AFFIMOE: FINE-GRAINED TEXT-VISUAL ALIGNMENT VIA TOKEN-WISE EXPERT ROUTING FOR LANGUAGE-GROUNDED AFFORDANCE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Affordance grounding aims to localize object regions that enable a specified action, requiring models to connect high-level task intent with fine-grained functional structures. Recent LLM- and VLM-based approaches improve open-world affordance reasoning through language supervision, while RAGNet further extends this paradigm with human-like instructions and reasoning-aware segmentation. However, these methods still compress rich multimodal semantics into compact affordance representations, leaving the correspondence between language-level actions and fine-grained functional regions largely implicit. To address this limitation, we propose **AffiMoE**, a language-grounded affordance reasoning framework that explicitly strengthens fine-grained text-visual alignment through token-wise expert routing. Specifically, AffiMoE uses cross-attention to align action semantics with dense visual features, producing localized representations that better capture action-relevant object parts. Since precise affordance grounding requires both global object context and local functional evidence, we further treat globally aligned multimodal features and locally aligned visual features as complementary experts, and introduce a token-wise mixture-of-experts router to adaptively integrate them at each spatial location. In this way, AffiMoE preserves coherent high-level reasoning while improving the localization of action-relevant functional regions. To rigorously evaluate this capability, we introduce a new benchmark, and extensive experiments show that AffiMoE achieves state-of-the-art performance across multiple benchmarks.

1 INTRODUCTION

Affordance grounding aims to identify object regions that support a specific action (Luo et al., 2022), addressing a fundamentally action-centric question: *where can an agent interact with an object to accomplish a task?* Early approaches (Li et al., 2023a; Luo et al., 2024) predominantly learn from human interaction demonstrations, directly associating actions with object regions. Despite their effectiveness on familiar objects and interactions, the limited diversity of such supervision often biases models toward object-action patterns observed during training, making generalization to novel objects challenging (Qian et al., 2024), e.g., transferring "graspability" from familiar handles to an unseen power drill. Natural language provides a more flexible interface for specifying actions and intentions (Liu et al., 2023b; Li et al., 2023b), offering semantic and functional cues beyond fixed object-action associations.

Motivated by the flexibility of language, recent works have incorporated LLMs and VLMs into affordance grounding. LLM-based methods such as WorldAfford (Chen et al., 2024) first reason about affordance concepts in language space and then localize them with CLIP/SAM. This decomposition benefits from the semantic knowledge of LLMs, but the inferred action semantics are transferred to vision only through high-level textual concepts, making the connection to local functional regions indirect. VLM-based approaches, including AffordanceLLM (Qian et al., 2024), NaturalVLM (Xu et al., 2024), and UniAFF (Yu et al., 2025), alleviate this separation by jointly modeling language and vision. However, they are commonly trained with relatively explicit object-action queries and constrained instruction formats, encouraging models to associate known actions with familiar object regions rather than reason compositionally from abstract task intent.

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.

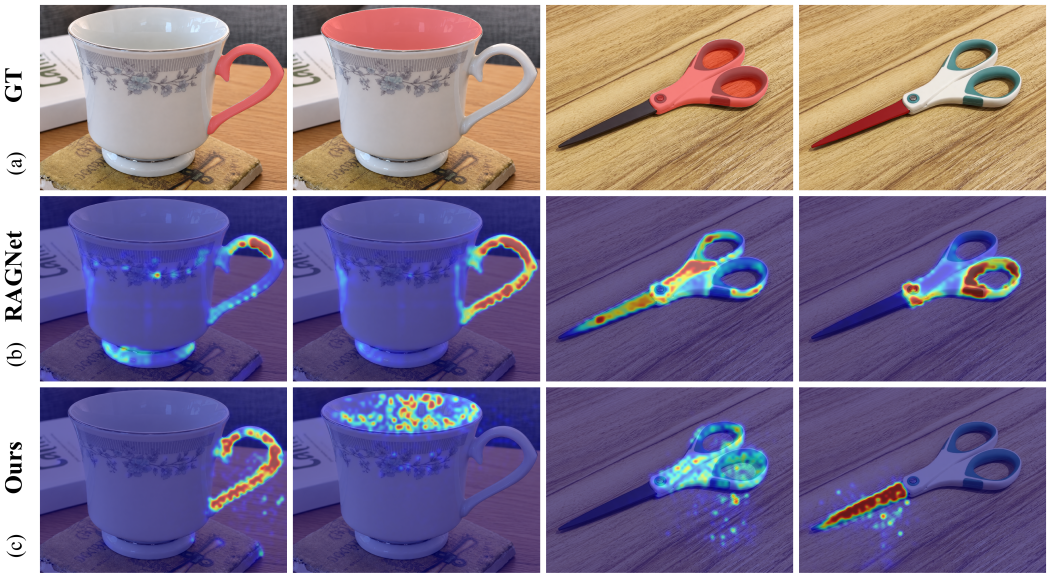


Figure 1: **Motivation of AffiMoE.** Under the same affordance instruction, RAGNet captures global task semantics but shows limited discrimination of action-specific functional regions, while AffiMoE enables finer-grained text–visual alignment for more precise local affordance grounding.

To address this limitation, RAGNet (Wu et al., 2025) advances affordance reasoning beyond explicit object-action queries through large-scale, human-like instructions. Its hard setting removes explicit object names, requiring the model to infer the relevant object and its function from high-level intent before predicting the interaction region. RAGNet further adopts the LISA (Lai et al., 2024) reasoning-segmentation paradigm, allowing multimodal reasoning to directly condition dense affordance prediction. This substantially strengthens intent-level reasoning, but does not fully resolve the localization problem. In particular, rich multimodal semantics are still summarized into a compact affordance-specific representation before being passed to the mask decoder. As a result, the decoder receives strong global task semantics, yet the correspondence between individual action cues and local functional structures remains largely implicit. This becomes especially limiting for affordance grounding, where the same object may expose different actionable parts under different actions (*e.g., the overlapping responses to grasping and cutting on scissors in Figure 1(b)*).

In this paper, we propose AffiMoE to explicitly bridge this gap between global semantic reasoning and local functional grounding. Built upon RAGNet, AffiMoE replaces the conventional SAM-based decoder (Kirillov et al., 2023) with SAM 3 (Carion et al., 2026) to obtain richer dense visual representations, and adopts LoRA (Hu et al., 2021) for efficient adaptation. Yet stronger visual features alone cannot establish which local structures correspond to a language-specified action. We therefore introduce a fine-grained text-visual alignment module that uses cross-attention to inject action semantics into dense SAM 3 features. Importantly, affordance grounding requires both object-level context and part-level functional evidence: global semantics identify the task-relevant object, whereas local cues distinguish the specific region that enables the action. To exploit this complementarity, we further introduce token-wise mixture-of-experts routing that adaptively integrates globally aligned multimodal representations and locally aligned visual features at each spatial location. This design enables AffiMoE to jointly reason about *what* object is relevant and *where* the intended action can be performed. In summary, our main contributions are as follows:

- We propose a **Fine-Grained Text-Visual Alignment** module that uses action-aware multimodal representations to retrieve corresponding functional evidence from dense visual

features. By progressively injecting local action-specific cues into multimodal reasoning, it explicitly connects global action semantics with fine-grained functional structures.

