RETHINKING CROSS-VIEW ALIGNMENT: INTERACTION-GROUNDED PART SEMANTIC GUIDANCE FOR WEAKLY SUPERVISED AFFORDANCE GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Affordance grounding aims to localize object regions that support specific actions. Since dense affordance annotation is expensive, weakly supervised affordance grounding (WSAG) has attracted increasing attention. Recent WSAG methods exploit exocentric human–object interactions as weak supervision for egocentric localization. They mainly extract visual cues from exocentric interactions and transfer them through cross-view feature or region alignment. However, large differences in appearance, viewpoint, and occlusion can make such visual transfer less reliable. Motivated by this limitation, we propose Interaction-Grounded Part Semantic Guidance (IGPSG), which explores interaction-grounded functional-part semantics as a complementary cross-view transfer signal. IGPSG uses foundation models to identify the functional part relevant to the Exo interaction and represents it with a semantic prototype. This prototype indicates which part should be transferred, while Ego visual features determine where it is localized. We further introduce Part–Verb Consensus Grounding (PVCG), which refines the transferred part query with multi-level Ego features and part–verb spatial consensus. Extensive experiments on AGD20K and HICO-IIF demonstrate consistent improvements over existing methods in both Seen and Unseen settings, supporting the effectiveness of functional-part semantics as a complementary cue for cross-view affordance grounding.

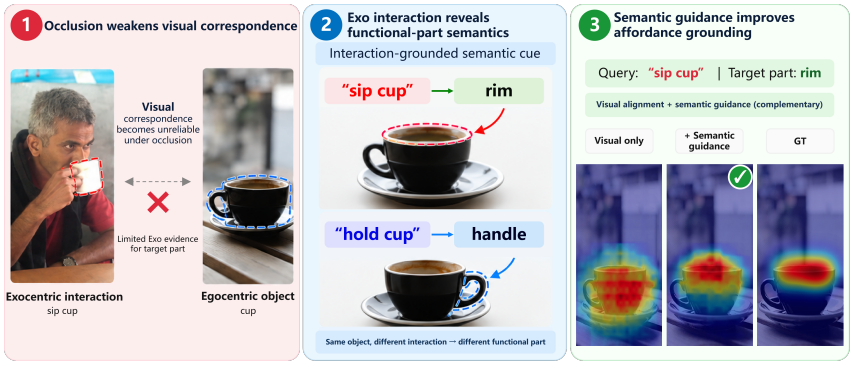


Figure 1: Motivation for interaction-grounded part semantics. Occlusion and cross-view appearance changes can weaken Exo–Ego visual correspondence (left). The Exo interaction provides action-conditioned functional-part semantics that identify the relevant part (middle), offering complementary guidance for Ego affordance localization (right).

1 INTRODUCTION

Affordance grounding (AG) aims to localize the object region that supports a queried action, such as the blade of a knife for cutting or its handle for holding (Myers et al., 2015; Do et al., 2018). The resulting spatial action cues are important for robotic manipulation and embodied interaction.

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.

Because dense region annotations are expensive, weakly supervised affordance grounding (WSAG) instead learns from cheaper image-level or interaction-level supervision (Sawatzky et al., 2017; Luo et al., 2022).

Among these weak supervisory signals, we focus on exocentric human–object interaction images. By showing an object being used, an Exo image can reveal action evidence that is absent or ambiguous in an egocentric object image. Cross-view WSAG exploits this evidence during training while requiring only the Ego image at inference. This formulation reduces annotation cost, but introduces a new challenge: which properties of the Exo interaction can be transferred reliably across views?

Cross-view WSAG has progressed from global feature transfer to increasingly fine-grained part- and interaction-level modeling (Luo et al., 2022; Li et al., 2023; Xu et al., 2024; Jang et al., 2024; Xu & Mu, 2025). Nevertheless, Exo knowledge is still largely transferred through visual correspondence between features, regions, or prototypes. Such correspondence can become unreliable under occlusion and cross-view appearance variation, allowing incorrect matches to propagate under weak supervision. As illustrated in Fig. 1, however, the Exo interaction can still reveal the functional identity of the relevant part, e.g., *sip* $\rightarrow$ rim and *hold* $\rightarrow$ handle.

This motivates using interaction-grounded functional-part semantics as a complementary cross-view transfer signal. Such semantics specify *which* functional part is relevant while being less tied to view-specific appearance; its spatial location is still determined from Ego visual features. A frozen retrieval diagnostic supports this hypothesis: on identical Exo–Ego pairs and the same Ego gallery, masked Exo visual queries achieve 3.86% Hit@1, compared with 41.78% for functional-part text queries (Section 4.2).

Based on this insight, we introduce Interaction-Grounded Part Semantic Guidance (IGPSG), consisting of Functional-Part Semantic Transfer (FPST) for semantic cross-view supervision and Part–Verb Consensus Grounding (PVCG) for Ego-side localization refinement. Experiments on AGD20K and HICO-IIF consistently improve Seen and Unseen affordance localization, validating the complementarity of interaction-grounded semantics and visual correspondence.

Our contributions are summarized as follows:

- We characterize the visual-correspondence bottleneck in fine-grained Exo-to-Ego transfer and provide a frozen, controlled diagnostic showing that functional-part text yields substantially more accurate category-level retrieval under large appearance changes.
- We introduce IGPSG, which converts object-conditioned Exo part evidence into a sample-adaptive semantic prototype and relocates its functional identity to Ego patches without assuming shared coordinates.
- We introduce a reliability-aware, query-guided Ego alignment objective that uses the semantic path only as training-time guidance. The resulting model improves matched-seed Seen and Unseen grounding while retaining the original Ego-only inference protocol.

2 Related Work

2.1 Visual Affordance Learning

Affordance describes the actions that an object or its functional parts enable for an agent (Gibson, 1979; Sahin et al., 2007). Early visual affordance studies focused on recognizing actionable object parts from geometric or appearance cues. For example, Myers et al. (Myers et al., 2015) modeled tool-part affordances from geometric features, while subsequent works extended affordance perception to RGB images, dense segmentation, and 3D interaction understanding (Roy & Todorovic, 2016; Do et al., 2018; Deng et al., 2021). These approaches established affordance learning as a visual prediction problem, but typically require explicit supervision such as pixel-level masks, part annotations, or interaction points.

To reduce annotation cost, weakly supervised methods infer affordance regions from coarse labels or indirect visual cues (Sawatzky et al., 2017; Zhou et al., 2016). While effective in avoiding dense annotations, such supervision often yields coarse localization and provides limited information about the functional parts that support an action. This motivates exploiting human–object interactions as a richer source of affordance supervision.

2.2 Weakly Supervised Affordance Grounding from Exocentric Interactions

Exocentric human–object interactions provide a natural source of weak supervision for affordance grounding. Luo et al. (Luo et al., 2022) introduced the exocentric-to-egocentric setting and AGD20K, where interaction cues observed in third-person views supervise affordance localization in egocentric images. However, such cross-view transfer remains challenging under large variations in viewpoint, appearance, occlusion, and object instance.

