ALIGNING THOUGHTS WITH ANSWERS: PROBABILITY REWARDS TO TAME THINKING DRIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper studies **thinking–answer consistency** in vision-language models. We focus on Visual Intention Grounding, where a model infers a target object based on a human intention query and predicts a bounding box. We reveal that previous IoU-based reinforcement learning (RL) frameworks suffer from "thinking drift", where the model produces a correct bounding box, despite having an incorrect reasoning process pointing to a different target object. Thus, we propose **Rita** (*ReInforcing Thinking–Answer consistency*) as a novel RL paradigm to tame the drift. Specifically, Rita introduces two reasoning-label-free RL rewards, constructed from the conditional probability of reference answers: a **thinking reward** and a **consistency reward**. It also adopts a difficulty-aware **data filtering** strategy that selects informative easy-to-medium samples for RL using rollout error rate and reward variance. Extensive experiments on EgoIntention and the new RefEgo-Int benchmarks show that Rita performs consistently superior to the supervised finetuning approaches and vanilla RL-finetuned frameworks.

1 INTRODUCTION

Given an egocentric image, Visual Intention Grounding (VIG) (Sun et al., 2025) requires a vision-language model (VLM) to identify and localize an object that can satisfy a user's intention. Unlike an explicit referring expression, an intention query may describe a goal without naming the target object. For example, a request for something to hold soapy water while rinsing dishes can refer to the bucket shown in Fig. 1. Resolving such a query requires connecting the intended use with the objects and their context in the image. This capability allows visual assistants to help users find suitable objects from descriptions of their goals, without requiring explicit object names or locations.

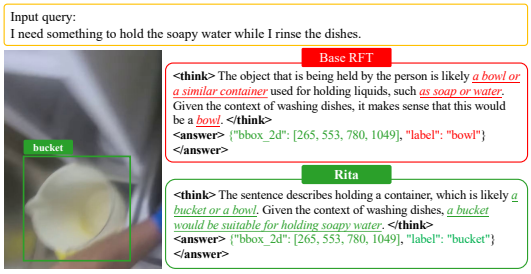


Figure 1: **Base RFT with IoU reward alone fails to enforce semantic consistency.** In the Base RFT (red), which relies solely on the IoU reward, the model predicts a reasonable bounding box but provides irrelevant reasoning, leading to a label error. In contrast, in our Rita (green), the integration of thinking and consistency rewards encourages consistency between the reasoning process and the final prediction, correcting the label while maintaining spatial accuracy.

Recent VLMs optimized via reinforcement learning (RL) (Liu et al., 2025c; Shen et al., 2025) adopt a think-then-answer mechanism, generating a reasoning trace before producing the final prediction. For VIG, this paradigm provides a natural decomposition: the model first interprets the user's intention and identifies a suitable target object in its reasoning trace, and then localizes that object in the final answer.

Existing RL-based grounding methods, such as VLM-R1 (Shen et al., 2025) and UniVG-R1 (Bai et al., 2025b), optimize the Intersection over Union (IoU) between predicted and ground-truth bounding boxes. Apart from an auxiliary reward for valid output formatting, their RL feedback evaluates only the final bounding box, without explicitly assessing consistency between the reasoning trace and the predicted target. Because the IoU reward depends only on the final bounding box,

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.

two responses receive the same reward when they predict the same box, even if their reasoning traces identify different target objects. This creates a risk of reward hacking, where the model can obtain a high reward through an inconsistent reasoning process, effectively predicting the correct box for the wrong reason. For example, in Fig. 1, with an IoU-only objective, the target object is correctly located, but the <think> trace identifies the wrong object category. We refer to this inconsistency as "thinking drift", inspired by Luo et al. (2025). Worse still, once such reward hacking occurs, RL may reinforce the entire high-reward rollout, thereby exacerbating the inconsistency. Indeed, when we train UniVG-R1 on VIG, IoU-only RL improves grounding precision, yet 15.8% of the boxes it newly corrects come with reasoning that targets a different object, $2.4\times$ the drift rate before RL (Appendix C). One remedy is to supervise the reasoning process, *e.g.*, using human-labeled rationales (Shao et al., 2024) or chains of thought (CoT) (Wei et al., 2022) distilled from stronger teacher models (Guo et al., 2025; Achiam et al., 2023), as in the cold-start stage of UniVG-R1 (Bai et al., 2025b). However, such supervision either requires costly annotation at scale (Zhao et al., 2025) or risks overfitting the model to the limited reasoning patterns of the teacher models (Chen et al., 2023b; Matos et al., 2025).