- We introduce a **Token-Wise Mixture-of-Experts** mechanism that treats global semantic and locally aligned representations as complementary experts and adaptively routes between them at each token. This selective fusion preserves reliable global semantics while activating local evidence where finer functional discrimination is required.
- We introduce **AffiMoE**, a progressive *what-how-where* paradigm for language-grounded affordance reasoning that connects high-level instruction understanding with action-specific functional grounding. We further construct **AffordBench** and validate the effectiveness of this paradigm under diverse human-like instructions.

2 Related Work

Language-Grounded Affordance Reasoning. Early affordance grounding methods learn action-relevant regions from human-object interaction demonstrations (Luo et al., 2022; Li et al., 2023a; Luo et al., 2024), but often generalize poorly beyond observed object-action patterns. Recent LLM- and VLM-based approaches introduce language supervision and pretrained semantic knowledge to improve open-world reasoning. WorldAfford (Chen et al., 2024) separates affordance reasoning from visual grounding using LLMs and CLIP/SAM, while AffordanceLLM (Qian et al., 2024), NaturalVLM (Xu et al., 2024), and UniAFF (Yu et al., 2025) more directly integrate vision-language representations into affordance prediction. RAGNet (Wu et al., 2025) further introduces large-scale, human-like instructions and LISA-style (Lai et al., 2024) reasoning segmentation to infer affordances from high-level task intent. However, these methods mainly improve *what* should be grounded, while fine-grained action-to-region correspondence remains less explored.

Text-Visual Alignment. Text-visual alignment has evolved from global semantic matching toward increasingly localized cross-modal interaction. CLIP (Radford et al., 2021) establishes image-text correspondence at the global level, while RegionCLIP (Zhong et al., 2022), DenseCLIP (Rao et al., 2022), and GroupViT (Xu et al., 2022) extend this paradigm to region-level and dense visual representations for open-vocabulary recognition and dense prediction. More recent approaches such as Grounding DINO (Liu et al., 2024), SEEM (Zou et al., 2023), and LISA (Lai et al., 2024) further couple language understanding with object localization and reasoning-driven segmentation. However, they mainly align language with objects or semantic regions, leaving action-dependent correspondence to fine-grained functional structures underexplored.

3 Method

3.1 Language-Grounded Affordance Reasoning

Object-Action Affordance Reasoning. Given an image $I \in \mathbb{R}^{H\times W\times 3}$ and a natural-language instruction T, language-grounded affordance reasoning aims to predict an action-conditioned affordance map $\hat{M} = \mathcal{F}_{\mathrm{AFF}}(I, T)$, where $\hat{M} \in [0,1]^{H\times W}$ indicates the relevance of each spatial location to the intended interaction. Unlike conventional segmentation, which primarily identifies *what* region corresponds to an object, affordance reasoning additionally requires understanding *how* the object should be interacted with and *where* the corresponding functional region lies (Luo et al., 2022). Thus, accurate affordance grounding inherently depends on first resolving the referred object and then reasoning about the action-specific functional region within it.

Motivated by this, we decompose the instruction as $T = (O, A)$, where O denotes the target object and A the intended action. This yields a progressive *what-how-where* reasoning process: O determines *what* to interact with, while A, conditioned on the grounded object, determines *how* and *where* the interaction should occur. As illustrated in Figure 2(a), we formulate this process as

$$M_{\mathrm{obj}} = \mathcal{F}_{\mathrm{obj}}(I, O), \qquad \hat{M} = \mathcal{F}_{\mathrm{AFF}}(I, M_{\mathrm{obj}}, A), \tag{1}$$

where M_{obj} provides the object-level spatial prior for subsequent action reasoning. This formulation explicitly constrains affordance prediction to the referred object while associating the intended action with its corresponding functional region.

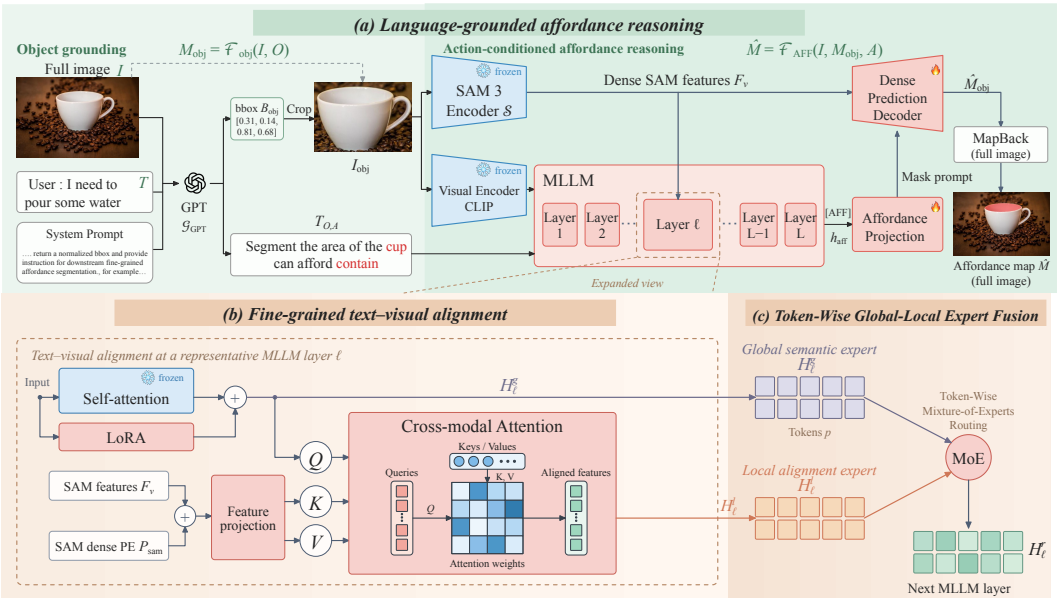


Figure 2: **Overview of our AffiMoE.** (a) AffiMoE follows an object-grounding-to-affordance-reasoning pipeline. Given an image I and instruction T, it first localizes the target object O to obtain M_{obj}, and then predicts the action-specific functional region through $\mathcal{F}_{\text{AFF}}(I, M_{\text{obj}}, A)$. The affordance reasoning stage contains two key components: (b) **Fine-Grained Text-Visual Alignment**, which grounds global action-aware MLLM representations into dense SAM features to capture local functional cues; and (c) **Token-Wise Global-Local Expert Fusion**, which adaptively integrates the global semantic expert and local alignment expert for subsequent affordance prediction.

