# WHAT TO ATTEND, WHERE TO ACT: REFERENT–AFFORDANCE CO-EVOLUTION FOR EMBODIED SPATIAL GROUNDING

**Anonymous authors**
Paper under double-blind review

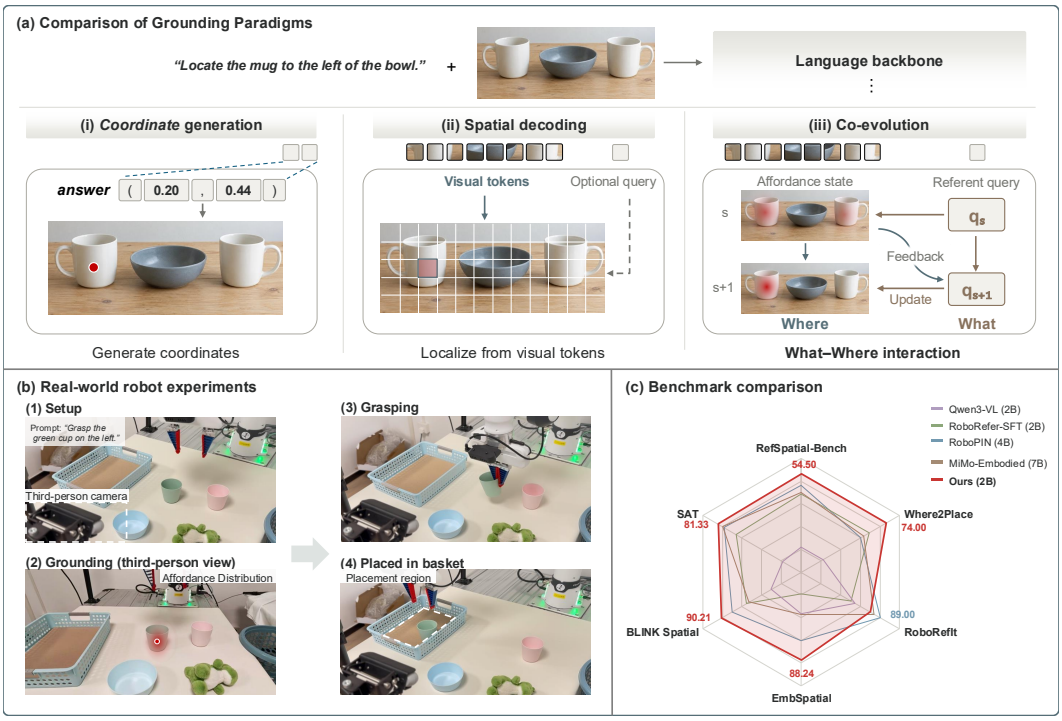



Figure 1: (a) **Comparison of spatial grounding paradigms** including (i) explicit coordinate generation, (ii) visual token-based spatial decoding, and (iii) our proposed iterative *What-Where* co-evolution framework. (b) **Real-world robotic execution** pipeline featuring language-guided grounding, grasping, and placement. (c) **Quantitative comparison** of our 2B model against baseline models across spatial grounding and reasoning benchmarks.

## ABSTRACT

Embodied spatial grounding refers to interpreting a language instruction and identifying the visual region for intended interaction. Existing methods broadly follow two paradigms: *coordinate-generation* methods produces discrete coordinate tokens as answers, whereas *spatial-decoding* methods predicts directly from visual representations to preserve spatial structure. However, both typically ground semantics into spatial predictions in a one-way manner, limiting reciprocal interaction between referent semantics and visual affordance evidence. To this end, we propose **RACE**, a **R**eferent–**A**ffordance **C**o-**E**volution framework that couples *referent semantics* specifying what to attend to with *an affordance state* indicating where to act, jointly refining referent understanding and spatial localization. Specifically, referent semantics guide updates to the affordance state, which in turn feeds back visual evidence to refine referent understanding. To supervise guide this co-evolution, we unify heterogeneous grounding annotations as target distributions and combine distribution matching with a geometry-aware con-



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.

straint penalizing spatially inconsistent predictions. Our model achieves 54.50% on RefSpatial-Bench and 74.00% on Where2Place while maintaining competitive general spatial-reasoning performance. These results suggest that referent—affordance co-evolution provides an effective interface between task-relevant semantics and fine-grained spatial grounding, offering a general pathway from task-level reasoning to actionable spatial representations.

## 1 INTRODUCTION

Embodied spatial grounding refers to the process of interpreting a natural-language instruction and localizing the region where the intended interaction should occur (Yuan et al., 2025; Zhou et al., 2025). Tasks such as `grasping a mug by its handle`, selecting an object specified by a spatial relation, or placing an item between two obstacles require more than recognizing an object in the scene (Li et al., 2023). The agent must determine *what to attend to* by interpreting the objects and spatial relations specified by the instruction, and *where to act* by identifying the region that satisfies these constraints (Huang et al., 2023), with the central challenge being to bridge high-level referent semantics and fine-grained spatial localization.

As illustrated in Figure 1(a)(i), coordinate-generation methods formulate spatial grounding as autoregressive coordinate-token prediction, leveraging the semantic understanding and reasoning capabilities of language models (Peng et al., 2024; Yuan et al., 2026b). However, this formulation compresses visual spatial judgments into a few output tokens and places both high-level referent understanding and fine-grained localization on the same language-model backbone (Zhang et al., 2024; You et al., 2024). Alternatively, spatial-decoding methods (Figure 1(a)(ii)) use language-model representations to guide the prediction of heatmaps or interaction locations from visual embeddings, preserving explicit spatial structure (Li et al., 2025; Shao et al., 2025). Yet conditioning spatial readout on a fixed semantic representation can limit the use of candidate-region evidence to refine referent understanding, making it difficult to identify the intended target among visually similar candidates under complex task constraints (Zhang et al., 2026; Dai et al., 2025).

To this end, we introduce **RACE** (Figure 1(a)(iii)), a *Referent–Affordance Co-Evolution* framework that jointly refines referent semantics specifying *what to attend to* and an affordance state expressed as a task-conditioned spatial distribution indicating *where to act*. The two co-evolve through alternating updates: referent semantics guide the accumulation of spatial evidence to update the affordance state, which in turn weights visual evidence from candidate regions to refine referent understanding for the next stage. To provide distribution-level supervision for this co-evolution process, we map point, box, and mask annotations into a shared distributional space and match predictions with the resulting target distributions at each stage. A complementary geometry-aware false-peak objective penalizes dominant responses in the final prediction based on their distance from valid target regions.

We evaluate RACE on benchmarks spanning embodied spatial grounding and general spatial understanding. Our 2B model achieves 54.50% accuracy on RefSpatial-Bench and 74.00% on Where2Place, attaining the best results among models of comparable scale, while remaining competitive in general spatial understanding with scores of 81.33% on SAT and 90.21% on BLINK Spatial. Controlled ablations support the contributions of referent-semantic refinement and spatially selective visual feedback to grounding accuracy, while real-world robot experiments demonstrate RACE's ability to support instruction-guided grasping. Our main contributions as follows:

- We introduce *RACE*, a formulation of embodied spatial grounding that bridges *what to attend to* and *where to act* through the co-evolution of referent semantics and an affordance state.
- We realize this co-evolution through reciprocal interaction: referent semantics guide affordance-state updates, while the evolving state feeds visual evidence back to refine referent understanding.
- We demonstrate strong grounding performance alongside competitive general spatial understanding, while enabling instruction-guided robotic grasping across varied real-world conditions.

## 2 RELATED WORK

**Grounding via Coordinate Generation.** This line of work formulates language-conditioned grounding as autoregressive coordinate generation. Qwen-VL and Kosmos-2 encode bounding