Subsequent works improve cross-view supervision through finer-grained correspondence and richer interaction cues. LOCATE (Li et al., 2023) introduces part-level localization and transfer, while other methods exploit multimodal cues, human–object relations, structured reasoning, selective correspondence, or closed-loop supervision (Xu et al., 2024; Yang et al., 2023; Jang et al., 2024; Luo et al., 2024; Wang et al., 2025; Moon et al., 2025; Tang et al., 2025). More recently, vision-language and segmentation models (Radford et al., 2021; Sun et al., 2023; Kirillov et al., 2023a) have enabled object parts to be localized and associated with explicit semantic concepts. PLSP (Xu & Mu, 2025) further incorporates part-level semantic priors to refine weak affordance supervision. Despite these advances, cross-view affordance transfer still relies heavily on visual feature correspondence, which can become unreliable under substantial cross-view variations. We therefore explore transferring interaction-grounded part semantics from Exo demonstrations across views as complementary guidance for Ego affordance localization.

3 Method

3.1 Problem Definition

Affordance grounding aims to localize image regions that support a queried action. We study the weakly supervised cross-view setting, where each training sample contains an egocentric image I^{ego}, a collection of exocentric human–object interaction images $\mathcal{I}^{\text{exo}} = \{I_i^{\text{exo}}\}_{i=1}^{N_e}$, and a queried verb v. The goal is to learn an affordance predictor whose inference function estimates an affordance distribution:

$$P^{\text{ego}} = f_\theta(I^{\text{ego}}, v), \tag{1}$$

where f_θ denotes the Ego localization network. During training, $\mathcal{I}^{\text{exo}}$ provides additional functional interaction guidance, while inference only requires the Ego image and verb query.

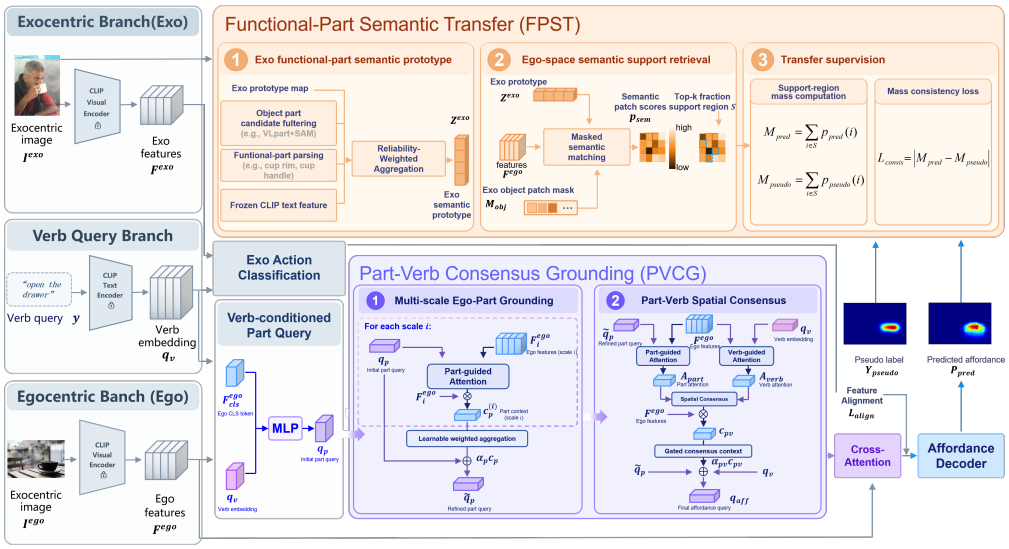


Figure 2: Overview of the proposed IGPSG framework. FPST transfers interaction-grounded part semantics from Exo demonstrations to provide additional supervision for Ego localization, while PVCG aligns part and verb semantics in the Ego feature space for affordance prediction.

3.2 METHOD OVERVIEW

As illustrated in Fig. 2, IGPSG complements visual Exo–Ego transfer with two components: **FPST**, which constructs a functional-part semantic prototype from Exo evidence and retrieves its support in the Ego object as training guidance, and **PVCG**, which refines the Ego-side part query with multi-level visual evidence and part–verb spatial consensus before affordance decoding. Following the base WSAG framework, the Exo branch includes an auxiliary action-classification head, while the Ego representation and verb embedding form the initial part query. Details of the inherited components are provided in Appendix A. At inference, the Exo branch and FPST are removed, preserving Ego-only prediction.

3.3 FUNCTIONAL-PART SEMANTIC TRANSFER (FPST)

Exo–Ego visual correspondence can become unreliable under large changes in viewpoint, appearance, and occlusion. Rather than directly transferring Exo visual regions, we use the functional identity of interaction-relevant parts as complementary semantic guidance for Ego affordance learning. Specifically, FPST extracts reliable functional-part evidence from the Exo interaction, converts it into a textual semantic prototype, and retrieves its semantic support in the Ego image to provide region-aware training supervision.

1) Exo functional-part semantic prototype. We first identify reliable functional-part evidence from the Exo interaction using pretrained open-vocabulary part and segmentation priors, VLPart Sun et al. (2023) and SAM Kirillov et al. (2023b). Given an Exo image I^{exo}, we denote by b_{obj} the bounding box of the manipulated object, obtained from the object detector or available instance metadata. For the queried verb–object pair, a predefined interaction rule activates a set of candidate functional-part prompts $\mathcal{P} = \{p_k\}_{k=1}^{K}$, where K is the number of candidate parts associated with the current interaction. For example, *cut–knife* may activate prompts such as *knife blade*, *cutting edge*, and *knife tip*. The complete prompt rules are provided in Appendix B.

VLPart and SAM generate candidate part boxes and masks, which are filtered by object association and evidence reliability (Appendix B). Each retained candidate is associated with a functional-part prompt p_k, a detection score s_k, and a reliability score r_k, where the latter accounts for its interaction role and evidence quality. Detailed prompt construction, candidate filtering, and reliability estimation are provided in Appendix B.

To transfer functional identity rather than Exo appearance or coordinates, we encode each accepted prompt with the frozen CLIP text encoder:

$$t_k = \mathrm{Norm}\left(E_{\mathrm{text}}(p_k)\right), \qquad k \in \mathcal{C}^{+}, \tag{2}$$

where t_k denotes the normalized semantic feature of candidate k.

Candidates with stronger and more reliable Exo evidence contribute more to the prototype. We therefore define

$$e_k = s_k r_k, \qquad w_k = \mathrm{softmax}_k\left(\frac{\log(e_k + \epsilon)}{\tau_{\mathrm{proto}}}\right), \tag{3}$$

where e_k measures the candidate evidence strength and τ_{proto} controls the concentration of the aggregation weights.

The Exo functional-part semantic prototype is then constructed as

$$z^{\mathrm{exo}} = \mathrm{Norm}\left(\sum_{k \in \mathcal{C}^{+}} w_k t_k\right). \tag{4}$$

Thus, z^{exo} represents the functional identity inferred from the Exo interaction, rather than an Exo spatial location. The Exo masks are used only to select and assess reliable part evidence and are not transferred to the Ego image.

2) Ego-space semantic support retrieval. We next identify where the transferred functional-part semantics appear in the Ego object. Let $F^{\mathrm{ego}} = [f_1^{\mathrm{ego}}, \ldots, f_N^{\mathrm{ego}}] \in \mathbb{R}^{N \times D}$ denote the Ego patch

features, where N is the number of patches and D is the feature dimension. Let $M_{\text{obj}}^{\text{ego}} \in \{0,1\}^N$ denote the corresponding object patch mask, where $M_{\text{obj}}^{\text{ego}}(j) = 1$ indicates that patch j belongs to the Ego object.