AffiMoE Overview. We instantiate the above object-action formulation with AffiMoE, as illustrated in Figure 2(a). Given an image I and instruction T, we first use GPT to parse the target object O and intended action A, while localizing the target object with a bounding box B_{obj}:

$$(O, A, B_{\text{obj}}) = \mathcal{G}_{\text{GPT}}(I, T). \tag{2}$$

This step establishes the object-level spatial context for subsequent affordance reasoning. The target region is then cropped from the original image, $I_{\text{obj}} = \text{Crop}(I, B_{\text{obj}})$, and the instruction is reformulated as an object-action prompt $T_{O,A} = \phi(O, A)$. Thus, the subsequent reasoning is explicitly conditioned on both the localized object and the intended interaction.

Given I_{obj}, SAM 3 extracts dense spatial features $F_v = \mathcal{S}(I_{\text{obj}})$, while the MLLM jointly processes I_{obj} and $T_{O,A}$. Within this reasoning process, **Fine-Grained Text-Visual Alignment** associates the action semantics with local visual structures, and **Token-Wise Global-Local Expert Fusion** selectively integrates global semantic and locally aligned representations. We denote the resulting multimodal representation and affordance token as

$$(H, h_{\text{aff}}) = \mathcal{E}_{\text{AffiMoE}}(I_{\text{obj}}, T_{O,A}; F_v). \tag{3}$$

The affordance token h_{aff} captures the action-conditioned functional semantics, while F_v retains dense spatial details. They are jointly fed into the decoder to obtain an object-level affordance prediction: $\hat{M}_{\text{obj}} = \mathcal{D}(F_v, h_{\text{aff}})$. Finally, the prediction is transformed back to the coordinate system of the original image: $\hat{M} = \text{MapBack}(\hat{M}_{\text{obj}}, B_{\text{obj}})$.

3.2 Fine-Grained Text-Visual Alignment

Given the grounded object crop I_{obj} and the object-action prompt $T_{O,A}$ defined above, the remaining challenge is to associate the intended action with the corresponding functional structure within the object. Although the MLLM can capture the global object-action relationship, its multimodal representations remain relatively coarse for distinguishing fine-grained functional regions. We therefore introduce a lightweight text-visual alignment module after the self-attention block of each MLLM

layer, as illustrated in Figure 2(b). At layer ℓ, the original MLLM branch produces a global action-aware representation H_ℓ^g, which is further aligned with dense SAM features to obtain a complementary local representation H_ℓ^l. The two representations are fused and propagated to the next layer, progressively injecting action-specific local evidence into multimodal reasoning.

Global Text-Visual Alignment. The MLLM jointly encodes I_{obj} and $T_{O,A}$, establishing the global correspondence between the grounded object and the intended action through multimodal self-attention. At layer ℓ, self-attention with LoRA adaptation (Hu et al., 2021) produces H_ℓ^g, which retains the high-level object-action semantics learned by the pretrained MLLM. For example, given a mug with the action *grasp*, H_ℓ^g captures that the interaction should be performed on the mug, but may not precisely distinguish its handle from other object parts. To preserve this pretrained semantic space, we reuse the query transformation of the host self-attention layer: $Q_\ell = \text{RMSNorm}(H_\ell^g)W_Q$, where W_Q is inherited from the pretrained MLLM. In this way, Q_ℓ carries the object-conditioned action semantics that guide subsequent local retrieval.

Local Text-Visual Alignment. In parallel, SAM 3 extracts dense spatial features $F_{\text{sam}} = \mathcal{S}(I_{\text{obj}})$, which preserve fine-grained part-level structures of the same grounded object. Let P_{sam} denote the corresponding dense positional encoding. We project the spatially encoded SAM representation into keys and values as $K = (F_{\text{sam}} + P_{\text{sam}})W_K^s$ and $V = (F_{\text{sam}} + P_{\text{sam}})W_V^s$. The action-aware queries then retrieve local visual evidence that is most relevant to the specified interaction:

$$H_\ell^l = W_O^s \,\text{Norm}\left[\text{Softmax}\left(\frac{Q_\ell K^\top}{\sqrt{d_h}}\right)V\right]. \tag{4}$$

This interaction establishes a global-to-local alignment: H_ℓ^g specifies *which action-specific evidence is required*, while the dense SAM representation provides *where that evidence is spatially supported* within the grounded object. Consequently, H_ℓ^l complements the global object-action semantics with fine-grained functional cues. For example, given the same mug, a *grasp* prompt retrieves evidence around the handle, whereas a *contain* prompt emphasizes the inner cup region. The integration of H_ℓ^g and H_ℓ^l is described in the following subsection.

3.3 Token-Wise Global-Local Expert Fusion

The alignment above yields two complementary representations: the *Global Semantic Expert* H^g and the *Local Alignment Expert* H^l. The key challenge is that local alignment is not uniformly beneficial across tokens. Some tokens are already reliably represented by the pretrained MLLM, whereas others require additional local evidence. Therefore, globally mixing the two experts can unnecessarily perturb already-correct representations.

Analysis of Global-Local Token Fusion. A straightforward solution is direct addition,At layer ℓ, $H_\ell^r = H_\ell^g + H_\ell^l$, which forces both experts to contribute to every token. This can perturb representations that are already reliable in H_ℓ^g by introducing unnecessary local responses. For example, for *grasping* a mug, the global representation may already encode the overall mug semantics correctly, while only tokens corresponding to the handle require additional local refinement. A shared weighted fusion, $H_\ell^r = \alpha H_\ell^g + (1-\alpha)H_\ell^l$, partially alleviates this interference, where $\alpha \in [0,1]$ is a scalar fusion coefficient shared across all tokens. However, a single α assumes the same global-local preference everywhere. This is restrictive because the demand for local refinement is token dependent: a weight suitable for handle-related tokens may unnecessarily modify well-represented tokens, whereas a weight favoring global semantics may suppress useful local evidence. Therefore, the fusion decision should be made independently for each token.

Token-Wise Mixture-of-Experts Routing. We therefore formulate H_ℓ^g and H_ℓ^l as two experts and perform hard routing at the token level, as illustrated in Figure 2(c). For each spatial token p, we first measure the response strength of the two experts as $s_p^g = \rho(H_p^g)$ and $s_p^l = \rho(H_p^l)$, where $\rho(\cdot)$ aggregates the responses of a token into a scalar score. The routing decision is then defined as $G_p = \mathbb{I}[s_p^l > s_p^g]$, with $G_p \in \{0,1\}$ indicating the selected expert. Accordingly, the routed representation is computed as $H_p^r = (1-G_p)H_p^g + G_p H_p^l$. When $G_p = 0$, the Global Semantic Expert is preserved; otherwise, the Local Alignment Expert is activated to provide stronger fine-grained functional evidence. Unlike direct addition or globally shared weighting, this routing makes an independent expert decision for each token, preserving reliable global semantics while selectively introducing local alignment where finer functional discrimination is required.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUPS

Datasets. We construct **AffordBench**, a language-grounded affordance dataset collected from HANDAL (Guo et al., 2023), Open-X (Open X-Embodiment Collaboration, 2024), EgoObjects (Zhu et al., 2023), GraspNet (Fang et al., 2020), and IIT-AFF (Nguyen et al., 2017). It contains 5,510 object-action instances from 3,144 images across seven affordance classes, with human-verified or manually refined functional-region annotations.For the main experiments, we use only AffordBench for fine-tuning. Selected ablation studies instead use only the original RAGNet (Wu et al., 2025) training set. Evaluation is conducted on UMD (Myers et al., 2015) and UMD-relabeled, which share the same evaluation images but use different affordance annotations. HANDAL is additionally used for selected RAGNet-only ablations.