To tackle thinking drift, we propose a reasoning label-free RL paradigm **Rita** (*ReInforcing Thinking–Answer consistency*). Motivated by Welch et al. (2026), who show that a VLM's post-reasoning answer likelihood increasingly reflects its reasoning-answer consistency rather than calibrated confidence in answer correctness, we turn this consistency signal into a correctness signal by scoring the reference answer instead of the sampled one. Specifically, our **thinking reward** encourages reasoning traces which directly contribute to higher conditional likelihood to the reference answer. Additionally, because this substitution detaches the thinking reward from the generated answer, simply adding it to the IoU reward does not ensure that a reference-supporting trace and a correct prediction occur within the same rollout; a high score from one reward may compensate for a low score from the other. We therefore introduce a **consistency regularization** that explicitly couples trace and answer quality, rewarding a reference-supporting trace when the rollout's predicted box overlaps the target and penalizing it when the box largely misses the target.

Beyond improving reasoning reliability, we investigate strategic data filtering to enhance training efficiency. Some reasoning-focused RL methods prioritize challenging but solvable problems to elicit deeper reasoning (Wang et al., 2025). For grounding, however, we find it more effective to concentrate on samples of low-to-moderate difficulty. Medium-difficulty samples show the greatest variation in rewards, while samples with no successful rollout provide little contrast on average for policy updates. By filtering toward samples with informative reward variation, our strategy improves grounding performance with RL fine-tuning on 5% of the original SFT training set.

We evaluate Rita on two VIG benchmarks: EgoIntention (Sun et al., 2025) and RefEgo-Int, a new dataset curated from RefEgo (Kurita et al., 2023). On Qwen2.5-VL-3B / 7B-Instruct (Bai et al., 2025a), Rita yields consistent gains in both reasoning and grounding accuracy. Our contributions are as follows:

- We identify *thinking drift* in IoU-rewarded RL for visual intention grounding: a rollout can localize the target while its reasoning trace selects a different object, and outcome-only rewards cannot distinguish it from a consistent rollout.
- We propose Rita, which adds two reasoning-label-free rewards computed from the policy's likelihood of the reference answer given a sampled trace: a thinking reward that scores how strongly the trace supports the reference answer, and a consistency reward that couples this score with the IoU of the rollout's own prediction.
- We introduce a rollout-based data filtering strategy that trains on easy-to-medium samples with high reward variance, using 5% of the SFT training set.
- Rita improves P@0.5 over SFT Qwen2.5-VL by +4.8 (3B) and +2.1 (7B) on EgoIntention and by +10.2 and +7.5 zero-shot on RefEgo-Int, raises grounded reasoning accuracy over IoU-only RL, and outperforms UniVG-R1 in P@0.5 without distilled reasoning traces.

2 Related Work

RL for Vision Language Models. Recent industrial foundation models such as DeepSeek-R1 (Guo et al., 2025), Kimi k1.5 (Team et al., 2025), and OpenAI's o1 system (Jaech et al., 2024) have

popularized GRPO-style reinforcement learning (RL) as a scalable post-training recipe, establishing the think-then-answer paradigm and triggering a wave of RL research on vision–language models (VLMs). Most subsequent academic efforts apply RL to tasks that naturally require long-chain reasoning, including mathematical and scientific QA, visual reasoning, and spatial QA, where multi-step CoT-style solutions are essential (Wang et al., 2026b; Jiang et al., 2025; Meng et al., 2025; Sarch et al., 2025; Tan et al., 2025). Closer to our work, recent perception-centric works (*e.g.*, recognition, OCR, grounding, segmentation) show that think-then-answer VLMs can benefit from RL when the final objective is a perceptual metric (Yu et al., 2025a; Ma et al., 2025; Shen et al., 2025; Bai et al., 2025b; Liu et al., 2025b).

These methods reward only the final prediction and leave the content of the reasoning trace unconstrained. Methods that supervise the trace require additional reasoning data, such as human-annotated rationales (Shao et al., 2024) or chains of thought distilled from stronger models, as in the cold-start stage of UniVG-R1 (Bai et al., 2025b). Such data are costly to collect at scale (Zhao et al., 2025), and distilled traces can inherit the limited reasoning patterns of the teacher (Chen et al., 2023b; Matos et al., 2025). Rita instead rewards the trace using only the existing answer annotations.

Label-free RL Rewards. To extend RL beyond tasks with rule-based verifiers, recent work trains language models on unverifiable data (Wang et al., 2026a; Tang et al., 2025). Reference-likelihood rewards replace the verifier with the policy's own likelihood of the reference answer after its reasoning, and use this signal to learn correct answers in question answering (Zhou et al., 2026; Yu et al., 2025b; Liu et al., 2025a). In grounding, IoU already verifies the correctness of the answer, so Rita uses reference likelihood for a different purpose. Because answer likelihood after reasoning reflects consistency with the trace (Welch et al., 2026), it measures whether the reasoning trace supports the correct target. Rita further adds a consistency reward that ties this trace score to the model's own prediction, which reference-likelihood rewards discard.