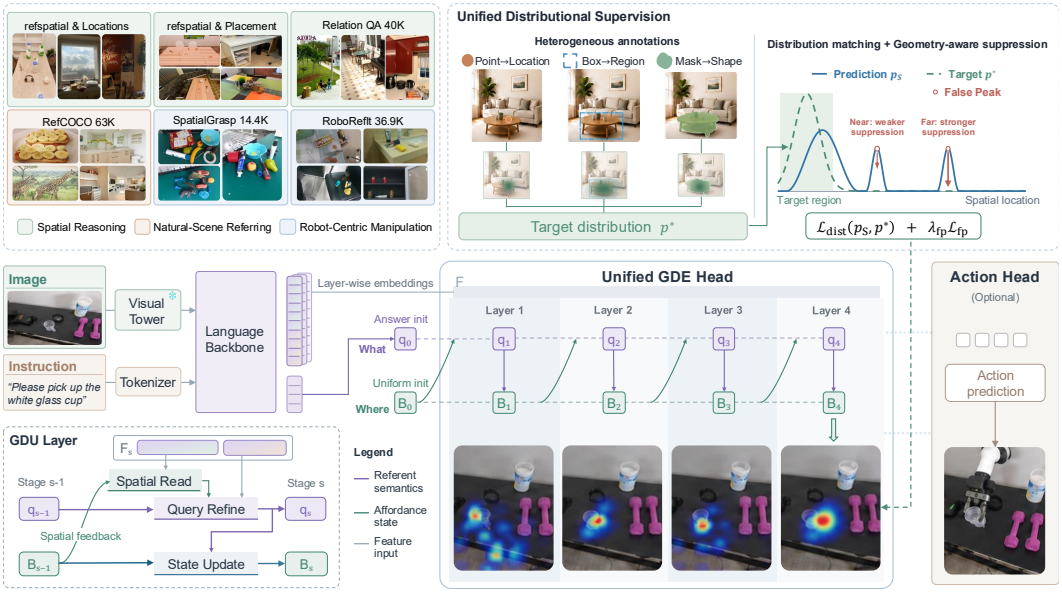



Figure 2: Overview of the proposed framework. Training data span spatial perception and reasoning, natural-scene referring, and robot-centric manipulation. Heterogeneous point, bounding-box, and segmentation-mask annotations are projected into a shared spatial distribution space, enabling unified supervision through distribution matching and geometry-aware false-peak suppression. Conditioned on layer-wise embeddings from the multimodal backbone, the GDE Head jointly evolves referent semantics and affordance states across successive GDU layers: spatial feedback guides semantic refinement, while refined semantics update spatial predictions. The resulting representations can further condition an optional action head for downstream manipulation.

boxes as token sequences (Bai et al., 2023; Peng et al., 2024), while Ferret enriches region inputs with continuous visual features but retains coordinate-based grounding outputs (You et al., 2024). Molmo and RoboPoint predict point coordinates for visual pointing and robotic affordance prediction, respectively (Deitke et al., 2025; Yuan et al., 2025). RoboRefer and Embodied-R1 further integrate spatial reasoning with reinforcement fine-tuning (Zhou et al., 2025; Yuan et al., 2026b). In contrast, we retain the multimodal language backbone for spatial reasoning while performing semantic refinement and spatial decoding with a grounding-specific module.

**Grounding based on Spatial Decoding.** Spatial decoding grounds language instructions through dedicated modules operating on visual features, with related designs used in reasoning segmentation (Lai et al., 2024) and GUI interaction (Wu et al., 2025). For visual grounding, semantic representations from a multimodal language model guide a separate decoder to select target regions from detector proposals (Kang et al., 2026). In robotics, grounding masks identify target objects and placement regions (Huang et al., 2025), while heatmap prediction aligns action locations with visual inputs (Li et al., 2025). Adaptive affordance heatmaps further represent spatial ambiguity in language-guided robotic tasks (Shao et al., 2025). Recent work also explores iterative perception–cognition interaction (Dai et al., 2025) and bidirectional updates between language conditions and visual decoder states (Zhang et al., 2026). Our method builds on this spatial-decoding paradigm by coupling referent-semantic refinement with progressive affordance-distribution evolution.

# 3 Method

## 3.1 Overall Framework

RACE represents referent semantics as a query $q_s$ and the affordance state as logits $B_s$, with $p_s = \mathrm{softmax}(B_s)$ denoting the corresponding affordance distribution (Figure 2). Given the initial query

and multi-depth memories from the multimodal backbone, the Grounding Distribution Evolution Head (GDE Head) jointly evolves the referent query and affordance state for localization.

Given an image $I$ and a language instruction $u$, the multimodal backbone $\mathcal{M}$ autoregressively generates an answer and its hidden representations. The generated answer establishes task-relevant referent semantics, and we fuse its associated representations into the initial referent query:

$$q_0 = \Phi\big(\{\bar{h}^{\text{ans}}_{\mathcal{M},\ell}\}_{\ell\in\mathcal{L}_q}\big)\,, \tag{1}$$

where $\bar{h}^{\text{ans}}_{\mathcal{M},\ell}$ is the mean hidden representation over the answer-associated hidden states at layer $\ell$, and $\mathcal{L}_q$ indexes the selected backbone layers. The fusion module $\Phi$ performs learned cross-layer aggregation followed by state gating to produce the initial referent query $q_0 \in \mathbb{R}^d$. Rather than treating generated coordinates as the grounding output, $q_0$ consolidates the backbone's high-level referent understanding into a semantic initialization for spatial grounding. We further extract multimodal memories $\{F_s\}_{s=1}^{S}$ comprising the hidden representations of visual and instruction tokens, which provide stage-specific visual–language evidence for the subsequent grounding stages.

Starting from $q_0$ and zero-initialized spatial logits $B_0 = 0$, corresponding to a uniform spatial prior, the GDE Head applies $S = 4$ Grounding Distribution Update (GDU) Blocks:

$$(q_s, B_s) = \text{GDU}_s\left(q_{s-1}, B_{s-1}, F_s\right), \qquad s = 1, \ldots, S. \tag{2}$$

Through the GDE Head, the preceding affordance state feeds spatial evidence back to refine the referent query, which in turn guides the affordance update, forming a referent–affordance evolution loop. The terminal logits $B_S$ are upsampled to produce the grounding heatmap $H$.

We collect spatial grounding data across general and embodied scenarios with heterogeneous point, box, and mask annotations, forming a unified corpus for spatial grounding. We further unify these annotations as target distributions and combine distribution matching with a geometry-aware constraint that penalizes spatially inconsistent predictions.

### 3.2 Grounding Distribution Evolution Head

The GDE Head comprises $S = 4$ independently parameterized GDU Blocks that progressively couple the referent query $q_s$ and spatial affordance state $B_s$ through reciprocal updates. Each stage operates on a multimodal representation $F_s$ from a different backbone depth, with its visual tokens restored and aligned to the current spatial grid as $V_s \in \mathbb{R}^{|\Omega_s|\times d_{v,s}}$.

**Affordance-guided referent refinement.** At stage $s$, the preceding affordance state $B_{s-1}$ provides spatial feedback by weighting visual evidence for referent refinement. We denote its logits resized to the current spatial grid $\Omega_s$ as $\widetilde{B}_{s-1}$. The resulting affordance-weighted visual readout is:

$$r_s = \sum_{x\in\Omega_s} \text{softmax}_{\Omega_s}(\widetilde{B}_{s-1})(x) V_s(x). \tag{3}$$

The distribution-weighted readout $r_s$ provides visual feedback for referent refinement. We first contextualize $q_{s-1}$ through cross-attention to the current multimodal memory $F_s$, retaining the incoming query through a residual connection. We then add the projected visual readout $W_s^r r_s$ to the contextualized query and apply an FFN with a residual connection to obtain $q_s$. In this way, the preceding spatial prediction influences the referent representation used to guide the current affordance update. With $B_0 \equiv 0$, the first-stage readout uniformly averages the visual features; subsequent readouts use spatial weights obtained by normalizing the aligned logits from the preceding stage.

**Referent-guided affordance update.** The refined referent query $q_s$ scores the current visual grid to produce incremental spatial evidence, which is accumulated with the aligned preceding logits:

$$B_s(x) = \widetilde{B}_{s-1}(x) + \frac{\langle W_s^q q_s,\, W_s^k V_s(x)\rangle}{\sqrt{d}}, \qquad x \in \Omega_s. \tag{4}$$

Here, $W_s^q$ and $W_s^k$ project the referent and visual representations into a common $d$-dimensional matching space. The resulting scores add referent-conditioned evidence to the aligned spatial logits, reinforcing or revising earlier predictions. Each non-terminal spatial state then guides visual evidence selection for the next referent update, sustaining the reciprocal interaction, while the terminal state yields the grounding output.

### 3.3 UNIFIED DISTRIBUTIONAL GROUNDING SUPERVISION

Grounding annotations provide spatial supervision at heterogeneous granularities, ranging from points to boxes and masks. To supervise the evolving affordance distributions within a shared prediction space, we convert these annotation types into target distributions and optimize them with two complementary objectives: stage-wise distribution matching and a geometry-aware constraint on spatially inconsistent dominant responses.