Evaluation Metrics. We evaluate our method using cumulative Intersection-over-Union (cIoU) (Wu et al., 2020) and average per-sample Intersection-over-Union (gIoU) (Liu et al., 2023a). cIoU measures the global pixel-level overlap by aggregating intersections and unions over all samples, while gIoU averages the IoU computed independently for each sample. Thus, cIoU emphasizes overall segmentation accuracy, whereas gIoU reflects balanced per-sample performance. Both metrics are reported as percentages, with higher values indicating better performance.

Implementation Details. AffiMoE is implemented based on AffordanceVLM (Wu et al., 2025), with the language backbone initialized from AffordanceVLM-7B and segmentation modules from pretrained SAM 3.1 (Carion et al., 2026). We freeze the CLIP (Radford et al., 2021) and SAM (Kirillov et al., 2023) image encoders and the CLIP-to-language projector, and adapt the language backbone using LoRA (Hu et al., 2021) with rank 8. We use AdamW (Loshchilov & Hutter, 2019) with learning rates of 10^{-5} for the backbone and 10^{-4} for the alignment modules, $(\beta_1, \beta_2) = (0.9, 0.95)$, and zero weight decay. Models are trained for 10 epochs with 100-step linear warmup and linear decay on two NVIDIA H200 GPUs using BF16 and DeepSpeed ZeRO-2 (Rajbhandari et al., 2020), with an effective batch size of 32.

Baseline and Comparisons. We use AffordanceVLM as the primary baseline and fine-tune it on AffordBench. To isolate architectural gains from additional data, we compare AffordanceVLM, its SAM 3 based variant, and our model under the same RAGNet-only training setting. All controlled comparisons use identical training and evaluation protocols. We further compare with Affordance-R1 (Wang et al., 2026), LOCATE (Li et al., 2023a), Cross-View-AG (Luo et al., 2022), and Uni-Aff (Yu et al., 2025) on both UMD annotation versions.

4.2 COMPARISON WITH STATE-OF-THE-ARTS

Quantitative Comparisons. Table 1 compares AffiMoE with prior affordance grounding methods across different instruction paradigms. While earlier approaches rely on structured action or object–affordance labels, AffordanceVLM and AffiMoE operate on more flexible human-like instructions, which require reasoning from high-level intent to the corresponding functional region. AffiMoE achieves 72.06/63.33 cIoU/gIoU on UMD and 83.04/79.76 on UMD-relabeled. Under the same human-like instruction setting, it outperforms AffordanceVLM by 45.41/34.22 and 51.39/41.85 points, respectively. It also surpasses Affordance-R1, the strongest competing method, by 21.90/8.97 points on UMD and 23.99/13.74 points on UMD-relabeled. These gains show that richer language supervision alone is insufficient: accurate affordance grounding requires explicitly connecting high-level action semantics to fine-grained local functional structures.

Qualitative Comparisons. Figure 3 provides further evidence for this global-to-local gap. Existing methods often identify the relevant object or approximate interaction region, but produce diffuse responses or confuse nearby functional parts. In contrast, AffiMoE yields compact, action-specific predictions that better match the ground truth, especially for small and elongated regions. This suggests that AffiMoE preserves global task semantics while resolving the local structures that determine *where* the intended interaction should occur. The consistent trend on both UMD and UMD-relabeled further shows that this advantage is robust to the two annotation protocols. Together, these results support our central hypothesis: global intent reasoning identifies *what* is relevant, while fine-grained text-visual alignment is critical for resolving *where* the interaction should occur.

Table 1: **Quantitative comparison of our AffiMoE with state-of-the-arts.**

Method	Instruction	UMD		UMD-relabeled	
		cIoU↑	gIoU↑	cIoU↑	gIoU↑
Cross-View-AG (Luo et al., 2022)	action label	1.61	2.63	1.90	3.00
LOCATE (Li et al., 2023a)	affordance + object label	11.59	10.42	13.33	12.61
UniAff (Yu et al., 2025)	language	22.55	38.30	25.35	44.19
AffordanceVLM (Wu et al., 2025)	human-like instruction	26.65	29.11	31.65	37.91
Affordance-R1 (Wang et al., 2026)	human-like instruction	50.16	54.36	59.05	66.02
AffiMoE(ours)	human-like instruction	**72.06**	**63.33**	**83.04**	**79.76**

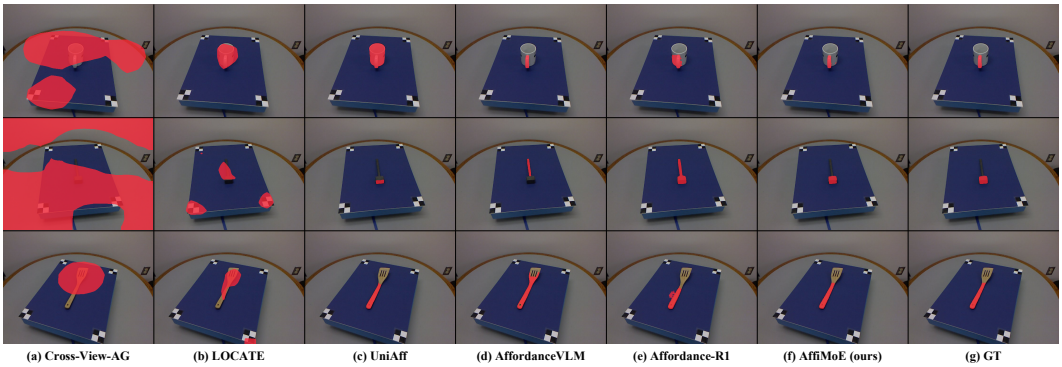


Figure 3: **Qualitative comparison across affordance grounding paradigms.** From conventional supervised methods in (a)-(b), to the LLM-based approach in (c), and VLM-based methods in (d)-(f), affordance grounding progressively benefits from richer semantic reasoning. AffiMoE yields more precise action-specific functional regions that better match the ground truth.

Table 2: **Quantitative evaluation against baseline.** (a) denotes the AffordanceVLM-based baseline with SAM replaced by SAM 3, using a frozen MLLM backbone without LoRA adaptation. '**FGA**' introduces fine-grained text-visual alignment with direct additive fusion of the global and local representations. '**MoE**' further replaces direct addition with token-wise mixture-of-experts routing, adaptively selecting the global or local expert for each token.