For each patch, we compute its semantic similarity to the Exo prototype z^{exo}:

$$a_j = \frac{\langle z^{\text{exo}}, f_j^{\text{ego}} \rangle}{\|z^{\text{exo}}\| \|f_j^{\text{ego}}\| \tau_{\text{sem}}}, \tag{5}$$

where a_j is the semantic matching score of patch j, and τ_{sem} is a temperature parameter. We normalize the scores only over patches inside the Ego object:

$$P^{\text{sem}}(j) = \frac{M_{\text{obj}}^{\text{ego}}(j) \exp(a_j)}{\sum_{n=1}^{N} M_{\text{obj}}^{\text{ego}}(n) \exp(a_n) + \epsilon}, \tag{6}$$

where $P^{\text{sem}}(j)$ denotes the semantic support score assigned to patch j, and ϵ is a small constant for numerical stability.

We retain the highest-scoring object patches as a compact semantic support set. The Top-γ selection, support-size bounds, and associated probability-mass definitions are given in Appendix C.1.

3) Support-mass transfer supervision. The retrieved support regularizes Ego prediction without being treated as a hard spatial label. Let m_{pred} and $\bar{m}_{\text{pseudo}}$ denote the predicted and calibrated pseudo probability masses over this support. For sample i, the support-mass transfer penalty is

$$\ell_i^{\text{transfer}} = \left| m_{\text{pred}} - \bar{m}_{\text{pseudo}} \right|, \tag{7}$$

where the support masses and clipped pseudo-mass are defined in Appendix C.1. This objective penalizes both insufficient and excessive probability mass within the retrieved semantic support.

3.4 Part–Verb Consensus Grounding (PVCG)

The base WSAG framework first predicts a functional-part representation q_p from the global Ego representation and the queried verb, and combines it with the verb representation q_v to construct an action-conditioned affordance query:

$$q_{\text{aff}} = q_p + q_v. \tag{8}$$

Although the subsequent affordance decoder further interacts with Ego visual features, the query itself is formed without explicitly adapting the predicted part semantics to local evidence from the current Ego instance. Moreover, directly combining q_p and q_v does not explicitly examine whether their visual evidence is spatially consistent.

To address these limitations, we introduce PVCG. Starting from the part query produced by the base model, PVCG first adapts it to the current Ego observation using local patch evidence, and then reinforces the affordance query using regions jointly supported by the refined part and the queried verb. Specifically, PVCG contains two sequential operations: (1) multi-level patch-conditioned part refinement and (2) part–verb spatial consensus.

1) Multi-level patch-conditioned part refinement. PVCG refines the initial part query with patch evidence from Transformer levels $\{6, 9, 12\}$. It computes part-conditioned attention at each level, aggregates the resulting contexts with learnable normalized level weights, and injects the multi-level context through a bounded residual. Detailed attention and level aggregation equations are provided in Appendix C.

The resulting refined query is

$$\widetilde{q}_p = q_p + \alpha_p c_p, \tag{9}$$

where α_p is a zero-initialized learnable gate bounded by $|\alpha_p| < 0.15$.

2) Part–verb spatial consensus. The refined query $\widetilde{q}_p$ identifies regions compatible with the functional part, but these regions are not necessarily consistent with the queried action. We therefore independently project the refined part query and the verb representation onto the same Ego patch space and retain their shared spatial support.

PVCG projects the refined part query and verb representation into the same Ego patch space, computes their spatial consensus, and pools the jointly supported visual context. The response maps, geometric-mean consensus, and value pooling are detailed in Appendix C.

The consensus context is injected into the affordance query through a bounded learnable residual gate:

$$q_{\text{aff}} = \widetilde{q}_p + q_v + \alpha_{pv} c_{pv}, \tag{10}$$

where α_{pv} is a bounded learnable coefficient initialized at zero. The resulting query preserves the refined part and verb semantics while incorporating consensus-aware Ego visual evidence.

4 Experiments

4.1 Experimental Setup

Dataset and evaluation. We evaluate on AGD20K (Luo et al., 2022), which contains 20,061 exocentric interaction images and 3,755 egocentric images spanning 36 affordance categories. We follow the standard Cross-View-AG protocol and report results on the Seen and Unseen splits, where Seen evaluates object categories observed during training and Unseen measures transfer to novel object categories. We additionally evaluate on HICO-IIF, a composite benchmark assembled from selected subsets of HICO-DET (Chao et al., 2018) and IIT-AFF (Nguyen et al., 2017). During training, the model receives an egocentric image, an action verb, and an associated exocentric interaction sample, while at test time only the egocentric image and action verb are used to predict the affordance heatmap. We report Kullback–Leibler divergence (KLD), similarity (SIM), and normalized scanpath saliency (NSS) between predicted and ground-truth affordance maps, where lower KLD and higher SIM/NSS indicate better localization.

Implementation details. We implement IGPSG on the WSAG-PLSP framework (Xu & Mu, 2025) with a CLIP ViT-B/16 visual encoder and the original egocentric affordance decoder. The encoder is initialized from pretrained weights, and the text features used by IGPSG are precomputed and fixed during training. Unless otherwise stated, we train the model with AdamW for 30 epochs using a learning rate of 10^{-4}, weight decay of 0.01, and batch size 20. **The detailed semantic filtering and optimization settings are provided in the supplementary material.**

4.2 Frozen Cross-View Transferability Diagnostic

We evaluate whether interaction-grounded functional-part semantics provide a transferable cue beyond Exo visual appearance using a frozen cross-view retrieval probe. The complete protocol and query definitions are deferred to Appendix F; the main retrieval results are reported below.

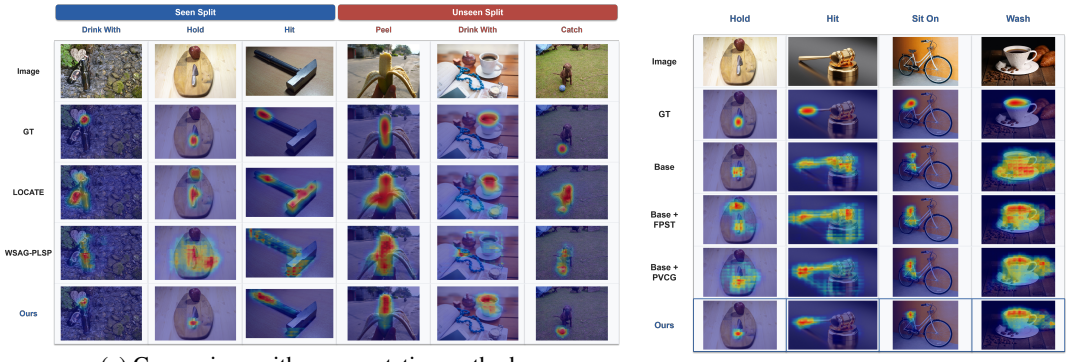
Table 1 shows that the masked visual query reaches 3.86% Hit@1, whereas the functional-part text query reaches 41.78%. The shuffled prompt reduces Hit@1 to 0.97%, indicating that the gain depends on the correct part semantics rather than merely adding an arbitrary text vector. The normalized fusion also outperforms the visual query, supporting the complementarity of semantic and visual cues.

Table 1: Diagnostic frozen cross-view retrieval on the Seen split. No grounding model or adapter is trained. All query variants use the same Exo–Ego pairs, frozen encoder, and Ego gallery.