**Unified distribution targets.** To express location preferences and valid-region constraints within a shared target space, we adopt an energy-based formulation that accommodates the spatial information provided by each annotation type. For an annotation $a$ with type $t_a \in \{\text{point}, \text{box}, \text{mask}\}$, we define its spatial energy over the stage grid $\Omega_s$ as:

$$E_a^{(s)}(x) = \begin{cases} \dfrac{\|x - y_a\|_2^2}{2(\sigma_p^{(s)})^2}, & t_a = \text{point}, \\ \dfrac{1}{2}(x - \mu_a)^\top \Sigma_a^{-1}(x - \mu_a) + \iota_{\mathcal{R}_a}(x), & t_a = \text{box}, \\ \iota_{\mathcal{R}_a}(x), & t_a = \text{mask}. \end{cases} \tag{5}$$

Here, $y_a$ is the annotated point location, and $\sigma_p^{(s)} > 0$ controls its target spread. For a box with center $\mu_a$, width $w_a$, and height $h_a$, we set $\Sigma_a = \text{diag}((\kappa w_a)^2, (\kappa h_a)^2)$, where $\kappa > 0$ is a dimensionless spread factor. For box and mask annotations, $\mathcal{R}_a$ denotes the box interior and mask foreground, respectively. The indicator energy $\iota_{\mathcal{R}_a}(x)$ is zero inside $\mathcal{R}_a$ and $+\infty$ outside. All annotation geometry is mapped to the current stage grid, with stage indices suppressed for brevity; coordinates and distance-valued scales use the same units. These definitions require positive box side lengths and nonempty rasterized support for box and mask targets.

Normalizing the energy as $T_a^{(s)} = \text{softmax}_{\Omega_s}(-E_a^{(s)})$ yields a discretized Gaussian around the annotated point, a box-truncated anisotropic Gaussian, or a uniform distribution over mask foreground. Thus, the point target encodes proximity to a preferred location, the box target combines a center preference with region support, and the mask target assigns equal weight to foreground locations. All three are expressed in the same normalized distributional space. At every stage, we match the predicted affordance distribution $p_s$ to this annotation-derived target using cross-entropy:

$$\mathcal{L}_{\text{dist}}^{(s)} = -\sum_{x \in \Omega_s} T_a^{(s)}(x) \log p_s(x). \tag{6}$$

This provides a shared distributional supervision across annotation types and directly constrains the evolving spatial predictions throughout the co-evolution process.

**Geometry-aware false-peak constraint.** Distribution matching captures the overall target geometry but does not explicitly penalize predicted spatial mass according to their distance from valid target locations. Let $d_a^{(S)}(x)$ denote the Euclidean distance from location $x$ to the valid set specified by annotation $a$. We convert this distance into a geometric affinity:

$$A_a(x) = \exp\left(-\frac{[d_a^{(S)}(x) - r_a]_+^2}{2\gamma_a^2}\right), \qquad [z]_+ = \max(z, 0), \tag{7}$$

where $r_a$ defines a tolerance around the valid region and $\gamma_a$ controls the decay with spatial distance. We sharpen the terminal prediction as: $p_{S,\tau} = \text{softmax}(B_S/\tau)$, with $0 < \tau < 1$, to emphasize dominant responses and penalize low expected geometric affinity:

$$\mathcal{L}_{\text{fp}} = \frac{-\log \mathbb{E}_{x \sim p_{S,\tau}}[A_a(x)]}{\max\{-\log \mathbb{E}_{x \sim U(\Omega_S)}[A_a(x)],\, c_{\min}\}}. \tag{8}$$

The denominator uses affinity under a uniform prediction as reference, reducing loss-scale variation across target geometries. The constraint emphasizes dominant responses while penalizing spatial mass geometrically inconsistent with valid targets. The overall grounding objective is:

$$\mathcal{L}_{\text{ground}} = \sum_{s=1}^{S} \alpha_s \mathcal{L}_{\text{dist}}^{(s)} + \lambda_{\text{fp}} \mathcal{L}_{\text{fp}}, \tag{9}$$

where $\alpha_s$ and $\lambda_{\text{fp}}$ weight stage-wise distribution matching and the geometry-aware constraint, respectively. Together, they supervise the evolving affordance distributions across stages while enforcing geometric consistency at the final prediction.

# 4 EXPERIMENTS

## 4.1 EXPERIMENTAL SETUP

**Datasets.** We construct a mixed-supervision grounding corpus for training from a curated subset of RefSpatial (Zhou et al., 2025), RoboRefIt (Lu et al., 2023), RefCOCO/RefCOCO+/RefCOCOg (Yu et al., 2016; Mao et al., 2016), and SpatialGrasp, covering object localization and spatially constrained placement with point, box, and mask annotations. SpatialGrasp extends GraspNet (Fang et al., 2020) with single-target referring expressions paired with filtered instance masks. We convert spatial annotations into stage-wise target distributions following Sec. 3.3. Dataset statistics, construction procedures, and quality control are detailed in Appendix B.

**Implementation Details.** We initialize the multimodal backbone from NVILA (Liu et al., 2025) and jointly optimize its language backbone and the proposed GDE Head under the grounding objective in Sec.3.3, while keeping the visual tower and projector frozen. The four GDU Blocks operate on intermediate representations from layers $(7, 14, 21, 28)$. For grounding evaluation, we decode points with SoftArgmax and report point-in-mask accuracy on all grounding benchmarks except the RefCOCO family, for which we report point-in-box accuracy. Optimization and evaluation details are provided in Appendix A.

## 4.2 COMPARISON WITH EXISTING METHODS

**Embodied Spatial Evaluation.** RACE achieves 54.50% on RefSpatial-Bench, the highest reported score among the listed 2B–3B models (Table 1). On Where2Place, our 2B model obtains the highest listed score of 74.00%, exceeding RoboMAP (3B) and RoboRefer-RFT (2B) by 1.00 and 3.00 percentage points, respectively. RACE also achieves the highest listed score of 88.24%

Table 1: Embodied spatial evaluation on RefSpatial-Bench, RoboRefIt, Where2Place, and EmbSpatial, covering embodied grounding and spatial understanding. All scores are reported as percentages; higher is better. Parameter counts denote total / active parameters where applicable. Bold and underlined values indicate the best and second-best comparable results, respectively. Baseline sources and protocol differences are summarized in Appendix A.

| Model | Params. | RefSpatial-Bench | RoboRefIt | Where2Place | EmbSpatial |
|---|---|---|---|---|---|
| RoboRefer-SFT (Zhou et al., 2025) | 2B | 47.50 | 72.80 | 66.00 | 70.66 |
| RoboRefer-RFT (Zhou et al., 2025) | 2B | 53.00 | 74.20 | 71.00 | 71.10 |
| RynnBrain 1.0 (Dang et al., 2026) | 2B | 52.70 | 80.36 | 62.98 | 78.71 |
| Si0 (Zhou et al., 2026) | 2B | 42.60 | 85.09 | 64.08 | <u>84.37</u> |
| Capek0.5 (Chen et al., 2026) | 2B | – | 79.89 | 35.86 | 78.68 |
| Cosmos-Reason2 (NVIDIA, 2025) | 2B | 30.70 | 53.35 | 33.00 | 81.65 |
| Qwen3-VL (Bai et al., 2025) | 2B | 28.90 | 74.89 | 45.00 | 75.90 |
| Embodied-R1 (Yuan et al., 2026b) | 3B | 39.70 | 85.58 | 69.50 | 67.40 |
| RoboMAP (Shao et al., 2025) | 3B | – | <u>88.73</u> | <u>73.00</u> | – |
| RoboPIN (Huang et al., 2026) | 4B | 50.50 | **89.00** | 62.00 | 83.10 |
| RoboBrain2.5 (Tan et al., 2026) | 4B | **56.00** | 82.65 | 65.00 | 73.80 |
| HY-Embodied-0.5 (Tencent Robotics X et al., 2026) | 4B / 2B | 45.80 | 82.80 | 68.00 | 82.80 |
| Molmo2-ER (Fang et al., 2026) | 5B | 52.50 | – | 54.00 | 78.80 |
| MiMo-Embodied (Hao et al., 2026) | 7B | 48.00 | 82.30 | 63.60 | 76.24 |
| FSD (Yuan et al., 2026a) | 13B | – | 56.73 | 45.81 | 63.30 |
| **Ours** | 2B | <u>54.50</u> | 81.33 | **74.00** | **88.24** |

Table 2: General spatial evaluation on CV-Bench, SAT, BLINK, and the RefCOCO family, covering spatial understanding and referring-expression grounding. All scores are reported as percentages; higher is better. Parameter counts denote total / active parameters where applicable. Bold and underlined values indicate the best and second-best comparable results, respectively. RefCOCO family scores average RefCOCO, RefCOCO+, and RefCOCOg accuracies, including test splits. Baseline results follow their source-reported protocols. Evaluation protocols and baseline sources are detailed in Appendix A.