Exp.	AffiMoE			Object-Action Alignment		UMD		UMD-relabeled	
	LoRA	FGA	MoE	Global	Local	cIoU↑	gIoU↑	cIoU↑	gIoU↑
(a)	✗	✗	✗	N	N	65.72	59.03	73.82	72.40
(b)	✓	✗	✗	Y	N	68.71	60.12	76.42	72.99
(c)	✗	✓	✗	N	Y	70.86	63.75	82.71	80.79
(d)	✗	✓	✓	N	Y	71.23	63.79	82.85	80.51
(e)	✓	✓	✗	Y	Y	69.83	63.86	81.56	80.61
(f)	✓	✓	✓	Y	Y	72.06	63.33	83.04	79.76

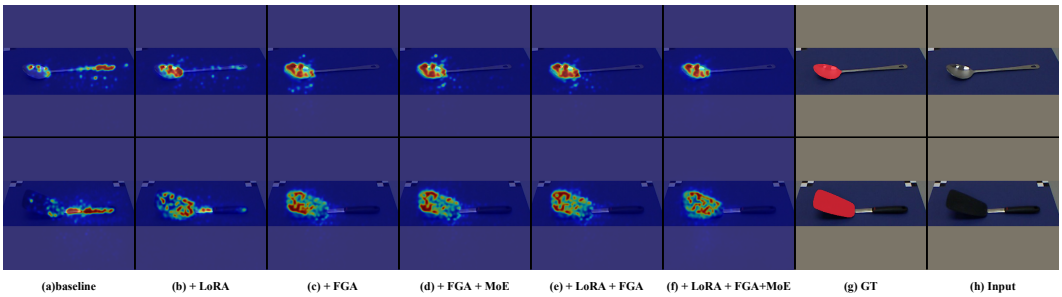


Figure 4: **Attention-map analysis of baseline variants.** 'FGA' sharpens responses toward action-specific functional regions, while token-wise 'MoE' further suppresses irrelevant activations, providing visual evidence for the effectiveness of fine-grained alignment and adaptive global-local fusion.

4.3 Analysis of Baseline Improvements

Table 2 and Figure 4 analyze the contributions of LoRA adaptation, fine-grained text–visual alignment (FGA), and token-wise global–local expert fusion. Adding LoRA from (a) to (b) yields modest gains of 2.99/1.09 cIoU/gIoU on UMD and 2.60/0.59 on UMD-relabeled, while both variants still exhibit diffuse responses over multiple object parts, indicating that adapting global semantic representations alone is insufficient for precise functional-region discrimination. In contrast, FGA provides the largest and most consistent improvement: without LoRA, (c) improves over (a) by 5.14/4.72 points on UMD and 8.89/8.39 on UMD-relabeled, while with LoRA, (e) improves over (b) by 1.12/3.74 and 5.14/7.62 points, respectively. Consistently, FGA shifts responses from diffuse object-level activation toward the ground-truth functional regions, supporting the need to explicitly align global action semantics with dense local structures. Replacing direct addition with token-wise routing further improves cIoU, with (d) exceeding (c) by 0.37/0.14 points and (f) exceeding (e) by 2.23/1.48 points on UMD/UMD-relabeled. Although gIoU does not consistently increase, the routed variants produce more concentrated responses with less activation in irrelevant regions, suggesting that global and local representations should be integrated selectively rather than uniformly. Overall, FGA is the key component for bridging global action semantics and local functional structures, while token-wise expert routing further improves spatial selectivity, supporting the central design of AffiMoE: global reasoning determines *what* is relevant, local alignment resolves *where* the interaction occurs, and adaptive routing balances the two.

Table 3: **Effect of SAM-to-SAM 3 visual encoder replacement.** Quantitative comparison under the baseline and AffiMoE settings on UMD and UMD-relabeled.

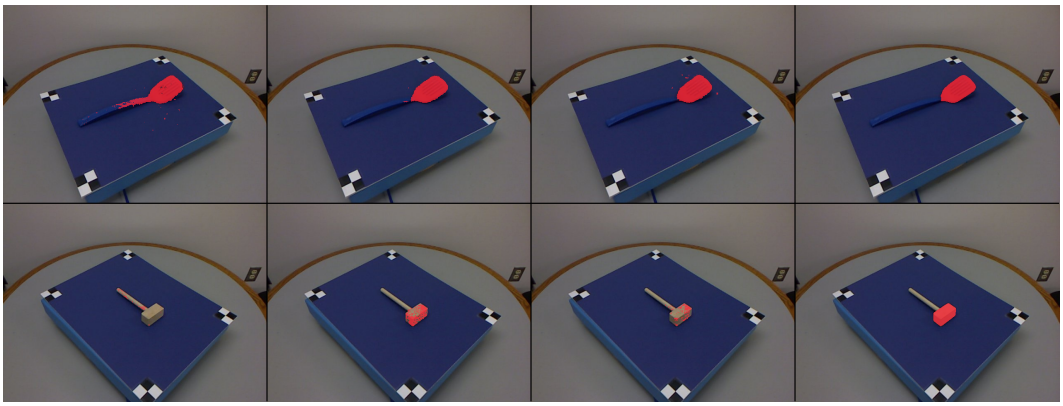
	UMD		UMD-relabeled	
Method	cIoU ↑	gIoU ↑	cIoU ↑	gIoU ↑
Baseline w/ SAM	26.65	29.11	31.65	37.91
Baseline w/ SAM 3	25.09	28.32	29.45	36.45
AffiMoE w/ SAM	27.89	27.72	31.42	33.06
AffiMoE w/ SAM 3	**31.61**	**35.45**	**38.09**	**44.20**



Figure 5: **Visual comparison of SAM and SAM 3 encoders.** SAM 3 provides stronger dense visual representations, enabling AffiMoE to better recover fine-grained functional regions.

4.4 Ablation Studies

Effect of SAM-to-SAM 3 Visual Encoder Replacement. Table 3 and Figure 5 compare SAM and SAM 3 under the same RAGNet training data. The first two rows of Table 3 show that replacing SAM with SAM 3 alone does not improve the baseline, indicating that stronger visual features require effective affordance alignment. In contrast, comparing the third and fourth rows, AffiMoE benefits substantially from SAM 3, with cIoU/gIoU improving from 27.89/27.72 to 31.61/35.45 on UMD and from 31.42/33.06 to 38.09/44.20 on UMD-relabeled. Figure 5 shows a consistent pattern: columns (a)-(b) focus on action-irrelevant regions, column (c) identifies the correct cup opening but with coarse boundaries, while column (d) yields a cleaner prediction closer to the GT in (e). These results show that stronger dense visual features are most effective when coupled with explicit action-conditioned alignment, demonstrating the complementarity between SAM 3 and AffiMoE.