Query	H@1	H@5	H@10	MRR	mAP
Visual	3.86	10.12	14.76	7.68	4.55
Shuffled Text	0.97	3.48	5.70	3.30	3.47
Visual+Text	28.03	58.37	68.19	42.40	20.25
Part Text	**41.78**	**86.96**	**87.97**	**62.18**	**35.68**

Table 2: Quantitative comparison on AGD20K and HICO-IIF. Published results are included for context; controlled component comparisons are reported in Table 3. Best values are shown in bold, and −− denotes unavailable results.

Method	AGD20K Seen			AGD20K Unseen			HICO-IIF		
	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
Cross-View-AG (Luo et al., 2022)	1.5380	0.3340	0.9270	1.7870	0.2850	0.8290	1.7790	0.2630	0.9460
Cross-View-AG+	1.4890	0.3420	0.9810	1.7650	0.2790	0.8820	1.8360	0.2560	0.8830
LOCATE (Li et al., 2023)	1.2260	0.4010	1.1770	1.4050	0.3720	1.1570	1.5930	0.3270	0.9660
WSMA (Xu et al., 2024)	1.1760	0.4160	1.2470	1.3350	0.3820	1.2200	1.4650	0.3580	1.0120
Strategies (Rai et al., 2024)	1.1940	0.4000	1.2230	1.4070	0.3620	1.1700	–	–	–
WorldAfford	1.2010	0.4060	1.2550	1.3930	0.3800	1.2250	–	–	–
INTRA (Jang et al., 2024)	1.1990	0.4070	1.2390	1.3650	0.3750	1.2090	–	–	–
WSAG-PLSP baseline (Xu & Mu, 2025)	0.9380	0.5030	1.4770	1.2560	0.4280	1.3460	–	–	–
WSAG-PLSP full (Xu & Mu, 2025)	0.8900	0.5100	1.5470	1.1530	0.4370	1.4180	0.8843	0.5616	1.8181
SelectiveCL (Moon et al., 2025)	1.1240	0.4330	1.2800	1.2430	0.4050	1.3680	–	–	–
LoopTrans (Tang et al., 2025)	1.0880	0.4450	1.3220	1.2470	0.4030	1.3150	–	–	–
Ours	**0.8472**	**0.5214**	**1.6081**	**1.0687**	**0.4633**	**1.5278**	**0.8059**	**0.5797**	**1.9503**



(a) Comparison with representative methods

(b) Component ablation

Figure 3: Qualitative results on AGD20K. (a) Compared with representative methods, Ours focuses more accurately on action-relevant functional regions. (b) FPST and PVCG provide complementary semantic and spatial guidance, producing the most focused response when combined.

4.3 Comparison with State-of-the-Art

4.3.1 Quantitative Comparison

Table 2 shows that IGPSG achieves the best reported performance among the listed methods on both AGD20K and HICO-IIF. On AGD20K, it consistently improves over WSAG-PLSP on both Seen and Unseen objects, with particularly clear gains on the Unseen split: KLD decreases from 1.1530 to 1.0687 (7.3% relative reduction), while SIM and NSS increase from 0.4370/1.4180 to 0.4633/1.5278 (6.0%/7.7% relative improvement). On HICO-IIF, IGPSG reduces KLD from 0.8843 to 0.8059 (8.9%) and improves SIM/NSS from 0.5616/1.8181 to 0.5797/1.9503 (3.2%/7.3%) over WSAG-PLSP. The consistent gains on the additional benchmark support the effectiveness of functional-part semantics as a complementary cue beyond visual correspondence.

4.3.2 Qualitative Comparison

Figure 3(a) shows that representative baselines often produce diffuse responses or attend to visually salient distractors, whereas IGPSG better concentrates on action-relevant parts across Seen and Unseen examples. This behavior supports the use of functional-part semantics for Ego localization. Additional results are provided in Appendix G.

4.4 Ablation Study

4.4.1 Component-wise Ablation Study

Table 3 shows that both FPST and PVCG contribute independently to IGPSG, while their combination achieves the best overall performance. Enabling either component alone improves all six

Table 3: Component ablation of IGPSG on AGD20K. FPST denotes Functional-Part Semantic Transfer and PVCG denotes Part–Verb Consensus Grounding. A checkmark indicates that the corresponding component is enabled; all other settings are fixed.

Variant	FPST	PVCG	Seen KLD / SIM / NSS	Unseen KLD / SIM / NSS
Base model	×	×	0.8875 / 0.5118 / 1.5140	1.1044 / 0.4438 / 1.4760
Base + FPST	✓	×	0.8594 / 0.5201 / 1.5917	1.0751 / 0.4619 / 1.5046
Base + PVCG	×	✓	0.8619 / 0.5150 / 1.5261	1.0829 / 0.4563 / 1.5104
Ours	✓	✓	**0.8472 / 0.5214 / 1.6081**	**1.0687 / 0.4633 / 1.5278**

Table 4: Ablation of Functional-Part Semantic Transfer on AGD20K. All variants use the same training setup and PVCG configuration. The "Ablated factor" column identifies the design choice changed from the full FPST model.

No.	Variant	Ablated factor	Seen KLD / SIM / NSS	Unseen KLD / SIM / NSS
1	**Ours**	–	**0.847** / 0.521 / **1.608**	**1.069 / 0.463 / 1.528**
2	Fixed prompt	Instance conditioning	0.889 / 0.496 / 1.593	1.114 / 0.451 / 1.519
3	Visual-only	Text semantic bridge	0.873 / **0.532** / 1.573	1.146 / 0.423 / 1.436
4	Direct alignment	Transfer route	0.877 / 0.508 / 1.573	1.124 / 0.442 / 1.473
5	w/o reliability	Reliability modeling	0.904 / 0.515 / 1.573	1.123 / 0.444 / 1.458

metrics over the Base Model, confirming that both semantic transfer and Ego-side grounding are effective individually. FPST yields consistent gains across Seen and Unseen settings, supporting the benefit of semantic guidance beyond visual correspondence, whereas PVCG provides complementary improvements by strengthening action-conditioned localization in the Ego view. Combining the two further improves every metric over either single-component variant, indicating that FPST and PVCG address different aspects of the problem and provide complementary rather than redundant benefits.

Figure 3(b) illustrates their complementary roles: FPST shifts responses toward functionally plausible parts, while PVCG sharpens localization using Ego evidence. Their combination is both semantically appropriate and spatially focused, consistent with Table 3.

4.4.2 Ablation Study of FPST Design Choices

Table 4 evaluates three key design choices in FPST: semantic transfer beyond visual correspondence, instance-conditioned and reliability-aware semantic construction, and Ego-space support supervision. The Visual-only variant removes the text bridge and transfers only the mask-pooled Exo visual feature. Its weaker Unseen performance indicates that raw appearance features are more sensitive to cross-view variation, while functional-part semantics provide complementary information about the identity of the relevant part. The Fixed prompt variant replaces instance-specific functional-part evidence with a fixed action–part prompt and leads to weaker performance, suggesting that the semantic cue should be grounded in the particular Exo interaction. Removing reliability modeling also reduces performance, showing that candidate functional parts should not be treated as equally trustworthy.

Direct alignment is further inferior to the full FPST, especially on Unseen samples. Direct alignment imposes a close correspondence between Exo and Ego features, whereas FPST constrains the probability mass assigned to retrieved Ego support without enforcing dense cross-view feature correspondence. Overall, these results show that FPST benefits not merely from introducing text semantics, but from transferring reliable, interaction-conditioned semantics as complementary guidance for Ego localization.