| Model | Params. | CV-Bench | SAT | BLINK | | RefCOCO |
|---|---|---|---|---|---|---|
| | | | | Depth | Spatial | |
| RoboRefer-SFT (Zhou et al., 2025) | 2B | – | 71.33 | **88.71** | 83.92 | 85.94 |
| RynnBrain 1.0 (Dang et al., 2026) | 2B | 86.49 | 64.67 | <u>87.90</u> | 79.70 | 78.48 |
| Si0 (Zhou et al., 2026) | 2B | 88.14 | 76.00 | – | – | 86.74 |
| Cosmos-Reason2 (NVIDIA, 2025) | 2B | 78.70 | 70.00 | 83.10 | 75.50 | 80.80 |
| Qwen3-VL (Bai et al., 2025) | 2B | 80.00 | 45.30 | 74.20 | 77.60 | 83.31 |
| Embodied-R1 (Yuan et al., 2026b) | 3B | 82.70 | 76.30 | 76.60 | 80.40 | – |
| RoboPIN (Huang et al., 2026) | 4B | 85.50 | 78.00 | 87.10 | <u>87.40</u> | – |
| RoboBrain2.5 (Tan et al., 2026) | 4B | 86.90 | 51.30 | – | – | 84.21 |
| HY-Embodied-0.5 (Tencent Robotics X et al., 2026) | 4B / 2B | **89.20** | 76.70 | – | – | – |
| Molmo2-ER (Fang et al., 2026) | 5B | 87.80 | 78.00 | – | – | – |
| MiMo-Embodied (Hao et al., 2026) | 7B | <u>88.82</u> | <u>78.67</u> | 81.50 | 83.20 | 74.30 |
| FSD (Yuan et al., 2026a) | 13B | 80.90 | 63.30 | 70.20 | 78.30 | – |
| **Ours** | 2B | 85.31 | **81.33** | 84.68 | **90.21** | **86.86** |

on EmbSpatial, surpassing Si0 by 3.87 percentage points, and obtains 81.33% on RoboRefIt. These results indicate competitive performance in embodied spatial grounding at comparable model scales.

**General Spatial Evaluation.** RACE achieves the highest listed scores on SAT (81.33%) and BLINK Spatial (90.21%), exceeding MiMo-Embodied and RoboPIN by 2.66 and 2.81 percentage points, respectively (Table 2). Our model also achieves 86.86% on the RefCOCO family, demonstrating its referring-expression grounding capability in general visual scenes. Together, these results highlight RACE's strong embodied grounding capability alongside competitive spatial reasoning and understanding in general visual scenes.

### 4.3 Ablation Study

**Grounding beyond text-coordinate generation.** As shown in Table 3, the GDE head consistently outperforms direct text-coordinate prediction across all evaluated settings, yielding gains of 5.00–9.22 points, including a 7.80-point improvement on the Unseen split of RefSpatial-Bench. These results support explicitly decoding spatial distributions for grounding in the evaluated settings. Figure 3 illustrates this advantage in both object localization and free-space grounding.

**Referent Query initialization and persistent evolution.** We examine the roles of referent query initialization and cross-stage persistence (Table 3). Replacing the answer-conditioned query with the pre-answer hidden state reduces accuracy from 54.50% to 32.85% on RefSpatial-Bench and from 74.00% to 50.00% on Where2Place, demonstrating the importance of informative referent semantics for query initialization. Resetting the query at each stage (*w/o Query persistence*) also degrades performance to 41.88% and 66.00%, respectively, showing that propagating the refined query provides gains beyond spatial-state propagation alone. Together, these results show that fine-grained localization relies on both informative initialization of referent semantics and their continuous refinement throughout the grounding process.

**Spatial feedback and selectivity.** To isolate the contribution of spatial feedback from query persistence, we remove the visual readout while retaining cross-stage query propagation (Table 3). Accuracy drops from 54.50% to 44.77% on RefSpatial-Bench and from 74.00% to 70.00% on




Figure 3: Qualitative comparison between direct text-coordinate generation and Ours. The first two examples expose failures in object localization and free-space grounding, where text-coordinate predictions fall outside the target region, while our method localizes the intended region. In the remaining examples, both approaches correctly localize the target region, with our predictions visually better aligned with the annotated regions.

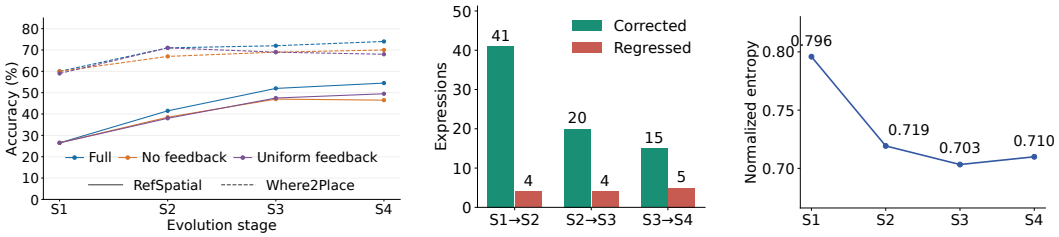
Where2Place, showing that query propagation alone is insufficient. We further replace spatially weighted readout with uniform averaging, which yields 44.04% and 68.00%, respectively, comparable to or worse than removing feedback. Together, these results show that where-to-what feedback benefits primarily from spatially selective readout rather than visual feedback alone. The evolving spatial state thus serves both as an intermediate grounding prediction and as a selection prior for referent refinement.

**Stage-wise Evolution Analysis** We extract and analyze intermediate predictions at each stage to examine how affordance distributions evolve with spatial feedback. Co-evolution consistently improves localization across stages on both benchmarks, with spatially selective feedback supporting sustained refinement (Figure 4(a)). Gains are largest in the early stages, while smaller later-stage improvements suggest a trend toward saturation. The RefSpatial-Bench transition analysis further shows that corrections substantially outnumber regressions at every update (Figure 4(b)). Later stages thus continue to repair errors, although individual predictions can regress. Selected four-stage visualizations appear in Figure 8 of Appendix C.

To examine how spatial concentration changes during refinement, we measure the mean normalized entropy across stages on RefSpatial-Bench. Entropy reaches its minimum of 0.703 at stage 3, then

Table 3: Grounding-head comparison and component ablations. Accuracy (%) is reported; higher is better. RefSpatial-Bench results are reported by subset. All ablated variants are trained independently under the same training and evaluation protocol as corresponding full configurations. *Text-Coordinate* uses coordinates generated by the language backbone. *Pre-answer query* initializes the query with a hidden state obtained before answer generation. *w/o Query persistence* keeps the query fixed at its initial state across all stages while retaining spatial logit accumulation. *No feedback* removes visual feedback from query refinement. *Uniform feedback* replaces visual pooling weighted by the previous spatial state with uniform spatial pooling. *RACE (full)* uses all components.

| Variant | RefSpatial-Bench | | | Where2Place | RoboRefIt |
|---|---|---|---|---|---|
| | Location | Placement | Unseen | | |
| Text-Coordinate | 53.00 | 45.00 | 36.36 | 66.00 | 72.11 |
| Pre-answer query | 36.00 | 30.00 | 32.47 | 50.00 | 78.97 |
| *w/o* Query persistence | 49.00 | 41.00 | 33.77 | 66.00 | 70.17 |
| No feedback | 54.00 | 40.00 | 38.96 | 70.00 | 78.16 |
| Uniform feedback | 52.00 | 40.00 | 38.96 | 68.00 | 78.74 |
| **RACE (full)** | **58.00** | **51.00** | **44.16** | **74.00** | **81.33** |




(a) Stage-wise accuracy (b) Prediction transitions (c) Distribution entropy

Figure 4: Stage-wise evolution of grounding predictions. (a) Stage-wise accuracy under different feedback settings on RefSpatial-Bench and Where2Place. (b) Counts of corrected (wrong-to-correct) and regressed (correct-to-wrong) predictions between consecutive stages on RefSpatial-Bench. (c) Mean normalized entropy of affordance distributions on RefSpatial-Bench, with lower values indicating greater spatial concentration. Additional distribution diagnostics are provided in Figure 6 of Appendix C.

rises slightly to 0.710 at the final stage. Meanwhile, accuracy increases from 52.00% to 54.50% over the final transition, indicating that further improvements in localization do not require the spatial distribution to become increasingly sharp. Later-stage refinement may instead improve alignment with the referent semantics while preserving spatial uncertainty, rather than simply sharpening the affordance distribution.

### 4.4 Real-world Robotic Experiments

The GDE Head supplies task-conditioned spatial features to an Action Head that predicts pick-and-place trajectories. The Action Head is trained on simulated demonstrations while the backbone and GDE Head remain frozen. We evaluate 50 real-world trials across five conditions, with 10 trials per condition. Success requires grasping the intended target and releasing it in the designated area. RACE completes 36 of 50 full tasks, with 6–8 completions per condition (Figure 5). These completed tasks indicate that the learned grounding representations can support downstream manipulation, providing an interface from language-conditioned spatial understanding to physical execution. Appendix C details deployment, training, and qualitative cases.

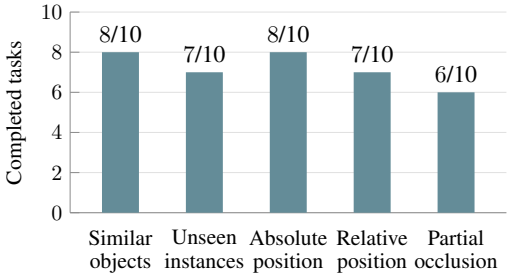



Figure 5: Real-world pick-and-place results. Bars and labels report completed tasks out of 10 trials per condition.