Data Filtering in RL. Early instruction-tuning studies already demonstrate that carefully filtered, task-specialized subsets can outperform using the full corpus (Xia et al., 2024; Liu et al., 2024; Li et al., 2024b;a; Jain et al., 2024; Chen et al., 2023a; Zhou et al., 2023). Particularly, recent work, such as Vision-G1 (Zha et al., 2026), QwenLong-L1 (Wan et al., 2025), and ThinkLite-VL (Wang et al., 2025), adopt various difficulty-aware or influence-based criteria to prioritize "hard-but-solvable" or unsolved examples, under the assumption that such instances provide the strongest learning signal for reasoning tasks (*e.g.*, math, logic, long-context QA). Different from these approaches, we propose a selection scheme that favors informative samples of easy-to-medium difficulty, avoiding the noisy, uninformative gradients from persistent failures like near-zero IoU rewards and thereby performing a more stable optimization curriculum.

3 METHOD

3.1 PRELIMINARY: RL FOR VISUAL INTENTION GROUNDING

Visual Intention Grounding maps an intention sentence to the target object and its location. Given an input image I and a human intention query Q, the model infers the underlying intention and predicts an answer A that contains the bounding box (x_1, y_1, x_2, y_2) of the target object. Recent RL-finetuned reasoning models first generate a thinking trace T, *i.e.*, $(T, A) \sim \pi_\theta(\cdot \mid I, Q)$, where T describes how the model links the query Q to the predicted answer A.

GRPO for VLMs. Group Relative Policy Optimization (GRPO) is a rule-based RL algorithm for post-training LLMs/VLMs that removes the need for a learned critic by using group-relative rewards. For each input q, the old policy samples a group of N candidate outputs $\{o_i\}_{i=1}^{N}$ and each reward r_i is normalized within the group into the relative advantage $A_i = (r_i - \mu)/\sigma$. The GRPO objective is

$$\mathcal{J}_{\mathrm{GRPO}}(\theta) = \mathbb{E}_{\{o_i\}\sim\pi_{\theta_{\mathrm{old}}}(q)}\left[\frac{1}{N}\sum_{i=1}^{N}\left\{\min\big[s_1 A_i,\ s_2 A_i\big] - \beta\,\mathbb{D}_{\mathrm{KL}}\big[\pi_\theta \,\|\, \pi_{\mathrm{ref}}\big]\right\}\right], \tag{1}$$

where $s_1 = \frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\mathrm{old}}}(o_i|q)}$ and $s_2 = \mathrm{clip}\Big(\frac{\pi_\theta(o_i|q)}{\pi_{\theta_{\mathrm{old}}}(o_i|q)},\, 1-\epsilon,\, 1+\epsilon\Big)$ and β weights the KL penalty toward the reference policy π_{ref}.

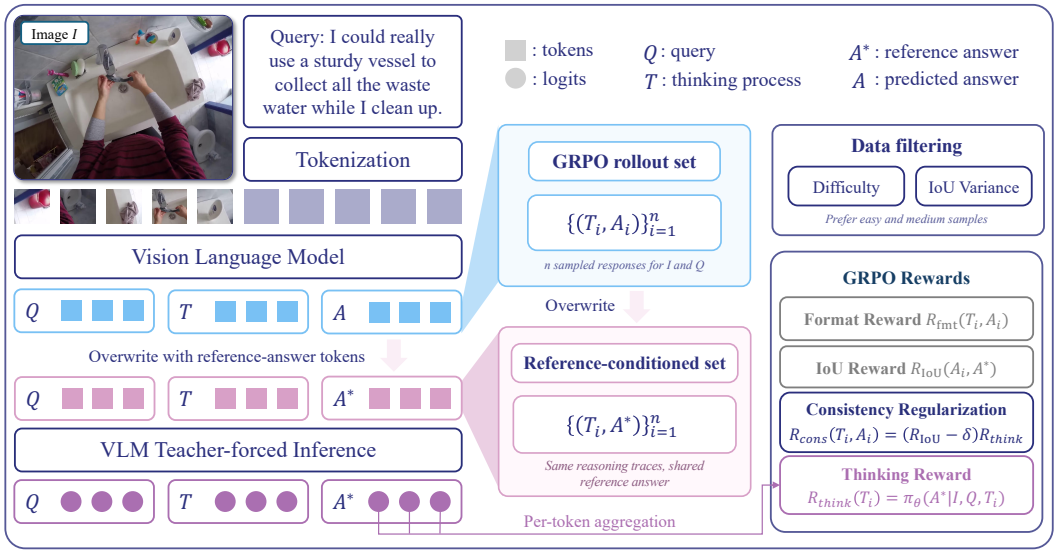


Figure 2: **Overview of the Rita framework.** Given an input image and human intention query, the VLM samples multiple think-then-answer rollouts (Q, T_k, A_k), from which we derive a grounding reward via IoU with the ground-truth box and a format reward. We then overwrite A_k with the reference answer $A^\star$ and run the VLM in teacher-forced mode to obtain logits for the answer tokens, which define the thinking reward R_{think}. The consistency reward R_{cons} scales this score by the predicted box's IoU minus a margin δ. During the GRPO-based optimization, we conduct difficulty-and-variance-aware data sampling computed from IoU rollouts to improve training efficiency.