4.4.3 Ablation Study of PVCG

Table 5 examines the complementarity between part refinement and P–V consensus, as well as the benefit of multi-level Ego features. Here, $L6$, $L9$, and $L12$ denote the corresponding CLIP Transformer blocks. At $L12$, part refinement is stronger on Seen samples, whereas P–V consensus

Table 5: Ablation of PVCG operations and Ego feature levels on AGD20K. All variants retain FPST. Part Refinement denotes patch-conditioned part refinement, and P–V Consensus denotes part–verb spatial consensus. $L6$, $L9$, and $L12$ are outputs of the corresponding CLIP Transformer blocks.

No.	Ego Levels	Part Refinement	P–V Consensus	Seen			Unseen		
				KLD↓	SIM↑	NSS↑	KLD↓	SIM↑	NSS↑
1	$L12$	✓	–	0.8698	0.5135	1.5599	1.1202	0.4396	1.4655
2	$L12$	–	✓	0.8881	0.5133	1.5539	1.0944	0.4446	1.5204
3	$L12$	✓	✓	0.8630	0.5163	1.5810	1.0812	0.4514	1.5017
4	$L9 + L12$	✓	✓	0.8501	0.5207	1.5925	1.0764	0.4560	1.5101
5	$\mathbf{L6 + L9 + L12}$	✓	✓	**0.8472**	**0.5214**	**1.6081**	**1.0687**	**0.4633**	**1.5278**

performs better on Unseen samples (Rows 1–2), indicating complementary strengths in instance adaptation and action-conditioned generalization. Their combination improves most metrics (Row 3), supporting their joint use.

Introducing earlier Ego features yields consistent gains across all six metrics (Rows 3–5). While the improvement from two to three levels is modest on Seen KLD and SIM, all Unseen metrics continue to improve, suggesting that earlier features provide complementary fine-grained evidence for cross-instance generalization. We therefore adopt $L6{+}L9{+}L12$ as the final PVCG configuration.

5 Conclusion

We investigate interaction-grounded functional-part semantics as a complementary signal for cross-view affordance grounding when Exo–Ego visual correspondence is unreliable. IGPSG combines FPST, which transfers reliability-aware functional-part semantics from Exo interactions as training guidance, with PVCG, which strengthens Ego-side localization through multi-level visual refinement and part–verb spatial consensus. Experiments on AGD20K and HICO-IIF consistently demonstrate the effectiveness of the proposed approach. The results suggest that semantics are most useful for indicating which interaction-relevant part should be sought, while Ego visual evidence determines where it is localized, enabling complementary semantic transfer without introducing additional Exo dependence at inference.

AI use statement

AI tools were used to assist with experimental code development, 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. Specifically, the appendix documents the base query construction (Appendix A), functional-part candidate generation (Appendix B), detailed FPST and PVCG operations (Appendix C), and the complete training and inference procedure (Appendix D).

References

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*, pp. 381–389. IEEE, 2018.

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, 1979.

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, 2023a.

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

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 Ferm"uller, 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.

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*, pp. 5908–5915. IEEE, 2017.

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 *Proceedings of the International Conference on Machine Learning*, pp. 8748–8763, 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 Sahin, Maya Cakmak, Mehmet R. Dogar, Emre Ugur, and Gokt"urk Ucoluk. 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

METHOD

A Action-Discriminative Exo Encoding and Verb-Conditioned Query 12

B Exo Functional-Part Candidate Generation and Metadata 13

C Detailed FPST and PVCG Equations 14

D Training and Inference Procedure 15

EXPERIMENTAL DETAILS

E Detailed FPST Ablation Definitions 15

DIAGNOSTIC STUDIES

F Detailed Diagnostic Analyses 16

QUALITATIVE RESULTS

G Additional Qualitative Results 18

METHOD

A ACTION-DISCRIMINATIVE EXO ENCODING AND VERB-CONDITIONED QUERY

This section details the two components summarized in the method overview. They belong to the underlying cross-view grounding pathway: the first provides action supervision for Exo representation learning during training, whereas the second constructs the Ego query used for affordance prediction.

A.1 EXO ACTION CLASSIFICATION

The Exo Action Classification module in Fig. 2 applies an action-classification head without requiring region-level annotations. The Exo encoder first pools visual evidence within the manipulated-object mask. When action-aware pooling is enabled, this object evidence is further used to aggregate action-relevant Exo patches, producing an Exo interaction token $z_i^{\text{exo}} \in \mathbb{R}^{512}$. A two-layer MLP predicts the action logits:

$$a_i^{\text{exo}} = W_2\ \text{GELU}(W_1 z_i^{\text{exo}})\,, \qquad a_i^{\text{exo}} \in \mathbb{R}^{36}, \tag{11}$$

where the classifier dimensions are $512 \to 256 \to 36$. It is optimized with the action label y_i:

$$\mathcal{L}_{\text{exo-cls}} = \frac{1}{B}\sum_{i=1}^{B} \text{CE}(a_i^{\text{exo}}, y_i)\,. \tag{12}$$

The resulting classification supervision encourages the pooled Exo token to retain action-discriminative information. When reliability weighting is enabled, it only reweights the per-sample cross-entropy terms. This classification loss is included in $\mathcal{L}_{\text{base}}$.

A.2 VERB-CONDITIONED PART QUERY CONSTRUCTION

Let $f_{\text{cls}}^{\text{ego}} \in \mathbb{R}^{512}$ be the global Ego token and $q_v \in \mathbb{R}^{512}$ the frozen CLIP embedding of the queried verb. The model first predicts a noun representation and then combines it with the verb embedding to obtain the initial part query:

$$q_n = h_n(\text{sg}\,(f_{\text{cls}}^{\text{ego}}))\,, \qquad q_p^{(0)} = h_p([q_n \| q_v])\,, \tag{13}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient, h_n is a $512 \to 512 \to 512$ MLP, and h_p is a $1024 \to 512 \to 512$ MLP; both use GELU activations. The two representations are supervised by frozen noun and part text features, e_n and e_p, using cosine losses:

$$\mathcal{L}_{\text{noun}} = 1 - \cos(q_n, e_n), \qquad \mathcal{L}_{\text{part}} = 1 - \cos(q_p^{(0)}, e_p). \tag{14}$$

In the base pathway, the affordance query is formed by adding the predicted part query and verb embedding. In IGPSG, PVCG first grounds $q_p^{(0)}$ with multi-level Ego patches and then combines the grounded part representation with q_v through part–verb spatial consensus. This query pathway remains available at inference and requires no Exo input.

B Exo Functional-Part Candidate Generation and Metadata

This section details how the functional-part evidence used by FPST is constructed. The procedure is applied to each Exo interaction and its manipulated object. It produces candidate metadata and frozen text features; it does not introduce an additional inference branch.

B.1 Prompt construction and candidate generation

Given the queried verb and object, we form a verb–object-aware set of functional-part prompts. For example, a cutting interaction with a knife may produce prompts such as "knife blade", "cutting edge", and "knife tip". Prompts are assigned interaction roles, including effect, grasp, support, and interface parts.

The candidate-generation sequence is

$$\text{Exo image and object box} \;\to\; \text{SAM object mask}, \qquad \text{part prompt} \;\to\; \text{VLPart box} \;\to\; \text{SAM part mask}. \tag{15}$$

VLPart uses the text prompt to detect a candidate part box, while SAM refines the object and part boxes into masks. The part mask is not transferred directly to the Ego view.