## 5 Conclusion

We introduced RACE, a framework for embodied spatial grounding that bridges what to attend to and where to act through the co-evolution of referent semantics and an affordance state. Referent semantics guide updates to the affordance state, which then selects visual evidence to refine referent semantics. To guide this reciprocal evolution, we map point, box, and mask annotations into a shared distributional space and directly supervise the evolving affordance distributions at every stage. Our 2B model achieves 54.50% on RefSpatial-Bench and 74.00% on Where2Place while remaining competitive on general spatial benchmarks. Ablations support multi-stage referent refinement and spatially selective feedback, while real-world pick-and-place indicates RACE's applicability to instruction-guided manipulation. Our results suggest that co-evolving referent semantics and visual evidence provides a general pathway toward extending language-model backbones with fine-grained spatial grounding capabilities.

## AI USE STATEMENT

In the preparation of this manuscript, relevant tools were used to assist with literature and information research, language polishing, and formatting adjustments. The implementation of the experiments, analysis and interpretation of the results, and preparation of the initial manuscript draft were completed independently by the author. The tools were used solely to improve the efficiency of information organization, language expression, and formatting, and were not involved in the experimental procedures, interpretation of results, or generation of the core academic content.

## REFERENCES

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966. arXiv preprint arXiv:2308.12966.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report, 2025. URL https://arxiv.org/abs/2511.21631. arXiv preprint arXiv:2511.21631.

Ying Chen, Weizhen Li, Zhe Hu, Zhenjiang Li, Rui Jiang, Zhifeng Gu, Lihuang Fang, Jiangping Liu, Lei Yi, and Jie Chen. Capek 0.5: An execution-centric vision-language model for embodied intelligence, 2026. URL https://arxiv.org/abs/2608.06756. arXiv preprint arXiv:2608.06756.

Ming Dai, Wenxuan Cheng, Jiang-jiang Liu, Sen Yang, Wenxiao Cai, Yanpeng Sun, and Wankou Yang. DeRIS: Decoupling perception and cognition for enhanced referring image segmentation through loopback synergy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19936–19946, 2025. URL https://arxiv.org/abs/2507.01738.

Ronghao Dang, Jiayan Guo, Bohan Hou, Sicong Leng, Kehan Li, Xin Li, Jiangpin Liu, Yunxuan Mao, Zhikai Wang, Yuqian Yuan, Minghao Zhu, Xiao Lin, Yang Bai, Qian Jiang, Yaxi Zhao, Minghua Zeng, Junlong Gao, Yuming Jiang, Jun Cen, Siteng Huang, Liuyi Wang, Wenqiao Zhang, Chengju Liu, Jianfei Yang, Shijian Lu, and Deli Zhao. RynnBrain: Open embodied foundation models, 2026. URL https://arxiv.org/abs/2602.14979. arXiv preprint arXiv:2602.14979.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 91–104, 2025. URL https://arxiv.org/abs/2409.17146.

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)*, 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.

Haoquan Fang, Jiafei Duan, Donovan Clay, Sam Wang, Shuo Liu, Weikai Huang, Xiang Fan, Wei-Chuan Tsai, Shirui Chen, Yi Ru Wang, Shanli Xing, Jaemin Cho, Jae Sung Park, Ainaz Eftekhar, Peter Sushko, Karen Farley, Angad Wadhwa, Cole Harrison, Winson Han, Ying-Chun Lee, Eli VanderBilt, Rose Hendrix, Suveen Ellawela, Lucas Ngoo, Joyce Chai, Zhongzheng Ren, Ali Farhadi, Dieter Fox, and Ranjay Krishna. MolmoAct2: Action reasoning models for real-world deployment, 2026. URL https://arxiv.org/abs/2605.02881. arXiv preprint arXiv:2605.02881.

Xiaoshuai Hao, Lei Zhou, Zhijian Huang, Zhiwen Hou, Yingbo Tang, Lingfeng Zhang, Guang Li, Zheng Lu, Shuhuai Ren, Xianhui Meng, Yuchen Zhang, Jing Wu, Jinghui Lu, Chenxu Dang, Jiayi Guan, Jianhua Wu, Zhiyi Hou, Hanbing Li, Shumeng Xia, Mingliang Zhou, Yinan Zheng, Zihao Yue, Shuhao Gu, Hao Tian, Yuannan Shen, Jianwei Cui, Wen Zhang, Shaoqing Xu, Bing Wang, Haiyang Sun, Zeyu Zhu, Yuncheng Jiang, Zibin Guo, Chuhong Gong, Chaofan Zhang, Wenbo Ding, Kun Ma, Guang Chen, Rui Cai, Diyun Xiang, Heng Qu, Fuli Luo, Hangjun Ye, and Long Chen. MiMo-Embodied: X-Embodied foundation model technical report, 2026. URL https://arxiv.org/abs/2511.16518v2. arXiv preprint arXiv:2511.16518.

Haifeng Huang, Xinyi Chen, Yilun Chen, Hao Li, Xiaoshen Han, Zehan Wang, Tai Wang, Jiangmiao Pang, and Zhou Zhao. RoboGround: Robotic manipulation with grounded vision-language priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22540–22550, 2025. URL https://arxiv.org/abs/2504.21530.

Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. VoxPoser: Composable 3D value maps for robotic manipulation with language models. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 540–562. PMLR, 2023. URL https://proceedings.mlr.press/v229/huang23b.html.

Yaoting Huang, Yifu Yuan, Linqi Han, Chengwen Li, Shuoheng Zhang, Xianze Yao, Hongyao Tang, Yan Zheng, and Jianye Hao. RoboPIN: Grounded embodied reasoning via pinned chain-of-thought, 2026. URL https://arxiv.org/abs/2606.15753. arXiv preprint arXiv:2606.15753.

Weitai Kang, Jason Kuen, Mengwei Ren, Zijun Wei, Yan Yan, and Kangning Liu. VGent: Visual grounding via modular design for disentangling reasoning and prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41160–41170, 2026. URL https://arxiv.org/abs/2512.11099.

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, 2024. URL https://arxiv.org/abs/2308.00692.

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, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Li_LOCATE_Localize_and_Transfer_Object_Parts_for_Weakly_Supervised_Affordance_CVPR_2023_paper.html.

Peiyan Li, Yixiang Chen, Hongtao Wu, Xiao Ma, Xiangnan Wu, Yan Huang, Liang Wang, Tao Kong, and Tieniu Tan. BridgeVLA: Input-output alignment for efficient 3D manipulation learning with vision-language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 63635–63673. Curran Associates, Inc., 2025. doi: 10.52202/085713-2137. URL https://arxiv.org/abs/2506.07961.

Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, Xiuyu Li, Haotian Tang, Yunhao Fang, Yukang Chen, Cheng-Yu Hsieh, De-An Huang, An-Chieh Cheng, Jinyi Hu, Sifei Liu, Ranjay Krishna, Pavlo Molchanov, Jan Kautz, Hongxu Yin, Song Han, and Yao Lu. NVILA: Efficient frontier visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4122–4134, 2025. URL https://arxiv.org/abs/2412.04468.

Yuhao Lu, Yixuan Fan, Beixing Deng, Fangfu Liu, Yali Li, and Shengjin Wang. VL-Grasp: a 6-DoF interactive grasp policy for language-oriented objects in cluttered indoor scenes. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 976–983, 2023. URL https://arxiv.org/abs/2308.00640.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Mao_Generation_and_Comprehension_CVPR_2016_paper.html.

NVIDIA. Cosmos-Reason2-2B, 2025. URL https://huggingface.co/nvidia/Cosmos-Reason2-2B. Official model card.

NVIDIA. Cosmos 3: Omnimodal world models for physical AI, 2026. URL https://arxiv.org/abs/2606.02800. arXiv preprint arXiv:2606.02800.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.14824.

Xinyu Shao, Yanzhe Tang, Pengwei Xie, Kaiwen Zhou, Yuzheng Zhuang, Xingyue Quan, Jianye Hao, Long Zeng, and Xiu Li. More than a point: Capturing uncertainty with adaptive affordance heatmaps for spatial grounding in robotic tasks, 2025. URL https://arxiv.org/abs/2510.10912. arXiv preprint arXiv:2510.10912.