RL Rewards. Early RL approaches (Liu et al., 2025c; Shen et al., 2025; Bai et al., 2025b) for VLM perception tasks optimize verifiable rewards aligned with the evaluation metric:

$$R(T, A \mid A^\star, \lambda_1, \lambda_2) \;=\; \lambda_1 R_{\text{IoU}}(A, A^\star) + \lambda_2 R_{\text{fmt}}(T, A). \tag{2}$$

Here, $R_{\text{IoU}} \in [0, 1]$ is the IoU of two bounding boxes, and $R_{\text{fmt}} \in \{0, 1\}$ is a binary indicator of whether (T, A) follows the required output format, *i.e.*, `<think>...</think> <answer>...</answer>`. This answer-oriented reward can produce inconsistent thinking traces and answers: the reasoning content is learned only indirectly through the IoU reward, and neither the trace nor its consistency with the answer is supervised. In our re-implementation of UniVG-R1, IoU-only RL indeed increases the share of correct boxes paired with drifted reasoning (Appendix C). To address this without reasoning labels, we directly optimize the thinking trace and thought–answer consistency with the probability-based rewards in Section 3.2.

3.2 Rita: Outcome-Conditioned Probability-based Rewards

Standard grounding rewards evaluate only the final prediction. Consequently, two rollouts receive the same IoU reward whenever they predict the same bounding box, even if one reasoning trace correctly identifies the intended object while the other refers to an unrelated target. To provide direct feedback on the generated trace without requiring reasoning annotations, Rita evaluates whether the trace supports the annotated answer.

Thinking reward. Let $x = (I, Q)$ denote an image–query pair, and let a rollout from the old policy be $o_k = (T_k, A_k)$, where T_k is the generated thinking trace and A_k is the sampled structured answer. Given the reference answer $A^\star = (a_1^\star, \ldots, a_L^\star)$, we preserve the sampled trace T_k, replace A_k with $A^\star$, and perform a teacher-forced forward pass, as illustrated in Fig. 2. We compute the length-normalized likelihood score

$$R_{\text{think}} = \frac{1}{L} \sum_{t=1}^{L} \pi_\theta \left(a_t^\star \mid x, T_k, a_{<t}^\star\right), \tag{3}$$

where π_θ is the policy snapshot used for reward computation. Intuitively, Eq. 3 averages the token-level probabilities assigned to the reference answer, conditioned on the given input (I, Q) and sampled thinking trace T_k.

Overwriting with the reference answer is important because, after a reasoning trace, the likelihood of the sampled answer A_k mainly reflects its consistency with the trace rather than its correctness (Welch et al., 2026), so it would reward confident incorrect predictions as well (see Appendix D.2 for an ablation). In contrast, Eq. 3 holds the target answer fixed across all rollouts of the same input. Differences in reward therefore reflect how strongly each sampled trace supports the annotated target. R_{think} thus provides a reasoning-label-free signal for comparing traces.

Thinking-answer consistency regularization. Although R_{think} evaluates whether T_k supports $A^\star$, it does not consider the answer A_k actually generated after that trace. Conversely, R_{IoU} evaluates the sampled answer without considering whether its preceding trace provides compatible reasoning. Simply adding these rewards therefore promotes trace quality and answer quality as two separate objectives, but does not explicitly reward their joint satisfaction within the same rollout. Consequently, a high score from one component can compensate for a low score from the other, so the sum does not address thinking drift.

To explicitly couple these two components, we introduce a thinking–answer consistency regularization. We define the box score

$$S_\delta(A_k, A^\star) = R_{\text{IoU}}(A_k, A^\star) - \delta, \tag{4}$$

where δ is an IoU margin that separates rollouts whose predicted box overlaps the target from rollouts whose box misses it. We use a small margin, $\delta = 0.1$, so that the penalty applies only to boxes with little or no overlap with the target; Appendix D.1 compares it with the standard grounding threshold $\delta = 0.5$. We define

$$R_{\text{cons}} = S_\delta(A_k, A^\star) R_{\text{think}}. \tag{5}$$

This multiplicative form couples trace quality and answer quality. When the predicted box overlaps the target by more than δ, $S_\delta > 0$, and a trace that strongly supports the reference answer receives a larger positive consistency reward. When the predicted box misses the target, $S_\delta < 0$, a high thinking score instead produces a stronger penalty, reflecting the inconsistency between the reference-supporting trace and the realized answer. If the trace itself provides little support for the reference answer, R_{cons} remains small regardless of the answer quality. Therefore, the consistency reward becomes strongly positive only when both components agree, while providing little or negative reinforcement when an inconsistency is observed.

Overall objective. For rollout k, Rita combines the proposed probability rewards with the standard grounding and formatting rewards:

$$R = \lambda_{\text{IoU}} R_{\text{IoU}} + \lambda_{\text{fmt}} R_{\text{fmt}} + \lambda_{\text{think}} R_{\text{think}} + \lambda_{\text{cons}} R_{\text{cons}}. \tag{6}$$

The four terms provide complementary supervision: R_{IoU} evaluates spatial accuracy, R_{fmt} enforces valid output structure, R_{think} evaluates whether the sampled trace supports the annotated target, and R_{cons} couples this compatibility signal to the answer produced by the rollout. All rewards are computed without reasoning labels and normalized within each rollout group by GRPO.

