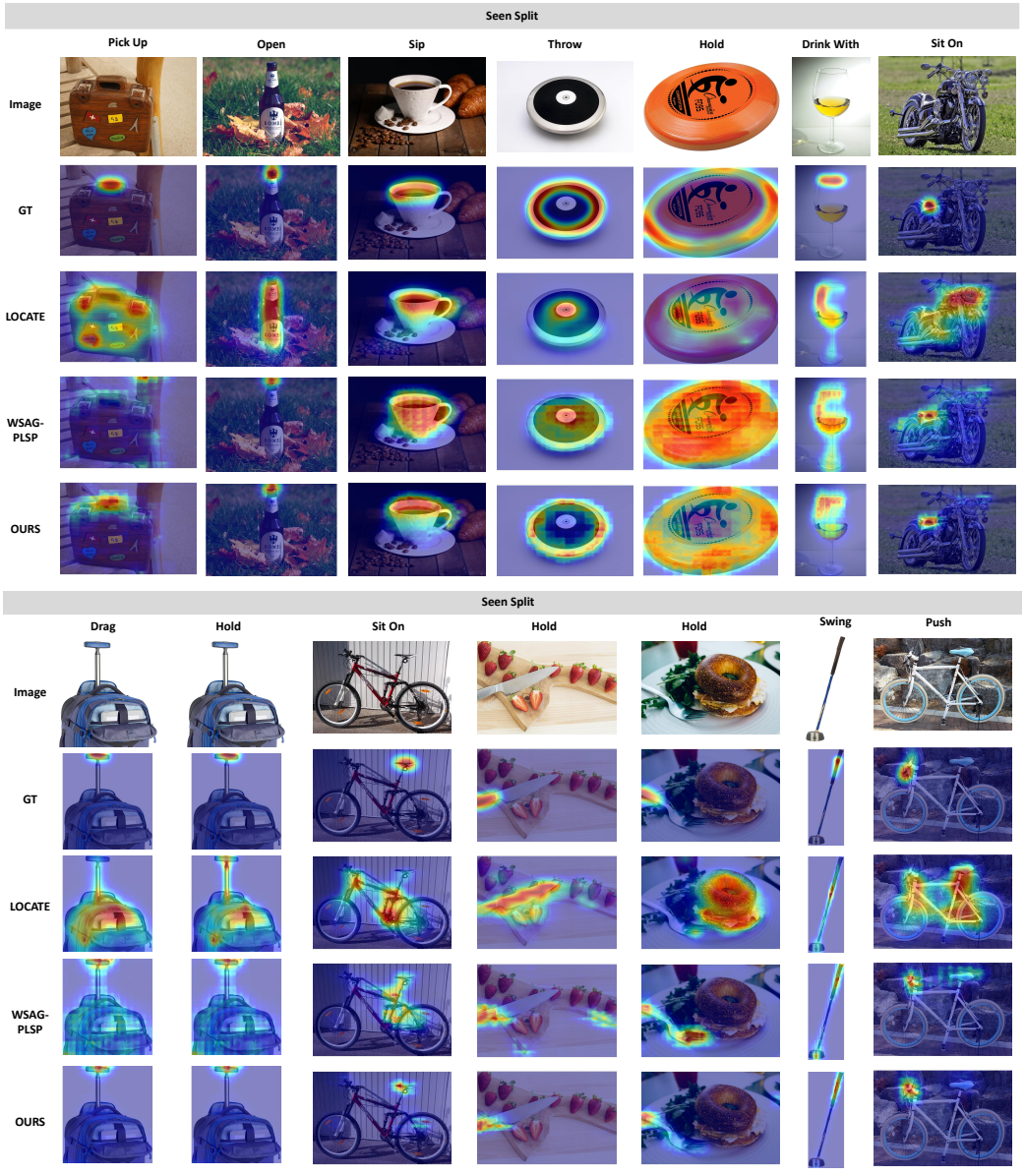

Figure S1: Additional Seen qualitative comparisons on AGD20K. Each panel compares LOCATE, WSAG-PLSP, and our method against the ground-truth affordance map.