Huajie Tan, Enshen Zhou, Zhiyu Li, Yijie Xu, Yuheng Ji, Xiansheng Chen, Cheng Chi, Pengwei Wang, Huizhu Jia, Yulong Ao, Mingyu Cao, Sixiang Chen, Zhe Li, Mengzhen Liu, Zixiao Wang, Shanyu Rong, Yaoxu Lyu, Zhongxia Zhao, Peterson Co, Yibo Li, Yi Han, Shaoxuan Xie, Guocai Yao, Songjing Wang, Leiduo Zhang, Xi Yang, Yance Jiao, Donghai Shi, Kunchang Xie, Shaokai Nie, Chunlei Men, Yonghua Lin, Zhongyuan Wang, Tiejun Huang, and Shanghang Zhang. RoboBrain 2.5: Depth in sight, time in mind, 2026. URL https://arxiv.org/abs/2601.14352. arXiv preprint arXiv:2601.14352.

Tencent Robotics X, HY Vision Team, Xumin Yu, Zuyan Liu, Ziyi Wang, He Zhang, Yongming Rao, Fangfu Liu, Yani Zhang, Ruowen Zhao, Oran Wang, Yves Liang, Haitao Lin, Minghui Wang, Yubo Dong, Kevin Cheng, Bolin Ni, Rui Huang, Han Hu, Zhengyou Zhang, Linus, and Shunyu Yao. HY-Embodied-0.5: Embodied foundation models for real-world agents, 2026. URL https://arxiv.org/abs/2604.07430. arXiv preprint arXiv:2604.07430.

Qianhui Wu, Kanzhi Cheng, Rui Yang, Chaoyun Zhang, Jianwei Yang, Huiqiang Jiang, Jian Mu, Baolin Peng, Bo Qiao, Reuben Tan, Si Qin, Lars Liden, Qingwei Lin, Huan Zhang, Tong Zhang, Jianbing Zhang, Dongmei Zhang, and Jianfeng Gao. GUI-Actor: Coordinate-free visual grounding for GUI agents. In *Advances in Neural Information Processing Systems*, volume 38, pp. 15101–15128. Curran Associates, Inc., 2025. doi: 10.52202/085713-0508. URL https://arxiv.org/abs/2506.03143.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.07704.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling context in referring expressions. In *Computer Vision – ECCV 2016*, pp. 69–85. Springer International Publishing, 2016. doi: 10.1007/978-3-319-46475-6_5. URL https://doi.org/10.1007/978-3-319-46475-6_5.

Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. RoboPoint: A vision-language model for spatial affordance prediction in robotics. In *Proceedings of The 8th Conference on Robot Learning*, volume 270 of *Proceedings of Machine Learning Research*, pp. 4005–4020. PMLR, 2025. URL https://proceedings.mlr.press/v270/yuan25c.html.

Yifu Yuan, Haiqin Cui, Yibin Chen, Zibin Dong, Fei Ni, Longxin Kou, Jinyi Liu, Pengyi Li, Yan Zheng, and Jianye Hao. From seeing to doing: Bridging reasoning and decision for robotic manipulation. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/d2eedb7ecf14b50967c5de98e30e0ffc-Paper-Conference.pdf.

Yifu Yuan, Haiqin Cui, Yaoting Huang, Yibin Chen, Fei Ni, Zibin Dong, Pengyi Li, Yan Zheng, Hongyao Tang, and Jianye Hao. Embodied-R1: Reinforced embodied reasoning for general robotic manipulation. In *International Conference on Learning Representations*, 2026b. URL https://arxiv.org/abs/2508.13998.

Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. GROUNDHOG: Grounding large language models to holistic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14227–14238, 2024. URL https://arxiv.org/abs/2402.16846.

Zekang Zhang, Guangyu Gao, Youyun Tang, ChengJing Wu, Xiaochao Qu, Chi Harold Liu, Jianbo Jiao, Yunchao Wei, Luoqi Liu, and Ting Liu. FlowSeg: Dynamic semantic guidance for LLM-conditioned segmentation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2605.29461.

Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, and Shanghang Zhang. RoboRefer: Towards spatial referring with reasoning in vision-language models for robotics. In *Advances in Neural Information Processing Systems*, volume 38, pp. 28404–28481. Curran Associates, Inc., 2025. doi: 10.52202/085713-0957. URL https://arxiv.org/abs/2506.04308.

Yanning Zhou, Wentao Yuan, Chi Yan, Honghui Wang, Guoyang Zhao, Feng Qiu, Kewei Wang, and Yixiao Ge. Si0: The capabilities that transfer, perceived through the robot's eyes. *XPENG Robotics Blog*, 2026. URL https://xpeng-robotics.github.io/si0/. Official technical blog.

# APPENDIX

## A Method, Training, and Evaluation Details

Table 4 gives the main GDE Head settings.

Table 4: Main GDE Head settings.

| Component | Configuration |
|---|---|
| Input feature width | 1536 for visual and instruction memories |
| Query / key / value width | 128 / 128 / 128 |
| Cross-attention per GDU | 4 heads with 32 dimensions per head |
| FFN | $128 \rightarrow 512 \rightarrow 128$; GELU activation |
| Stage grids | $8 \times 8$, $16 \times 16$, $32 \times 32$, $32 \times 32$ |
| Output heatmap | $224 \times 224$ |

The four GDUs are independently parameterized. Visual tokens from each selected backbone layer are aligned to that stage's grid, while the associated instruction tokens remain in the multimodal memory used for query cross-attention. The last two stages share a $32 \times 32$ grid but read different layer memories and apply separate updates; the query width stays fixed as the grid resolution changes.

Table 5 lists the loss settings. Coordinates and distance-valued parameters use normalized image coordinates; $H_s$ denotes the stage-grid height, and the terminal constraint uses $r_a = r_a^{(S)}$.

Table 5: Supervision and loss settings.

| Parameter | Symbol | Value |
|---|---|---|
| Stage-wise loss weights | $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ | $(0.1, 0.1, 0.1, 0.1)$ |
| False-peak loss weight | $\lambda_{\text{fp}}$ | $0.3$ |
| Point-target spread: object / placement | $\sigma_p^{(s)}$ | $0.3$ / $0.5$ |
| Box spread factor | $\kappa$ | $0.15$ |
| Geometric tolerance | $r_a^{(s)}$ | $\max\{0.08, 1/(H_s - 1)\}$ |
| Geometric decay scale | $\gamma_a$ | $6/223$ |
| Sharpening temperature | $\tau$ | $0.5$ |
| Normalization floor | $c_{\min}$ | $1.0$ |

Loss targets are reconstructed at each stage's grid resolution. Point spread has separate settings for object localization and placement, whereas box spread scales with the annotated box dimensions. The geometric tolerance is the larger of a fixed radius and one grid interval; the terminal false-peak constraint uses this tolerance.

At inference, only the terminal heatmap is decoded into an output point; intermediate distributions remain available for stage-wise analysis. For output logits $H \in \mathbb{R}^{h \times w}$, SoftArgmax gives the normalized point prediction

$$P = \mathrm{softmax}_{i,j}(H), \qquad (\widehat{x}, \widehat{y}) = \sum_{i=0}^{h-1} \sum_{j=0}^{w-1} P_{ij} \left( \frac{j}{w-1}, \frac{i}{h-1} \right). \tag{10}$$

For an image of width $W_I$ and height $H_I$, the corresponding pixel coordinates are $((W_I - 1)\widehat{x}, (H_I - 1)\widehat{y})$.

We train for one epoch over the full training corpus using the settings in Table 6. All language-backbone and GDE Head parameters are optimized; the visual encoder and multimodal projector remain frozen. The language and visual backbones use FP16, while the GDE Head, optimizer master parameters, and optimizer states use FP32. We use the final checkpoint after this epoch for evaluation.

Table 6: Training configuration.

| Parameter | Setting |
|---|---|
| Optimizer | AdamW, $\beta = (0.9, 0.999)$, $\epsilon = 10^{-6}$ |
| Language-backbone learning rate | $1 \times 10^{-5}$ |
| GDE Head learning rate | $1 \times 10^{-4}$ |
| Learning-rate schedule | Linear warmup followed by cosine decay |
| Warmup | $3\%$ of total training steps |
| Minimum learning rate | $20\%$ of each base learning rate |
| Weight decay | $0$ |
| Hardware | 8 NVIDIA L20 GPUs |
| Global batch size | 256 (32 per GPU across 8 GPUs) |

Baseline scores are taken from original reports and cross-model evaluations in Si0 (Zhou et al., 2026), Cosmos 3 (NVIDIA, 2026), and HY-Embodied-0.5 (Tencent Robotics X et al., 2026), except for the dagger-marked RoboRefer-SFT result, which we re-evaluate on the RefCOCO family using

point-in-box accuracy (85.94%, pooled over all 57,457 examples). Capek0.5's RoboRefIt score uses box IoU@0.5. Source-reported scores retain their original evaluation settings and are not assumed to be directly comparable.

## B Dataset Construction

The training corpus contains 311,315 image–instruction records from RefSpatial (Zhou et al., 2025), the RefCOCO family (Yu et al., 2016; Mao et al., 2016), RoboRefIt (Lu et al., 2023), and SpatialGrasp (Table 7). Each record has unit sampling weight.