3.3 RITA: DATA FILTERING

While our probability-based rewards provide richer feedback for individual rollouts, effective policy optimization also requires informative training samples. Rollout groups dominated by unsuccessful grounding predictions often provide limited outcome-level contrast. We therefore introduce a difficulty- and variance-aware data selection strategy that prioritizes learnable samples with diverse grounding outcomes, improving data efficiency while complementing our reward design.

For each training sample i, we sample N rollouts and compute an IoU-based verifiable reward $r_i^{(k)} \in [0, 1]$ for rollout k, with correctness indicator $c_i^{(k)} = \mathbb{I}\Big[\text{IoU}_i^{(k)} \geq \tau\Big]$ under a fixed threshold τ. We define a sample difficulty score as the number of correct rollouts $d_i = \sum_{k=1}^{N} c_i^{(k)}$ (smaller d_i means harder), and partition the dataset into three buckets: hard $B_h = \{i \mid d_i \in [0, \gamma_h N]\}$,

medium $B_m = \{i \,|\, d_i \in (\gamma_h N, \gamma_e N)\}$, and easy $B_e = \{i \,|\, d_i \in [\gamma_e N, N]\}$, where γ_h, γ_e are hyperparameters for bucket thresholds with $0 \leq \gamma_h < \gamma_e \leq 1$.

In addition to difficulty, we quantify a sample's informativeness by the reward spread $\sigma_i = \mathrm{Std}\Big(\{r_i^{(k)}\}_{k=1}^N\Big)$, which equals the dispersion of the rollout advantages before normalization: for any baseline b_i, the unnormalized advantages $\tilde{A}_i^{(k)} = r_i^{(k)} - b_i$ satisfy $\mathrm{Std}(\{\tilde{A}_i^{(k)}\}_{k=1}^N) = \sigma_i$. Intuitively, a larger σ_i indicates a greater separation between good and bad rollouts, yielding stronger policy gradients for GRPO-style updates.

Our filtering-and-sampling scheme specifies a portion vector $\boldsymbol{\pi} = (\pi_h, \pi_m, \pi_e)$ with $\pi_h + \pi_m + \pi_e = 1$, and draws $M\pi_b$ samples from bucket B_b, where M is the size of the RL training set, according to a variance-biased distribution

$$P(i \mid B_b) = \frac{\sigma_i^\alpha}{\sum_{j \in B_b} \sigma_j^\alpha}, \quad \alpha > 0, \tag{7}$$

where α is a temperature controlling within-bucket sharpness and $b \in \{h, m, e\}$. To avoid misleading updates from unlearnable cases, we discard samples with $d_i = 0$ regardless of high σ_i; this prevents the optimizer from chasing high-variance but consistently incorrect rollouts. In practice, $\boldsymbol{\pi}$ emphasizes B_m for learnability, adds B_h for exploration and B_e for stability, and the σ_i^α weighting favors samples with larger advantage gaps.

4 Experiments

4.1 Implementation Details

Datasets. Our evaluation focuses on egocentric VIG datasets derived from head-mounted-camera imagery (Grauman et al., 2022). Such imagery presents characteristic challenges, including frequent camera motion, motion blur, and off-center target objects. Our primary benchmark is EgoIntention (Sun et al., 2025), a visual grounding benchmark pairing Ego4D images with intention sentences. It contains two splits: *context* (typical intents, *e.g.*, "I need to sit down" $\rightarrow$ chair) and *uncommon* (atypical intents, *e.g.*, "I need a boost to change the bulb" $\rightarrow$ chair). To assess generalization, we construct RefEgo-Int from RefEgo (Kurita et al., 2023) by sampling frames (one frame per video clip) and collecting context intention sentences following the EgoIntention dataset protocol. RefEgo-Int preserves RefEgo key challenges: (i) frequent viewpoint motion, (ii) no-target cases, and (iii) long-tail object categories. Refer to Appendix A for RefEgo-Int construction details.

Models. We use Qwen2.5-VL-3B-Instruct as the main backbone and report scalability with Qwen2.5-VL-7B-Instruct (Bai et al., 2025a). We follow VLM-R1 (Shen et al., 2025) for the GRPO settings, with N=8 rollouts, temperature 0.9, iterations 1, KL ratio 0.04, and learning rate 1e$-$6. All reward weights λ_i's are set to 1, and the IoU margin of the consistency reward is $\delta = 0.1$. For data filtering, we set $\tau = 0.5$, $\gamma_h = 0.2$ and $\gamma_e = 0.8$. The grounding template and thinking prompts also follow VLM-R1 (Appendix E).