B.2 Object association and reliability filtering

For a candidate part mask M_k^{part} and manipulated-object mask $M_{\text{obj}}^{\text{exo}}$, we compute

$$\text{Assoc}_k = \frac{\left| M_k^{\text{part}} \cap M_{\text{obj}}^{\text{exo}} \right|}{\left| M_k^{\text{part}} \right| + \epsilon}. \tag{16}$$

Candidates with insufficient object association, implausible area, low VLPart confidence, or low reliability are rejected. The reliability score is an evidence-quality heuristic rather than a calibrated probability:

$$r_k = \alpha_s q_k^{\text{score}} + \alpha_i q_k^{\text{inside}} + \alpha_a q_k^{\text{area}} + \alpha_r q_k^{\text{role}}, \qquad \sum_u \alpha_u = 1. \tag{17}$$

The accepted metadata records the prompt, role, detection score, reliability, area statistics, object-inside ratio, object overlap, acceptance flag, and the corresponding Exo instance identifiers.

B.3 CLIP feature cache and training-time use

For every accepted prompt, a frozen CLIP text encoder produces a normalized feature. These features can be cached before training and retrieved by prompt identifier. During training, the data loader selects role-compatible accepted candidates, combines their detection scores and reliabilities, and forms z^{exo} using Eq. equation 4. The prototype is then matched with Ego object patches as described in Section 3.3.

In summary, Exo masks are used to filter and assess candidate semantics, whereas z^{exo} is the quantity that interacts with Ego patch features. The Exo spatial masks and coordinates are never copied to the Ego view.

C Detailed FPST and PVCG Equations

This section collects implementation-level equations omitted from the main paper for readability.

For FPST, VLPart and SAM produce candidate boxes and masks:

$$\{(b_k, s_k)\}_{k=1}^{K} = \text{VLPart}(I^{\text{exo}}, \mathcal{P}), \qquad M_{\text{obj}}^{\text{exo}} = \text{SAM}(I^{\text{exo}}, b_{\text{obj}}), \quad M_k^{\text{part}} = \text{SAM}(I^{\text{exo}}, b_k). \tag{18}$$

Candidates are retained when

$$\mathcal{C}^+ = \left\{ k \mid \text{Assoc}(M_k^{\text{part}}, M_{\text{obj}}^{\text{exo}}) \geq \delta_{\text{obj}} \right\}. \tag{19}$$

The semantic support map and compact support set are

$$a_j = \frac{\langle z^{\text{exo}}, f_j^{\text{ego}} \rangle}{\|z^{\text{exo}}\| \|f_j^{\text{ego}}\| \tau_{\text{sem}}}, \qquad P^{\text{sem}}(j) = \frac{M_{\text{obj}}^{\text{ego}}(j) e^{a_j}}{\sum_n M_{\text{obj}}^{\text{ego}}(n) e^{a_n} + \epsilon}, \tag{20}$$

$$\mathcal{R}_{\text{sem}} = \text{Top}_\gamma \left(P^{\text{sem}} \mid M_{\text{obj}}^{\text{ego}} = 1 \right). \tag{21}$$

For PVCG, let $\mathcal{L}_{\text{feat}} = \{6, 9, 12\}$. At each level,

$$a_{p,j}^{(i)} = \text{softmax}_j \left(\frac{\text{sim}(W_q q_p, W_k f_j^{(i)})}{\tau_p} \right), \quad c_p^{(i)} = \sum_j a_{p,j}^{(i)} \phi_{\text{val}}(f_j^{(i)}), \tag{22}$$

with $c_p = \sum_{i \in \mathcal{L}_{\text{feat}}} w_i^p c_p^{(i)}$ and $w_i^p = \text{softmax}(\lambda^p)_i$. The part and verb responses, consensus weights, and pooled context are

$$a_{p,j} = \text{softmax}_j \left(\frac{\text{sim}(W_{pq} \widetilde{q}_p, W_{pk} f_j)}{\tau_p} \right), \qquad a_{v,j} = \text{softmax}_j \left(\frac{\text{sim}(W_{vq} q_v, W_{vk} f_j)}{\tau_v} \right), \tag{23}$$

$$s_j = \frac{\sqrt{a_{p,j} a_{v,j}}}{\sum_n \sqrt{a_{p,n} a_{v,n}} + \epsilon}, \qquad c_{pv} = \sum_j s_j \phi_{\text{val}}(f_j). \tag{24}$$

C.1 Batch Aggregation of Support-Mass Transfer

For a mini-batch, let ℓ_i^{transfer} denote the support-mass penalty for sample i defined in Eq. equation 7. We include a sample only when valid semantic support is available and use its transfer reliability to modulate the contribution:

$$\mathcal{L}_{\text{transfer}} = \frac{\sum_i \nu_i g_i \ell_i^{\text{transfer}}}{\sum_i \nu_i + \epsilon}. \tag{25}$$

Here, $\nu_i \in \{0, 1\}$ indicates support validity and g_i is the reliability of the transferred semantic cue. The normalization averages over valid samples, while g_i downweights less reliable semantic evidence.

C.2 Training Objective Details

The overall training objective combines the standard WSAG objective with the semantic transfer supervision introduced by FPST:

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda_{\text{transfer}}(t) \mathcal{L}_{\text{transfer}}, \tag{26}$$

where $\mathcal{L}_{\text{base}}$ denotes the training objective inherited from PLSP (Xu & Mu, 2025), t denotes the current training epoch, and $\lambda_{\text{transfer}}(t)$ is a scheduled weight that gradually introduces the semantic transfer supervision.

Since the semantic support retrieved by FPST provides approximate functional-part guidance rather than pixel-level ground truth, we regularize the probability mass assigned to the retrieved support instead of matching a dense cross-view map. The per-sample support-mass penalty is defined in Eq. 7; its batch aggregation and reliability weighting are given in Appendix C.1. This design treats the transferred semantics as complementary guidance rather than pixel-level supervision.

Algorithm 1 Training and inference procedure of IGPSG

Require: Ego image I^{ego}, Exo interaction I^{exo}, queried verb v, and weak action labels
Ensure: Trained Ego affordance predictor f_θ

```
Precompute Exo functional-part candidates, reliability scores, and frozen CLIP text features
for each training epoch t do
    Compute the scheduled transfer weight λ_transfer(t)
    for each training batch do
        Encode I^ego and I^exo
        Construct the verb-conditioned part query q_p^(0)
        Aggregate valid Exo candidates into z^exo
        Refine q_p^(0) using PVCG and Ego features
        Predict the affordance map Ŷ
        Compute L_base
        if valid semantic candidates exist then
            Compute L_transfer
        else
            Set L_transfer = 0
        end if
        Compute L = L_base + λ_transfer(t) L_transfer
        Update θ using L
    end for
end for
At inference, discard the Exo branch
Predict Ŷ = f_θ(I^ego, v)
return f_θ
```

D Training and Inference Procedure

Algorithm 1 summarizes the complete training and inference workflow. The Exo branch and FPST provide training-time guidance only; they are discarded during inference.

EXPERIMENTAL DETAILS

E Detailed FPST Ablation Definitions

All FPST ablations use the same training setup and fixed PVCG configuration. Each variant changes one semantic-transfer design choice relative to the full model, except that Direct Alignment is a route-level control described below. These controls cover the semantic-source, evidence-quality, and transfer-route choices in FPST. Fixed Prompt, Visual-only, Object-level Text, and Without Reliability examine how the Exo cue is constructed; Direct Alignment examines how the cue is used to supervise the Ego prediction.