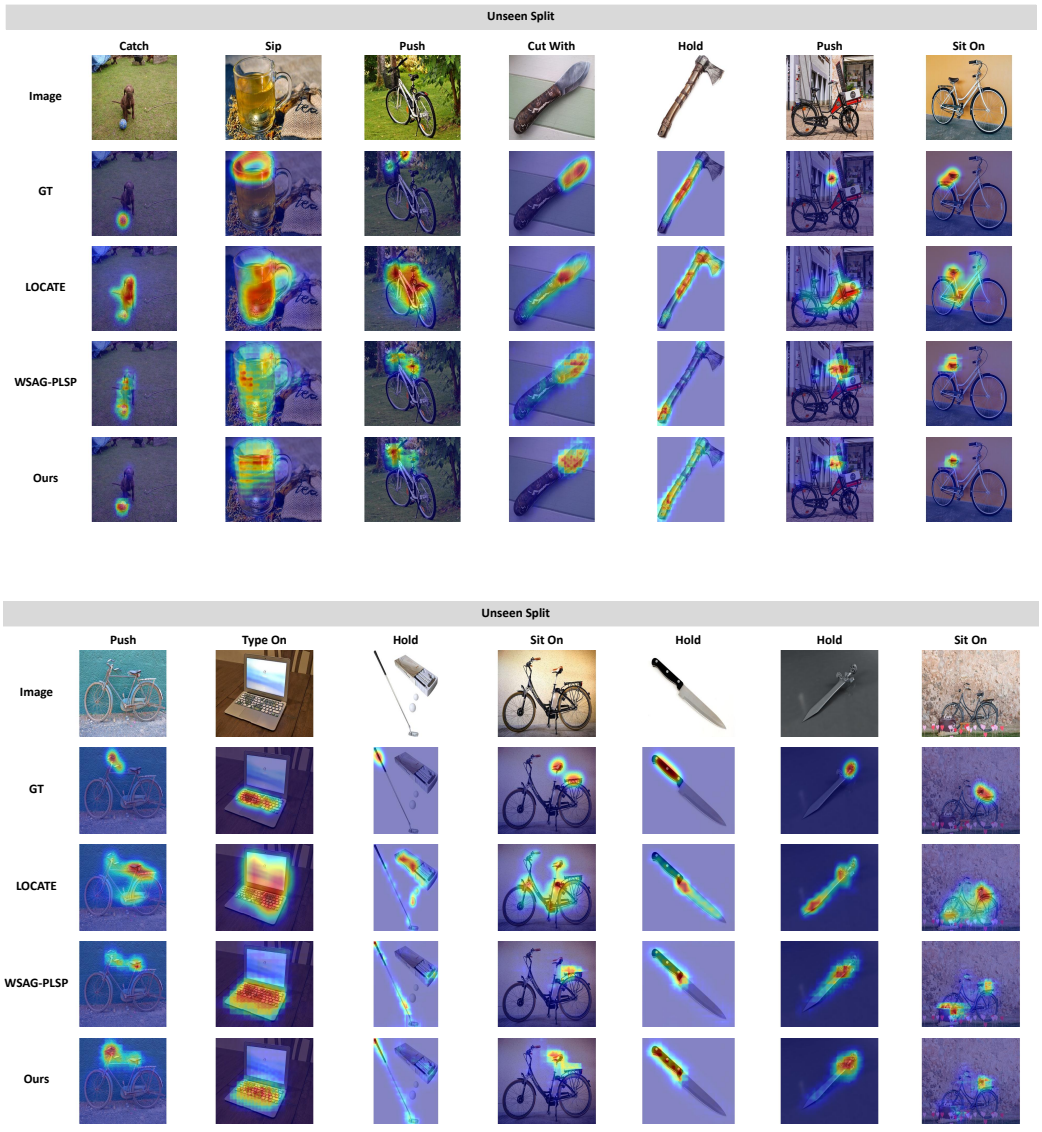

Figure S2: Additional Unseen qualitative comparisons on AGD20K. Each panel compares LOCATE, WSAG-PLSP, and our method against the ground-truth affordance map.

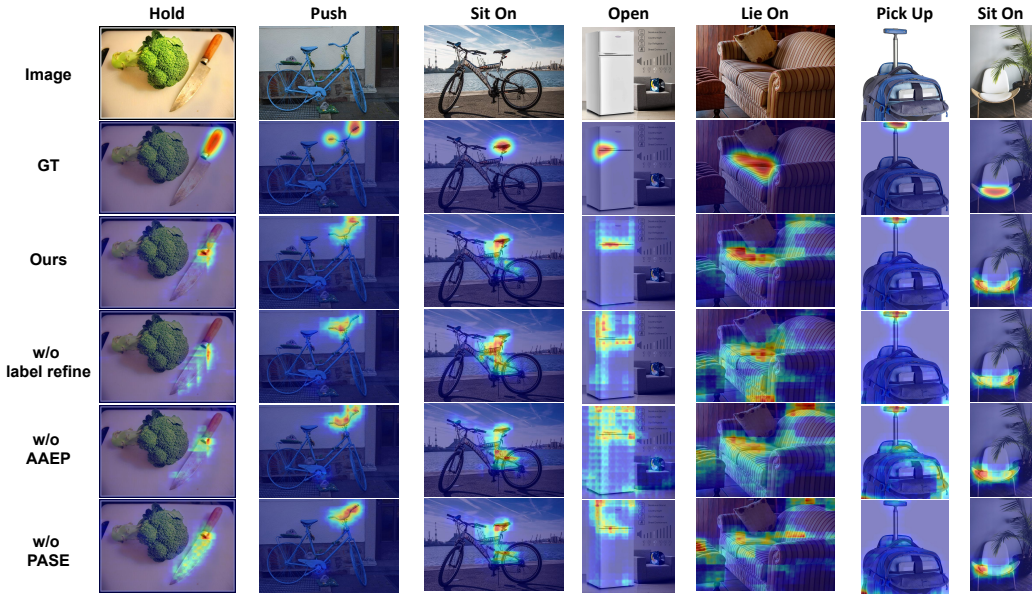

Figure S3: Additional qualitative ablations on AGD20K. Rows show the full model and variants without label refinement, AAEP, or PASE; columns show representative action–object queries.