For the SFT baseline, we perform full finetuning with LLaMA-Factory (Zheng et al., 2024) on the EgoIntention training set for 1 epoch, whereas for RL finetuning we apply Low-Rank Adaptation (LoRA) (Hu et al., 2022) (rank 64) for 500 update steps. We re-implement UniVG-R1 (Bai et al., 2025b) on EgoIntention with the same Qwen2.5-VL-3B-Instruct backbone. Following its two-stage recipe, we first fine-tune the backbone for one epoch on chain-of-thought traces distilled from Qwen2.5-VL-72B-Instruct (Bai et al., 2025a) (replacing the proprietary Qwen-VL-Max teacher), keeping only traces whose final answers are correct. We then run its GRPO stage with the official implementation, including its IoU and format rewards and difficulty-aware weight adjustment, under the same RL configuration as Rita.

Metrics. We evaluate our models with three metrics: *Grounding Precision* (P@0.5) measures spatial accuracy, where a predicted box is correct if its IoU with the ground truth is ≥ 0.5. *Mean Intersection over Union* (mIoU) provides a fine-grained measure of localization quality by averaging the IoU across all samples. *Thinking Accuracy* ($\mathrm{Acc}_{\mathrm{think}}$) assesses the logical correctness of the thinking process by checking if the predicted object category matches the ground-truth label. *Grounded*

Table 1: **EgoIntention results split by Context / Uncommon and Overall.** Metrics are *P@0.5* (↑) and *mIoU* (↑). Numbers in parentheses give the P@0.5 gain of Rita over the Qwen2.5-VL-Instruct baseline of the same size.

Method	Context		Uncommon		Overall	
	P@0.5	mIoU	P@0.5	mIoU	P@0.5	mIoU
Qwen-VL (Bai et al., 2023)	32.1	0.350	26.1	0.298	29.1	0.324
MiniGPT-v2 (Chen et al., 2023c)	46.0	0.429	40.9	0.373	43.4	0.401
Reason-to-Ground (Sun et al., 2025)	49.9	0.445	44.7	0.396	47.3	0.421
Qwen2.5-VL-3B-Instruct (Bai et al., 2025a)	64.8	0.581	56.1	0.483	60.4	0.532
Qwen2.5-VL-7B-Instruct (Bai et al., 2025a)	67.2	0.594	61.8	0.514	64.5	0.554
UniVG-R1 (Qwen2.5-VL-3B) (Bai et al., 2025b)	65.9	0.604	59.6	0.549	62.7	0.576
Rita-3B	67.9 (+3.1)	0.598	62.4 (+6.3)	0.518	65.2 (+4.8)	0.558
Rita-7B	68.9 (+1.7)	0.605	64.3 (+2.5)	0.527	66.6 (+2.1)	0.566

Table 2: **Effectiveness of reward components on reasoning performance.** We evaluate Acc_{think} and Acc_{align} (%) across *Context* and *Uncommon* splits. R_{IoU}, R_{think}, and R_{cons} represent grounding, thinking, and consistency rewards used during RFT. Best results are in **bold**.

RFT Rewards			Context Split (%)		Uncommon Split (%)	
R_{IoU}	R_{think}	R_{cons}	$\text{Acc}_{\text{think}}$	$\text{Acc}_{\text{align}}$	$\text{Acc}_{\text{think}}$	$\text{Acc}_{\text{align}}$
✓			83.7	61.4	71.5	54.4
✓		✓	83.1	61.2	70.9	54.0
✓	✓		85.1	62.3	73.7	55.6
✓	✓	✓	**86.6**	**63.7**	**75.0**	**56.5**

Reasoning Accuracy ($\text{Acc}_{\text{align}}$) is a holistic metric that requires a sample to have both a correct bounding box ($IoU \geq 0.5$) *and* the correct object category.

4.2 Quantitative Results

Localization Evaluation. We evaluated our method on the EgoIntention *Context* split and *Uncommon* split, reporting overall accuracy as shown in Table 1. Prior methods, Qwen-VL (Bai et al., 2023), MiniGPT-v2 (Chen et al., 2023c), and Reason-to-Ground (Sun et al., 2025), which disentangles intention reasoning from localization, achieve limited performance. In comparison, our base model Qwen2.5-VL-3B-Instruct (Bai et al., 2025a) achieves substantial improvements. Building on the supervised finetuning models, Rita-3B improves precision@0.5 by 3.1 on the context split and by 6.3 on the uncommon split. For 7B models, the context split increases by 1.7 and the uncommon split by 2.5. These results demonstrate that Rita is effective even when compared to strong supervised fine-tuning baselines at both 3B and 7B scales. Relative to these baselines, mIoU follows the same trend as precision@0.5. Rita-3B also outperforms UniVG-R1 (Bai et al., 2025b) on the same backbone by 2.0/2.8 precision@0.5 on the context/uncommon splits without teacher-distilled reasoning traces, although UniVG-R1 attains a higher mIoU.