Table 7: Composition of the training corpus. Counts refer to image–instruction records used in training.

| Source | Task or subset | Records | Spatial supervision |
|---|---|---|---|
| RefSpatial | Choice-based QA | 40,000 | None |
| RefSpatial | Localization | 80,000 | Point |
| RefSpatial | Placement | 80,000 | Point |
| RefCOCO | Referring localization | 25,000 | Box |
| RefCOCO+ | Referring localization | 20,000 | Box |
| RefCOCOg | Referring localization | 15,000 | Box |
| RoboRefIt | Train split | 36,915 | Mask |
| SpatialGrasp | Two-camera collection | 14,400 | Mask |
| Total | | 311,315 | |

RefSpatial samples require more than two annotated reasoning steps; public RefCOCO and RoboRefIt annotations are used as supplied. Spatial targets are mapped to stage-wise distributions as in Sec. 3.3.

SpatialGrasp uses the official GraspNet-1Billion training scenes (Fang et al., 2020) (official dataset site). We select 3,600 RGB images from `scene_0000`–`scene_0099` across Kinect and RealSense views; each contributes four distinct targets, yielding 14,400 records. Frame spacing and image-similarity checks reduce redundant views, and each selected image must contain four usable targets.

We retain the largest mask component and require at least 256 foreground pixels, 0.05%–25% image area, a tight-box shorter side of at least 10 pixels, and nonempty foreground at $224 \times 224$. Boxes are recomputed from cleaned masks. Small or occluded targets remain eligible when identifiable; masks supervise visible foreground only.

Instructions use reviewed object phrases and geometric templates, without per-example language-model generation. Each image assigns its four targets to category, attribute, absolute-position, and relative-relation roles. Relative expressions use image-plane directions; ambiguous or multi-target instructions are excluded. Reference objects only locate the single manipulated target. In Ref v2, RGB images and masks are rotated by $180^\circ$, and spatial descriptions are recomputed in the final frame.

Automatic checks cover target identity, mask geometry, and language relations. Model-assisted visual review covers 720 records (5%); it is not a full human audit. No placement-destination or multi-step instructions were found in the final 14,400-record audit.

The original release partitions scenes into 80 training, 10 validation, and 10 audit scenes, keeping both camera views together. All 14,400 records are used here, so the original validation and audit sets are not independent holdouts. Historical checks against 2,137 RoboRefIt testA and RefSpatial-Bench images found no duplicate selected source or then-final images. They do not audit the rotated Ref v2 assets or rule out all semantic overlap.

## C Additional Analyses and Real-World Experiments

Table 8 reports our point-in-mask accuracy on each RefCOCO-family evaluation split.

Table 8: Our point-in-mask accuracy (%) on individual RefCOCO-family splits.

| RefCOCO | | | RefCOCO+ | | | RefCOCOg | |
|---|---|---|---|---|---|---|---|
| val | testA | testB | val | testA | testB | val | test |
| 91.79 | 93.65 | 90.72 | 82.20 | 87.34 | 78.16 | 85.27 | 85.43 |

Table 9: Multi-stage design choices. Accuracy (%) under the evaluation protocol of Table 3. Memory results report means over three independent runs. Grid variants retain four stages. † The default-grid row reports main-model scores rather than a verified matched-run control for the fixed-grid configuration. The memory and grid comparisons use distinct experimental runs.

| Design | RefSpatial-Bench | Where2Place |
|---|---|---|
| *Layer memory* | | |
| Full | 48.62 | 74.00 |
| Mean-layer memory | 43.68 | 71.67 |
| *Grid schedule* | | |
| $(8, 16, 32, 32)^{\dagger}$ | 54.50 | 74.00 |
| $(32, 32, 32, 32)$ | 46.93 | 75.00 |

Averaging visual and prefix memories across backbone layers preserves the answer-query tokens but removes stage-specific memory selection. Across three independently trained run pairs, the mean paired gains are 4.93 percentage points on RefSpatial-Bench and 2.33 on Where2Place (Table 9). These comparisons favor supplying each stage with its corresponding layer's contextual evidence over reusing a cross-layer average; they do not identify which individual layer is necessary.

Table 9 reports two four-stage grid configurations. The default $(8, 16, 32, 32)$ row uses the main-model scores of 54.50% on RefSpatial-Bench and 74.00% on Where2Place; the fixed $(32, 32, 32, 32)$ configuration achieves 46.93% and 75.00%, respectively. The default row is a main-model reference rather than a verified matched-run control for the fixed-grid configuration, so these results do not isolate the effect of the grid schedule.

Figure 6 reports six distribution diagnostics on all 277 RefSpatial-Bench expressions, complementing the prediction-transition analysis on the same evaluation set. From stage 3 to stage 4, target-region probability mass increases from 32.71% to 33.92%, and expected distance to the target decreases from 4.76% to 4.40% of the image diagonal. Over the same transition, SoftArgmax hit rate rises from 46.57% to 50.18% despite the increase in mean normalized entropy from 0.703 to 0.710. These aggregate observations distinguish target alignment from concentration; they do not establish that increasing entropy causes better localization or that every sample becomes more diffuse.

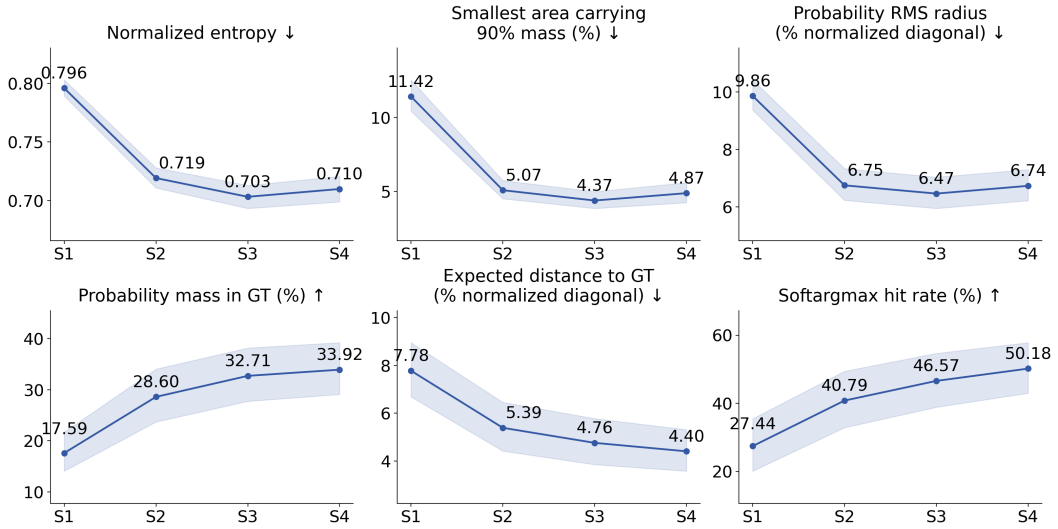



Figure 6: Distribution diagnostics across stages on all 277 RefSpatial-Bench expressions. Entropy, probability coverage area, and spatial spread measure concentration; target-region probability mass, expected distance to the target, and hit rate measure target alignment. The final stage improves alignment despite a slight increase in distribution spread. Shading shows 95% confidence intervals from 5,000 bootstrap resamples grouped by image.

Figure 7 provides a complementary distribution-level analysis on the 200 Location and Placement expressions. The number of expressions with no valid target peak among the top five decreases from 92 to 58, while the number whose highest-ranked peak lies in the target grows from 53 to 108. Peak coverage and point-localization accuracy measure different properties and are reported separately. For each $224 \times 224$ heatmap, we repeatedly select the highest-valued pixel and apply non-maximum suppression within a six-pixel radius to obtain five candidate peaks. A candidate is valid when its location falls inside the ground-truth mask; the suppression radius is not a hit tolerance.

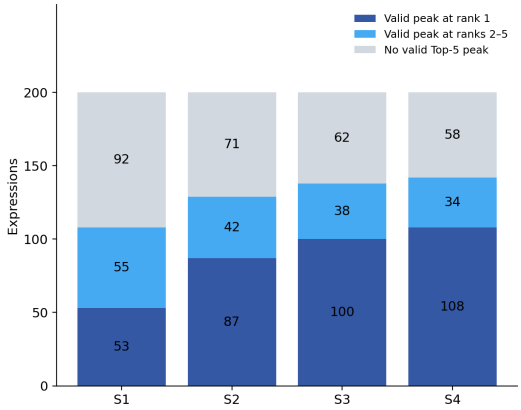



Figure 7: Target coverage among the top five spatial peaks on refspatial-bench. Expressions are grouped by whether a valid target peak is ranked first, is ranked between second and fifth, or is absent from the top five.