(a) Baseline | (b) AffiMoE w/ Token-wise MoE | (c) AffiMoE w/ Shared Weight | (d) GT

Figure 6: **Qualitative comparison of token-wise MoE and shared-weight fusion.** Token-wise MoE yields more focused action-specific responses and suppresses irrelevant activations, producing functional-region predictions closer to the GT than shared-weight fusion.

Effect of Token-wise MoE. Table 4 compares token-wise MoE with shared-weight fusion under the same cross-modal alignment setting. In the shared-weight variant, all locally aligned features share a single learned weight within each layer, i.e., $(1-a)H_\ell^g + aH_\ell^l$, whereas token-wise MoE independently selects between the global and local experts for each token. Token-wise MoE achieves 83.04 cIoU and 79.76 gIoU, outperforming shared-weight fusion by 1.20 cIoU while maintaining comparable gIoU (79.76 vs. 79.66). Figure 6 (b) and (c) show that both fusion strategies improve over the baseline, while token-wise MoE produces more complete functional regions and fewer irrelevant responses, particularly in the *pound* example. These results suggest that global-local preference varies across tokens and cannot be fully captured by a single shared weight. Token-wise expert selection therefore preserves global semantics while adding local evidence where finer discrimination is needed.

Table 4: Ablation of Token-wise MoE vs. shared-weight fusion on UMD-relabeled.

Method	cIoU ↑	gIoU ↑
Baseline	73.82	72.40
AffiMoE w/ Token-wise MoE	**83.04**	**79.76**
AffiMoE w/ Shared Weight	81.84	79.66

Effect of RoPE on Queries in Fine-Grained Alignment. Table 5 studies whether RoPE (Su et al., 2024) should be further applied to the MLLM-derived queries in cross-modal alignment. In our design, the query Q is derived from the MLLM representation, while the keys and values (K, V) are projected from SAM 3 features that already encode spatial position. Applying RoPE only to Q introduces language-sequence positional information into the cross-modal similarity, $s_{ij}^{\text{Q-RoPE}} = (R_i q_i)^\top k_j / \sqrt{d_h}$, but no corresponding rotation exists for the SAM 3 keys. As a result, the sequence coordinate system of the MLLM is not aligned with the spatial coordinate system of SAM 3, and the additional query-side rotation can distort the semantic matching between action-aware queries and local visual features. Consistently, applying RoPE to Q reduces cIoU/gIoU from 83.04/79.76 to 81.71/79.04. Therefore, we omit RoPE on Q in the fine-grained alignment.

Table 5: Ablation of query-side RoPE in the cross-modal alignment of AffiMoE on UMD-relabeled.

Method	cIoU ↑	gIoU ↑
AffiMoE	83.04	79.76
AffiMoE w/ Query-side RoPE	81.71	79.04

5 Conclusion

In this work, we presented AffiMoE for language-grounded affordance reasoning with human-like instructions. AffiMoE follows a progressive *what-how-where* paradigm that connects high-level task understanding with precise functional-region grounding. To bridge global semantics and local structures, we introduced fine-grained text-visual alignment and token-wise mixture-of-experts routing for adaptive global-local integration. We further constructed AffordBench for fine-grained affordance evaluation. Extensive experiments and ablations demonstrate the effectiveness of the proposed design, showing that AffiMoE more accurately translates high-level interaction intent into action-specific functional regions. Thus, AffiMoE provides a promising basis for embodied AI.

AI USE STATEMENT

AI models assisted with dataset pre-annotation by proposing object names, bounding boxes, affordance labels, and candidate segmentation masks. The retained annotations were verified or refined by human annotators, as described in Appendix B. The object-grounding pipeline described in Section 3.1 also uses GPT to interpret instructions, identify the target object and intended action, and estimate the object bounding box. Additionally, we used Codex for grammar refinement, formatting checks, and methodological feedback. The authors are responsible for verifying AI-assisted outputs.

REPRODUCIBILITY STATEMENT

To support reproducibility, we describe the object-grounding pipeline, fine-grained text-visual alignment, and token-wise expert fusion in Sections 3.1–3.3. Section 4.1 reports the pretrained model initialization, frozen components, LoRA configuration, optimization settings, training schedule, hardware, and evaluation metrics. Appendix B.1 documents the source datasets, annotation procedure, filtering and human review process, and statistics of AffordBench. Appendix B.2 describes the instruction templates and training-time image and mask preprocessing. Appendix B.3 explains the construction of UMD-relabeled and its relationship to the original UMD annotations. Component analyses and ablation results are presented in Sections 4.3–4.4, with additional HANDAL results in Appendix A. We will release our source code, model weights, training and evaluation configurations, the complete AffordBench training set, and the UMD-relabeled evaluation set upon publication.

REFERENCES

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations (ICLR)*, 2026.

Changmao Chen, Yuren Cong, and Zhen Kan. WorldAfford: Affordance grounding based on natural language instructions. In *IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 822–828. IEEE, 2024.

Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. GraspNet-1Billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11444–11453. IEEE, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Fang_GraspNet-1Billion_A_Large-Scale_Benchmark_for_General_Object_Grasping_CVPR_2020_paper.html.

Andrew Guo, Bowen Wen, Jianhe Yuan, Jonathan Tremblay, Stephen Tyree, Jeffrey Smith, and Stan Birchfield. HANDAL: A dataset of real-world manipulable object categories with pose annotations, affordances, and reconstructions. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023. URL https://nvlabs.github.io/HANDAL/.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685, 2021. URL https://arxiv.org/abs/2106.09685.

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 (ICCV)*, pp. 4015–4026. IEEE, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Kirillov_Segment_Anything_ICCV_2023_paper.html.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9579–9589. IEEE, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Lai_LISA_

Reasoning_Segmentation_via_Large_Language_Model_CVPR_2024_paper.html.

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 (CVPR)*, pp. 10922–10931. IEEE, 2023a. URL https://openaccess.thecvf.com/content/CVPR2023/html/Li_LOCATE_Localize_and_Transfer_Object_Parts_for_Weakly_Supervised_Affordance_CVPR_2023_paper.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023b. URL https://proceedings.mlr.press/v202/li23q.html.

Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23592–23601. IEEE, 2023a. URL https://openaccess.thecvf.com/content/CVPR2023/html/Liu_GRES_Generalized_Referring_Expression_Segmentation_CVPR_2023_paper.html.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023b. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision (ECCV)*, pp. 38–55. Springer, 2024. doi: 10.1007/978-3-031-72970-6_3.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

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 (CVPR)*, pp. 2252–2261. IEEE, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Luo_Learning_Affordance_Grounding_From_Exocentric_Images_CVPR_2022_paper.html.

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.

Austin Myers, Ching L. Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1374–1381. IEEE, 2015. doi: 10.1109/ICRA.2015.7139369. URL https://users.umiacs.umd.edu/~fermulcm/affordance/Affordance.html.

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 *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5908–5915. IEEE, 2017. doi: 10.1109/IROS.2017.8206484. URL https://sites.google.com/site/ocnncrf/home.