Reasoning evaluation. Table 2 ablates the reward components used during RFT and reports reasoning quality ($\text{Acc}_{\text{think}}$) and thought–answer consistency ($\text{Acc}_{\text{align}}$) on both *Context* and *Uncommon* splits. Adding the thinking reward R_{think} yields a clear and consistent improvement over the R_{IoU}-only baseline: $\text{Acc}_{\text{think}}$ increases from 83.7% to 85.1% on *Context* and from 71.5% to 73.7% on *Uncommon*, while $\text{Acc}_{\text{align}}$ improves from 61.4% to 62.3% and from 54.4% to 55.6%, respectively. In contrast, using the consistency reward R_{cons} alone does not help and slightly degrades performance (*e.g.*, 83.7%→83.1% on $\text{Acc}_{\text{think}}$ for *Context*). We attribute this to the fact that R_{cons} is defined relative to the direction induced by R_{think}; without R_{think} participating in GRPO updates, the consistency signal cannot effectively steer the model toward better reasoning. In other words, enforcing consistency between reasoning traces and answers, without learning informative reasoning, may instead hinder optimization and degrade the answer performance. Finally, combining R_{cons} with R_{think} delivers the best results across both splits, improving $\text{Acc}_{\text{think}}$ to 86.6%/75.0%

and $\text{Acc}_{\text{align}}$ to 63.7%/56.5% on *Context*/*Uncommon*. Relative to R_{think} alone, adding R_{cons} raises $\text{Acc}_{\text{think}}$ by 1.5/1.3 and $\text{Acc}_{\text{align}}$ by 1.4/0.9 points, showing that the consistency reward complements the thinking reward even though it does not help on its own.

4.3 Data Filtering

We broadly explore multiple data sampling methods for Rita on visual intention grounding tasks. *Random.* The baseline randomly samples the RL training set. *Unsolved.* Following prior RL-for-reasoning works, we keep only samples that the SFT model fails on, these form the hard unsolved RL set. *Near threshold* (IoU $\approx$ 0.5). We select samples whose SFT IoU is near the 0.5 decision boundary, assuming they are most correctable by RL. *Normal distribution sampling.* Using our difficulty (wrong rollout count) and informativeness (reward variance) estimates, we sample with a normal distribution centered at 4 wrong rollouts (midpoint of 0–8). *Rita Strategy.* Motivated by the drop with hard–unsolved data, we downweight hard cases and exclude all-wrong samples, focusing RL on informative easy–medium instances.

As shown in Fig. 3, we first observe that directly borrowing strategies from previous RL work on reasoning tasks does not work well. Using unsolved hard samples for RL finetuning cannot even boost the final grounding precision, performance drops from 61.7 at step 0 to 60.8 at step 500. Compared with the baseline method without any data filtering strategy, we see a clear gap at step 500, dropping from 64.5 to 61.7. For training samples with IoU near 0.5, there is a consistent slight decrease across updating steps compared with baseline results. For the normal distribution approach based on the difficulty level we estimated for the training set, our strategy prefers samples with higher informativeness levels within each difficulty bin. We observe better results compared with the baseline method. At step 500, the overall grounding performance on EgoIntention outperforms the baseline method by 0.9. Finally, we exclude samples with all wrong rollouts and use a sample ratio of 1:6:3 for hard, medium, and easy level samples. This achieves the best overall performance compared with baseline.

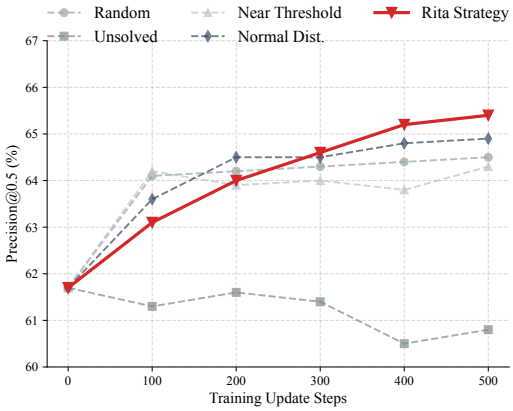


Figure 3: Grounding accuracy (Precision@0.5) on the EgoIntention dataset for five data filtering strategies across training update steps.

4.4 Qualitative Results

As shown in Fig. 4, we compare paired outputs to assess consistency between reasoning and grounding. The first three examples illustrate **misalignment** in the baseline: it correctly localizes the target but describes a different object. Specifically, it describes *tissue paper* as a *sponge*, a *tray* as a *bucket*, and *scissors* as a *knife*. Rita retains these localizations while producing descriptions that agree with the localized targets. The last example shows a grounding error accompanied by an incorrect object label: the baseline predicts a *ladder* with an incorrect bounding box, whereas Rita identifies and localizes the *stool* needed to reach the shelf.