Fixed prompt. This variant replaces the instance-conditioned functional-part semantics with one fixed action–part prompt. It tests whether selecting the functional part from the current Exo interaction is necessary beyond using category-level language semantics.

Visual-only. This variant removes the text semantic bridge and uses the mask-pooled Exo visual feature as the transfer cue. It tests whether raw Exo appearance alone is sufficient for cross-view transfer. Its higher Seen SIM but lower Unseen performance indicates that appearance correspondence can fit familiar samples but is less reliable under category and viewpoint changes.

Direct alignment. This variant directly aligns the mask-pooled Exo feature with the Ego affordance token. It bypasses the functional-part text prototype and the support-region transfer route, and therefore serves as a route-level control for FPST. It should not be interpreted as a loss-only ablation.

Without reliability. This variant retains structurally valid semantic-transfer samples but removes the reliability-based candidate filtering or weighting. It tests whether semantic transfer benefits from evidence-quality modeling rather than from injecting all available Exo candidates.

Object-level text control. As an additional control, we query the role-text bank with a whole-object visual token instead of a part-mask-pooled feature. This variant obtains $0.884/0.506/1.576$ on Seen and $1.128/0.440/1.479$ on Unseen. Its lower Unseen SIM and NSS than the full model suggest that object identity alone does not sufficiently identify the active functional part. We omit this control from the main table because it tests semantic granularity rather than one of the primary FPST transfer operations.

DIAGNOSTIC STUDIES

F Detailed Diagnostic Analyses

The main paper reports the diagnostic protocol, retrieval table, and primary interpretation. This appendix provides additional analyses of appearance variation, evidence incompleteness, paired rank changes, and qualitative cases.

F.1 Diagnostic Protocol

This probe uses 912 unique Ego images, 1,490 unique Exo part observations, and 2,669 fixed Exo–Ego pairs. We compute all Ego global features, masked Exo visual features, and part-prompt text features once using the same frozen CLIP ViT-B/16 encoder. All four query variants are evaluated against the identical Ego bank. The rank of a query is the first position occupied by any Ego image with the same verb–noun category; mAP additionally accounts for all positives in the bank.

The strict protocol is reported only on Seen. Under the standard Unseen split, test interactions have no overlapping verb–noun categories with the training Exo pool, so enforcing the same positive definition yields no valid pairs. Constructing a same-verb, cross-noun probe would answer a different question and is therefore not mixed into Table 1.

This experiment measures preservation of category-level functional semantics. Because the prompt names a functional part and positives are defined by the interaction category, it should not be interpreted as evidence that text alone solves dense localization. The shuffled-prompt control rules out an arbitrary text-vector effect, while the FPST ablation tests whether the semantic prior improves the actual localization model.

F.2 Frozen Queries and Retrieval Definitions

The masked visual and functional-part text queries are defined as

$$q_{\text{vis}} = \text{Norm}\left(\text{MaskedPool}(E_I(I^{\text{exo}}), M_{\text{part}})\right), \tag{27}$$

and

$$q_{\text{text}} = \text{Norm}(E_T(p_{\text{part}})). \tag{28}$$

where E_I and E_T are the frozen CLIP image and text encoders, M_{part} is the detected Exo functional-part mask, and p_{part} is its associated prompt.

Appearance variation. The semantic advantage becomes larger as the Exo–Ego appearance gap increases. In the largest-gap quartile, Visual obtains 0.90Part Text obtains 39.07Text recovers a Visual failure increases from 61.6quartile to 86.7

F.3 Evidence Incompleteness and Appearance Gap

For an Exo observation with detected part area A_{part} and target object area A_{object}, we define a functional-part evidence incompleteness proxy

$$o = 1 - \min\left(1, \frac{A_{\text{part}}/A_{\text{object}}}{Q_{90}(A_{\text{part}}/A_{\text{object}} \mid v, n, p)}\right), \tag{29}$$

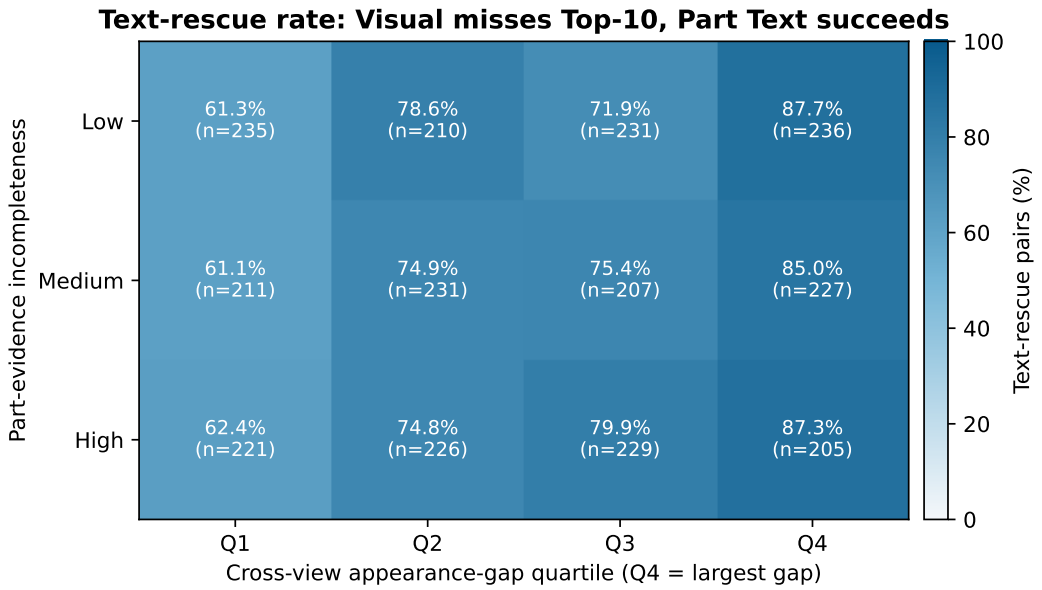


Figure 4: Text-rescue rate for fixed Exo–Ego pairs. A rescue denotes a pair for which the masked Visual query misses all positives in the Top-10 but the functional-part Text query succeeds. Q4 contains the largest Exo–Ego appearance gaps. Evidence incompleteness is an automatically computed part-evidence proxy, not physical-occlusion ground truth.

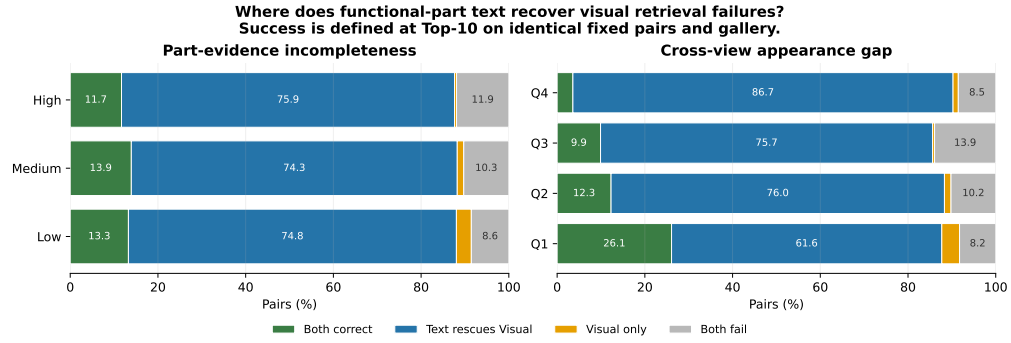


Figure 5: Paired Top-10 outcome composition under part-evidence incompleteness and cross-view appearance gap. The fraction labeled Text-rescues-Visual grows most clearly with appearance gap, reaching 86.7% in Q4.

where the reference quantile is computed within each verb–noun–part group. We use Low ($o < 0.2$), Medium ($0.2 \leq o < 0.5$), and High ($o \geq 0.5$). This value measures the incompleteness of detected functional-part evidence; it is not an amodal human-occlusion annotation. Samples without a valid object-conditioned candidate are excluded from this probe rather than counted as highly occluded. The appearance gap is $1 - \cos(g^{\text{ego}}, g^{\text{exo}})$ using frozen global CLIP features and is divided into split-specific quartiles.