Figure 8 follows three selected RefSpatial-Bench expressions through four GDU stages: an object specified relative to the shovels, a free-space placement in front of the litter box, and the leftmost toy on the lower shelf. The green outline marks the target region, the yellow circle marks the evaluated SoftArgmax point, and the cyan cross marks the distribution maximum. In the first two rows, target-region probability mass rises from 30.5% to 81.1% and from 24.3% to 56.3%, respectively, between

stages 1 and 2, when the decoded point first enters the target. In the third row, target-region mass increases from 1.0% to 60.1% over all four stages, and the decoded point enters the target only at stage 4. These selected trajectories illustrate different rates of spatial alignment; they do not estimate their prevalence across the benchmark.

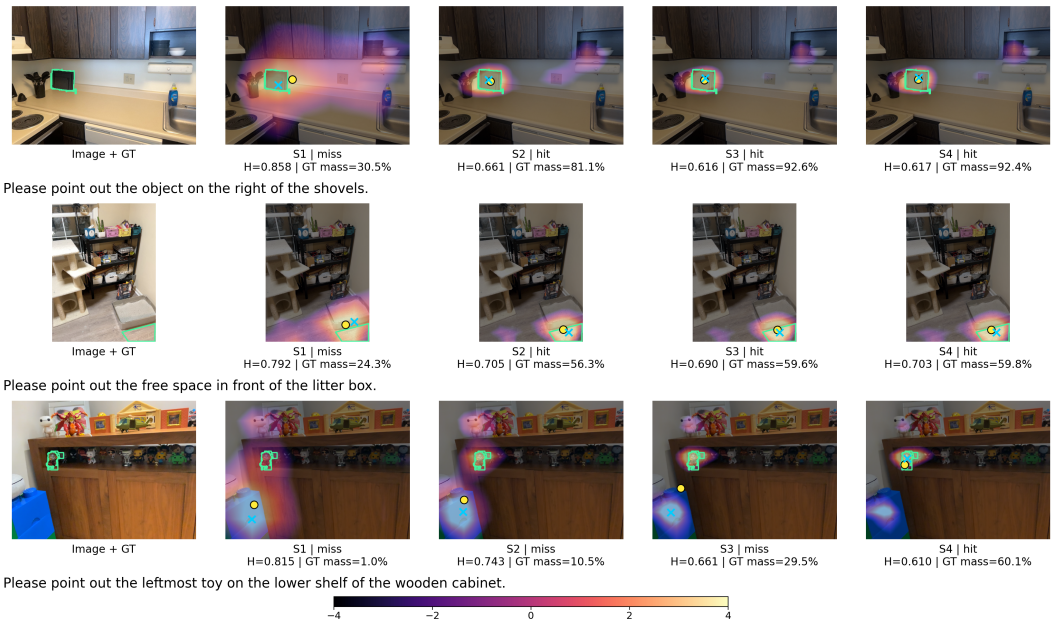



Figure 8: Four-stage distributions for three selected RefSpatial-Bench expressions. Green outlines mark target regions, yellow circles mark SoftArgmax points, and cyan crosses mark distribution maxima. Each panel reports normalized entropy and target-region probability mass on a common $224 \times 224$ grid.

Figure 9 compares four same-scene instruction pairs involving kitchen objects, bedroom pillows, table placement regions, and desk objects. The requested referent and predicted distribution change within each pair. Since the targets differ, these selected cases illustrate instruction conditioning rather than paraphrase consistency.




Figure 9: Grounding under different instructions for the same scenes. Each adjacent pair has distinct targets. Green outlines mark targets, yellow circles SoftArgmax points, and cyan crosses distribution maxima; hit/miss labels indicate point-in-target outcomes.

For real-world evaluation, we use a seven-joint Franka FR3 with a parallel gripper and a fixed third-person RGB camera. Each trial encodes its initial image and English instruction once; the features are cached during execution. The Action Head reads frozen multistage semantic and spatial features and the localization distribution from the GDE Head, together with the initial end-effector pose and gripper state. Depth is disabled, and heatmaps are not back-projected through camera calibration; action training and deployment therefore use consistent viewpoints.

A four-layer asymmetric MoT (width 128, four heads, one state token, eight action tokens) and a two-layer Transformer decoder predict 112 base-frame six-dimensional pose nodes (48 approach, 16 lift, 48 placement) and three phase durations. Each node contains XYZ position and a rotation vector relative to a fixed reference rotation. After de-normalization, durations set linear position interpolation and rotation SLERP. Placement starts after a measured 3 cm rise following gripper closure. The controller tracks the interpolated poses using measured-state feedback, bounded Cartesian increments, MoveIt inverse kinematics, and collision and joint-jump checks. It closes the gripper within 15 mm and 0.15 rad of the predicted closure pose; a release condition fitted from demonstrations determines automatic opening during placement.

The Action Head is trained from scratch on same-view simulated demonstrations with the backbone and GDE Head frozen. The 10 Hz dataset has 1,278/163/164 train/validation/test segments, split by original recording. After removing pre-action waiting frames, trajectories are aligned to gripper closure and resampled into approach, lift, and placement nodes. The objective is

$$\mathcal{L} = \mathcal{L}_{\text{pose}} + 0.2\mathcal{L}_{\text{duration}} + 0.2\mathcal{L}_{\text{difference}}, \tag{11}$$

where the difference term penalizes adjacent-node changes. Pose errors at closure and placement endpoints receive weight 8; all others receive weight 1. AdamW uses batch size 48, learning rate $3 \times 10^{-4}$, weight decay $10^{-4}$, and a 240-epoch cosine schedule with validation every 10 epochs. We deploy seed 17 at epoch 120 (2.486 million trainable parameters).

We test five conditions with 10 independent trials each: *similar objects* (e.g., pink and green bottles), *unseen instances* absent from the Action Head's simulated training data (e.g., a toy dinosaur), *absolute position* (the leftmost object), *relative position* (the object right of a red can), and *partial occlusion* (at most 50% of the target occluded in the camera view). Object poses vary across trials. Only the first grasp attempt is allowed, without retries; success requires grasping the intended object and releasing it in the designated area. Operators set the prompt and scene and initialize the robot; execution then proceeds autonomously.

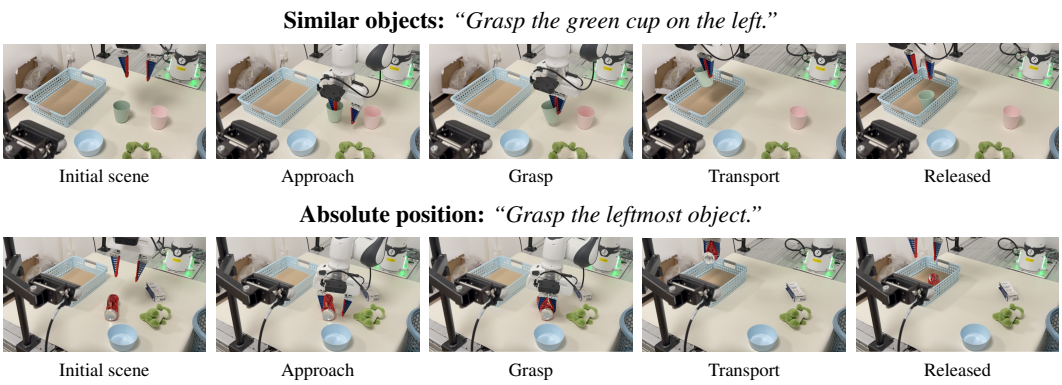



Figure 10: Two successful real-world pick-and-place sequences, ordered left to right. The green cup and the selected can are each released into the designated basket.

Qualitative inspection suggests that the intended object is usually localized correctly even with visually similar distractors or more complex spatial relations. We do not report a separate localization accuracy for the robot trials. In a small number of relative-position cases, the model instead localizes the reference object, such as the red can, rather than the target specified relative to it.

Most observed failures arise during grasp execution rather than target selection. Initial object pose strongly affects grasp success. For elongated objects, a single fixed third-person RGB view without

depth input can leave the grasp point ambiguous, causing a grasp-position error even when the target object is localized correctly. Object orientation and mild occlusion can also restrict gripper access.

## D Discussion and Limitations

RACE formulates spatial grounding as the iterative co-refinement of referent semantics and a spatial affordance state, rather than a one-way prediction from semantics to location. Stage-wise analysis shows that more accurate localization does not require the affordance distribution to become progressively sharper. Later refinement may instead redistribute spatial probability to better align with the evolving referent representation. These observations suggest that semantic understanding and spatial evidence can serve as complementary states that are jointly refined through interaction, providing a viable way to equip language-model backbones with fine-grained spatial grounding capabilities.

RACE currently uses a fixed four-stage refinement process and primarily localizes in two-dimensional image space. Future work could adapt the refinement process to state uncertainty or convergence, and extend referent–affordance co-evolution to 3D, temporal perception, and closed-loop visual robotic interaction. The current update mechanism is one realization of bidirectional refinement; alternative state representations and feedback mechanisms remain to be explored.