Open X-Embodiment Collaboration. Open X-Embodiment: Robotic learning datasets and RT-X models. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024. doi: 10.1109/ICRA57147.2024.10611477. URL https://robotics-transformer-x.github.io/.

Shengyi Qian, Weifeng Chen, Min Bai, Xiong Zhou, Zhuowen Tu, and Li Erran Li. AffordanceLLM: Grounding affordance from vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 7587–7597. IEEE, 2024. URL https://openaccess.thecvf.com/content/CVPR2024W/OpenSUN3D/html/Qian_AffordanceLLM_Grounding_Affordance_from_Vision_Language_Models_CVPRW_2024_paper.html.

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 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models. In *International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, pp. 1–16. IEEE, 2020. doi: 10.1109/SC41405.2020.00024. URL https://doi.org/10.1109/SC41405.2020.00024.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. DenseCLIP: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18082–18091. IEEE, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rao_DenseCLIP_Language-Guided_Dense_Prediction_With_Context-Aware_Prompting_CVPR_2022_paper.html.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Hanqing Wang, Shaoyang Wang, Yiming Zhong, Zemin Yang, Jiamin Wang, Zhiqing Cui, Jiahao Yuan, Yifan Han, Mingyu Liu, and Yuexin Ma. Affordance-R1: Reinforcement learning for generalizable affordance reasoning in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 9738–9746, 2026. doi: 10.1609/aaai.v40i12.37937.

Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. PhraseCut: Language-based image segmentation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10216–10225. IEEE, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wu_PhraseCut_Language-Based_Image_Segmentation_in_the_Wild_CVPR_2020_paper.html.

Dongming Wu, Yanping Fu, Saike Huang, Yingfei Liu, Fan Jia, Nian Liu, Feng Dai, Tiancai Wang, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Jianbing Shen. RAGNet: Large-scale reasoning-based affordance segmentation benchmark towards general grasping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11980–11990. IEEE, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wu_RAGNet_Large-scale_Reasoning-based_Affordance_Segmentation_Benchmark_towards_General_Grasping_ICCV_2025_paper.html.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. GroupViT: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18134–18144. IEEE, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Xu_GroupViT_Semantic_Segmentation_Emerges_From_Text_Supervision_CVPR_2022_paper.html.

Ran Xu, Yan Shen, Xiaoqi Li, Ruihai Wu, and Hao Dong. NaturalVLM: Leveraging fine-grained natural language for affordance-guided visual manipulation. *IEEE Robotics and Automation Letters*, 9(12):10842–10849, 2024. doi: 10.1109/LRA.2024.3477095.

Qiaojun Yu, Siyuan Huang, Xibin Yuan, Zhengkai Jiang, Ce Hao, Xin Li, Haonan Chang, Junbo Wang, Liu Liu, Hongsheng Li, Peng Gao, and Cewu Lu. UniAff: A unified representation of affordances for tool usage and articulation with vision-language models. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8980–8987. IEEE, 2025.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. RegionCLIP: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16793–16803. IEEE, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Zhong_RegionCLIP_Region-Based_Language-Image_Pretraining_CVPR_2022_paper.html.

Dingyi Zhou, Mu He, Zhuowei Fang, Xiangtong Yao, Yinlong Liu, Alois Knoll, and Hu Cao. AffordanceGrasp-R1: Leveraging reasoning-based affordance segmentation with reinforcement learning for robotic grasping. arXiv preprint arXiv:2602.03547, 2026. URL https://arxiv.org/abs/2602.03547.

Chenchen Zhu, Fanyi Xiao, Andres Alvarado, Yasmine Babaei, Jiabo Hu, Hichem El-Mohri, Sean Culatana, Roshan Sumbaly, and Zhicheng Yan. EgoObjects: A large-scale egocentric dataset for fine-grained object understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20110–20120. IEEE, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhu_EgoObjects_A_Large-Scale_Egocentric_Dataset_for_Fine-Grained_Object_Understanding_ICCV_2023_paper.html.

Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 19769–19782. Curran Associates, Inc., 2023. doi: 10.52202/075280-0868. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3ef61f7e4afacf9a2c5b71c726172b86-Paper-Conference.pdf.

A Additional Experiments

State-of-the-art performance on HANDAL. We further evaluate AffiMoE on HANDAL using exactly the same training data as AffordanceVLM, without introducing any additional training data. We include AffordanceGrasp-R1 (Zhou et al., 2026) as a key baseline because it represents the existing state of the art on HANDAL. As shown in Table 6, AffiMoE achieves 71.32% cIoU and 69.33% gIoU, surpassing AffordanceGrasp-R1 by 5.32 and 3.73 percentage points, respectively, and establishing a new state of the art on this benchmark. Compared with AffordanceVLM, AffiMoE improves cIoU and gIoU by 10.52 and 9.03 percentage points, respectively. These gains, achieved using the same training data as AffordanceVLM, highlight the effectiveness of our architectural improvements.

Table 6: Comparison with state-of-the-art methods on HANDAL. Results are reported in cIoU and gIoU (%; higher is better).

	HANDAL	
Method	cIoU ↑	gIoU ↑
AffordanceVLM	60.80	60.30
AffordanceGrasp-R1	66.00	65.60
AffiMoE	**71.32**	**69.33**

B Data

We construct AffordBench to provide supervision for fine-grained functional regions and manually reannotate UMD for evaluation. The two resources serve complementary purposes: AffordBench

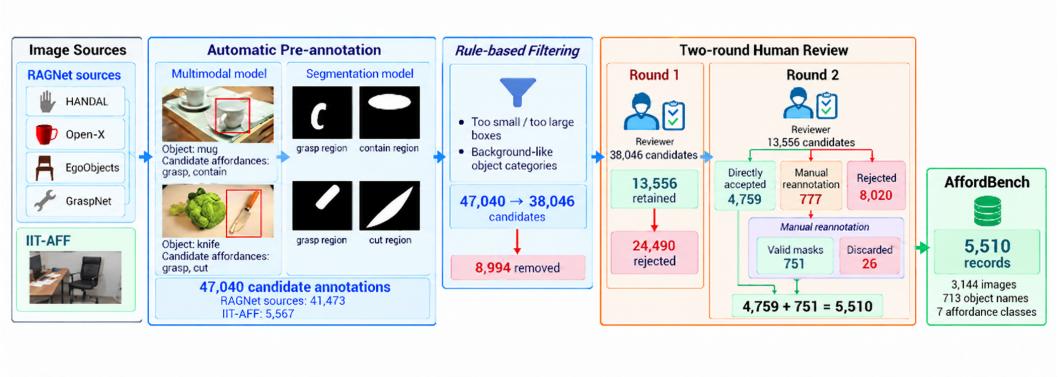


Figure 7: Construction pipeline of AffordBench. Images from the RAGNet sources and IIT-AFF undergo automatic pre-annotation, rule-based filtering, and two rounds of human review. Of 47,040 initial candidate annotations, 5,510 records are retained, including 751 masks obtained through manual reannotation. Candidate and record counts refer to object-instance–affordance annotations rather than unique images.

supplies training annotations from multiple image sources, while UMD-relabeled provides revised affordance masks for the UMD images used in our evaluation. All statistics below refer to the AffordBench training set used in our experiments.