Figure 5 decomposes Top-10 outcomes into both queries succeeding, Text rescuing Visual, Visual alone succeeding, and both failing. Across Low, Medium, and High evidence incompleteness, Text rescues respectively 74.8%, 74.3%, and 75.9% of all pairs. This stability supports robustness to incomplete evidence, but the natural-data proxy does not show a monotonic occlusion effect. We therefore reserve the stronger causal claim for a future controlled synthetic-occlusion intervention.

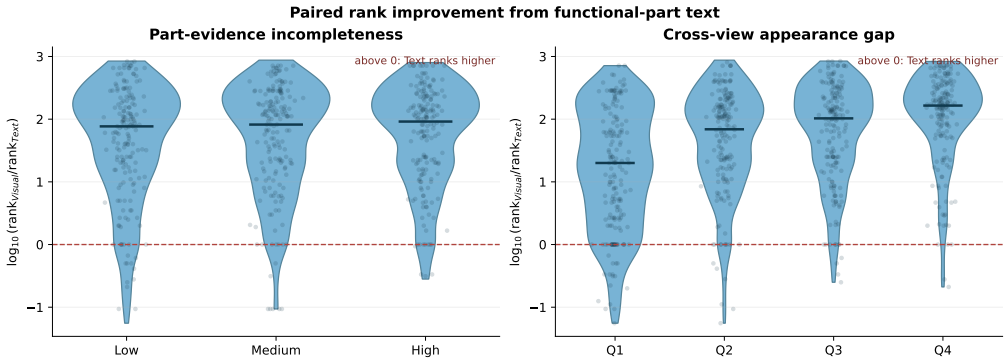


Figure 6: Distribution of paired rank improvements. Values above zero indicate that the functional-part Text query ranks the first positive Ego image above the masked Exo Visual query.

F.4 Paired Rank and Qualitative Analysis

For each fixed pair we additionally measure

$$\Delta_{\text{rank}} = \log_{10} \frac{\text{rank}_{\text{Visual}}}{\text{rank}_{\text{Text}}}. \tag{30}$$

Positive values indicate a better rank for Part Text. As shown in Figure 6, the median Δ_{rank} increases from 1.301 in Q1 to 2.217 in Q4. This pairwise analysis avoids conflating the appearance-gap result with changes in gallery composition.

QUALITATIVE RESULTS

G Additional Qualitative Results

This section provides additional qualitative comparisons across actions, object categories, and Seen/Unseen settings. In both splits, our model produces more concentrated responses on the action-relevant functional part, while LOCATE and WSAG–PLSP often activate broader object regions or visually salient distractors. The examples complement the quantitative results by illustrating that the semantic guidance remains useful for both familiar and novel object categories.

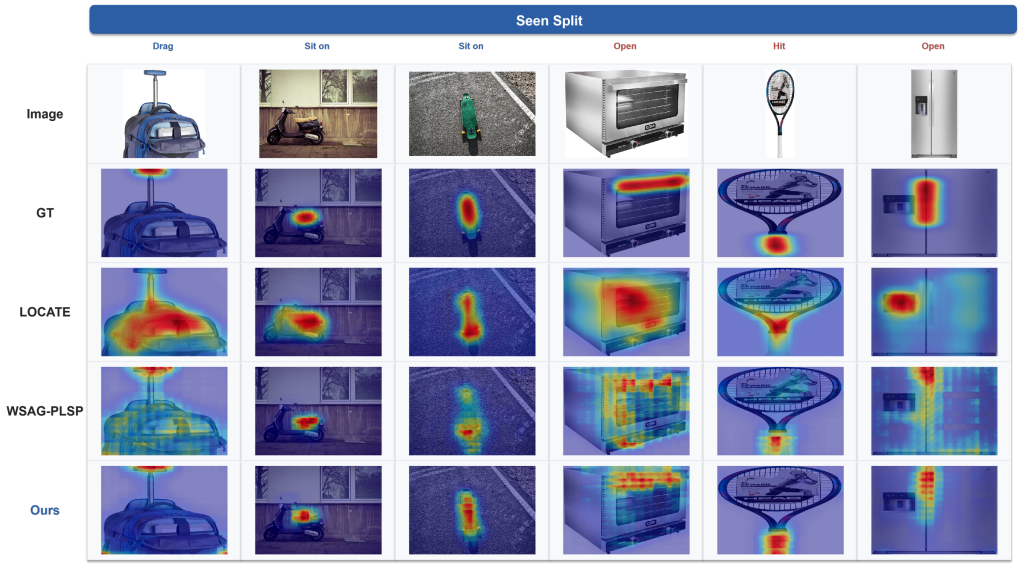


Figure 7: Additional qualitative comparisons on the AGD20K Seen split.

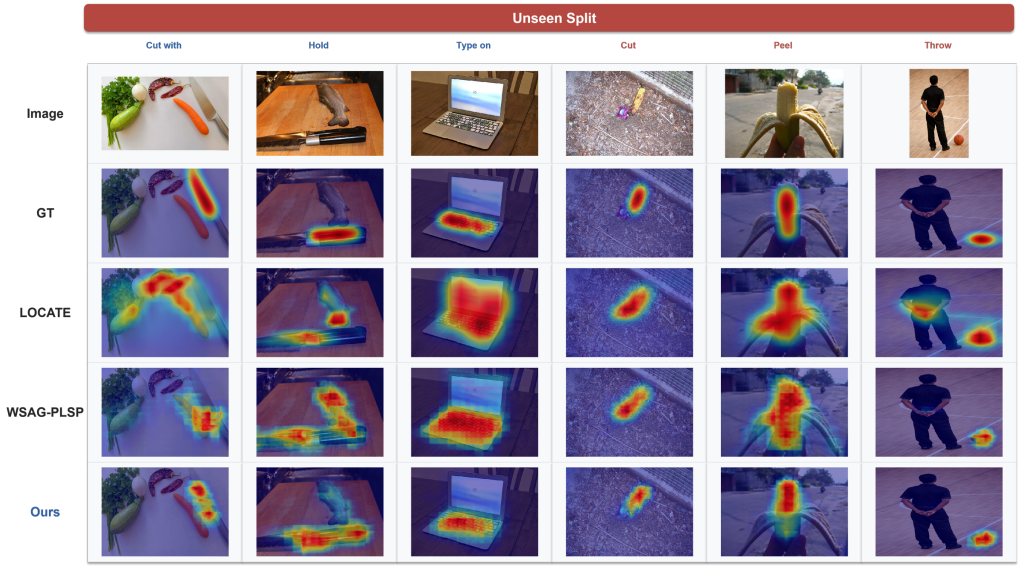


Figure 8: Additional qualitative comparisons on the AGD20K Unseen split.