B.1 AffordBench Training Dataset

Data collection and annotation. Figure 7 summarizes the construction of AffordBench, from automatic pre-annotation to filtering and human review. Each record associates an image with an object instance, an affordance label, an object bounding box, and a binary mask of the corresponding functional region. An image may contribute multiple records.

We draw source images from the HANDAL, Open-X, EgoObjects, and GraspNet data assembled in RAGNet, together with IIT-AFF. The automatic pre-annotation pipeline first uses a multimodal model to propose object names, bounding boxes, and candidate affordances, then uses a segmentation model to produce candidate masks. This process yields 47,040 candidate annotations: 41,473 from the RAGNet sources and 5,567 from IIT-AFF. Rule-based filtering removes 8,994 unsuitable candidates, including instances with excessively small or large bounding boxes and object categories predominantly associated with background regions. The remaining 38,046 candidates enter human review.

Human review proceeds in two rounds. The first round retains 13,556 candidates for further inspection and excludes the remaining 24,490. The second round reviews all 13,556 retained candidates: 4,759 are directly accepted, 777 are sent for manual reannotation, and 8,020 are rejected. Of the 777 reannotation tasks, 751 yield valid masks and 26 are discarded. The final training set therefore contains 5,510 records, comprising 4,759 human-verified pre-annotations and 751 manually reannotated masks.

Dataset statistics. AffordBench contains 5,510 records from 3,144 unique images, covering 713 distinct object names and 959 object–affordance combinations across seven affordance classes (Table 7). Figure 8 shows the distribution of the final records by image source: HANDAL contributes 1,524 records (27.66%), Open-X 1,492 (27.08%), EgoObjects 1,242 (22.54%), GraspNet 364 (6.61%), and IIT-AFF 888 (16.12%). These counts describe the final training set after filtering and human review.

Each image has an average of 1.75 records, and 1,302 images contain annotations for more than one affordance. The annotated functional regions occupy a median of 0.81% of the full image area and 19.7% of the object bounding-box area, highlighting the need to localize small functional regions within objects. All masks are non-empty binary maps with the same spatial dimensions as their corresponding source images.

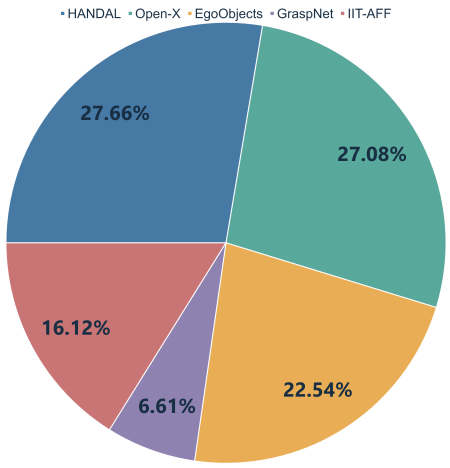


Figure 8: Distribution of the 5,510 AffordBench training records by image source after filtering and human review. Percentages are computed over records rather than unique images and may not sum to 100% due to rounding.

Table 7: Affordance distribution in the AffordBench training set. Counts denote object-instance–affordance records rather than unique images. Percentages may not sum to 100 due to rounding.

Affordance	Records	Percentage (%)
Grasp	2,797	50.76
Contain	1,405	25.50
Wrap-grasp	474	8.60
Support	354	6.42
Cut	345	6.26
Pound	86	1.56
Hit	49	0.89
Total	5,510	100.00

B.2 Instruction Construction and Training Preprocessing

Each training record is paired with four question templates that explicitly specify the object and its affordance, such as "Please segment the region of the {object} that affords {affordance}." One template is randomly selected whenever a record is loaded. The target response is "Sure, the requested affordance region is `[AFF]`." The component ablations use these standard templates; a separate GPT-rewritten instruction version of the same records is not used in these experiments.

We crop the original image and mask online using the object bounding box. Expansion ratios r_w and r_h are sampled independently from $\mathcal{U}(0.2, 0.4)$, giving target crop dimensions of $(1 + r_w)w$ and $(1 + r_h)h$ for a box of width w and height h, with a minimum crop side length of 128 pixels. The cropped image is padded to a square using RGB $(123, 117, 104)$, and the mask undergoes the corresponding spatial transformation.

B.3 UMD-Relabeled Evaluation Set

We observe that the original UMD annotations are often coarse in their delineation of functional regions, motivating manual reannotation for fine-grained affordance evaluation. We refer to the resulting annotation version as *UMD-relabeled*. It retains the underlying UMD evaluation images and replaces their affordance masks with manually revised annotations. The reannotation aims to improve the spatial precision of the evaluation targets while preserving the image content used for comparison. Figure 9 shows representative examples of common discrepancies between the original and revised masks.

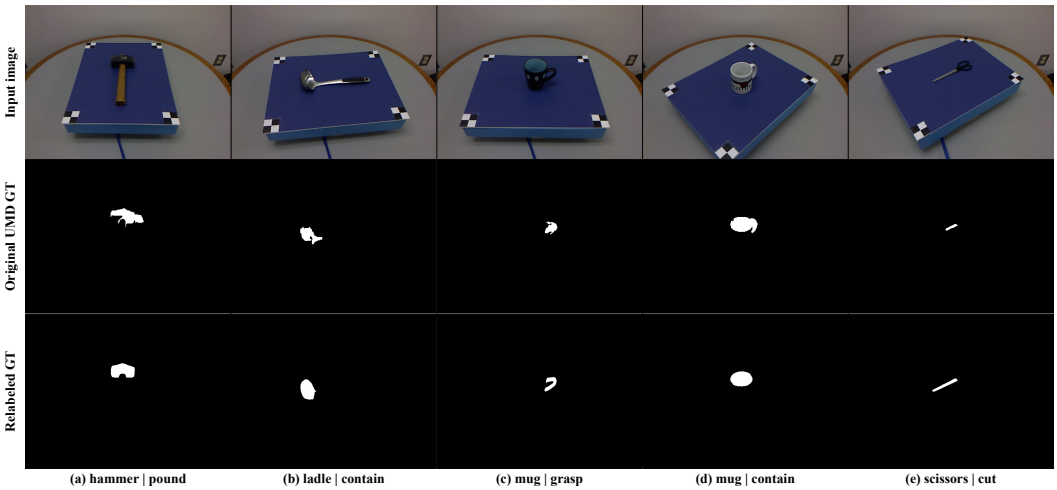


Figure 9: Qualitative comparison between the original UMD affordance masks and our UMD-relabeled annotations on the same images. The relabeled masks tighten boundaries around the functional region and correct missing or over-extended areas, enabling more reliable fine-grained evaluation.

We report results using both the original UMD annotations and UMD-relabeled. These results correspond to two annotation versions of the same image set.