4.5 Zero-shot Evaluation

To verify the generalization capabilities of our Rita framework, we evaluate the models on the challenging RefEgo-Int dataset in a zero-shot setting as shown in Fig. 5. Initially, the SFT baselines exhibit limited performance, with the 3B and 7B models achieving only 14.2% and 18.0% P@0.5, respectively. This can be attributed to the domain gap and inherent challenges in egocentric vision, such as motion blur, long-tail categories, and the requirement to reject queries when no target is present, a capability the original SFT models lack. However, after applying our RL framework on the EgoIntention dataset, we observe substantial zero-shot improvements. Compared to the SFT

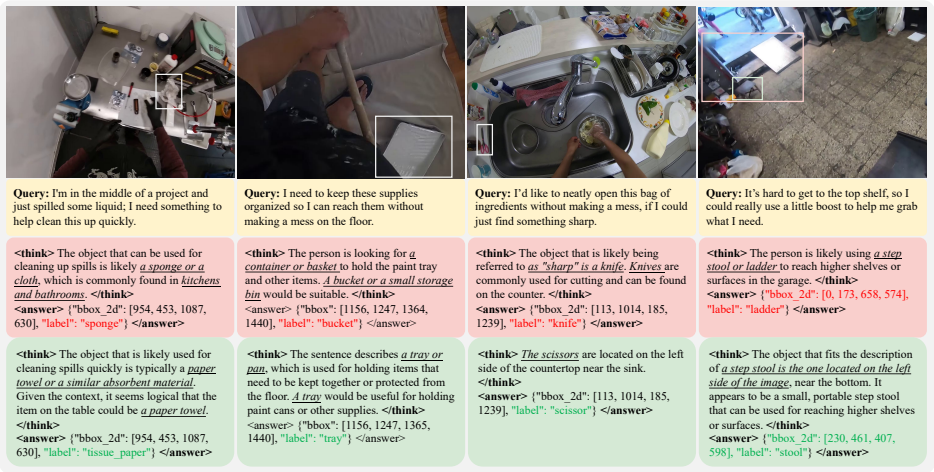


Figure 4: Qualitative comparison of the IoU-reward baseline (pink) and Rita (green), with key reasoning phrases underlined. White boxes show overlapping predictions in the first three examples; colored boxes show differing predictions in the last example.

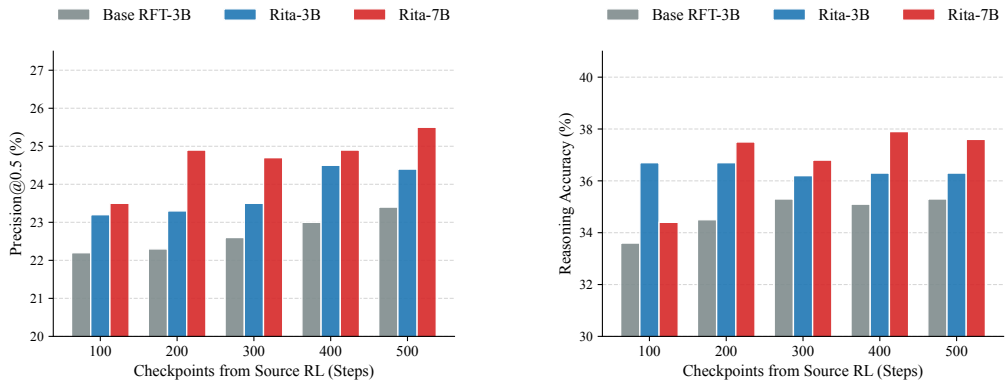


Figure 5: Zero-shot generalization across source RL training checkpoints on RefEgo-Int. We evaluate (a) **visual grounding precision** (P@0.5) and (b) **intention reasoning accuracy** using model checkpoints saved at different steps during RL training on the EgoIntention source dataset.

baselines, our RL-tuned 3B and 7B models yield gains of +10.2% and +7.5% in P@0.5 respectively. Furthermore, our method incorporating thinking and consistency rewards consistently outperforms the Base RFT model (IoU and format rewards only). At step 500, our 3B model achieves a 1.0% lead in intention reasoning accuracy over the baseline, while the 7B variant achieves a peak reasoning accuracy of 37.6%, demonstrating that our framework effectively fosters robust grounding-reasoning alignment that generalizes beyond the training distribution.

5 Conclusion

In this paper, we identify a mismatch between incorrect thinking processes and correct bounding boxes in visual intention grounding. To address this issue, we propose Rita, which reinforces reasoning traces based on their causal contribution to predicting the correct answer. Beyond the proposed probability-based rewards, we introduce a difficulty- and informativeness-aware data filtering strategy tailored to visual intention grounding, which focuses GRPO training on informative easy-to-medium samples. Extensive experiments on EgoIntention and RefEgo-Int show that our Rita framework consistently improves both grounding accuracy and intention reasoning, achieving state-of-the-art performance over strong Qwen2.5-VL baselines. We hope our framework will inspire future work toward more thought–answer consistent VLMs.

AI USAGE STATEMENT

We used AI tools to assist in drafting sections of this paper and to polish its wording, grammar, and clarity. Teacher-generated training traces used for baseline reproduction are described in Section 4.1 and Appendix C. The authors take responsibility for the final content.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025a.

Sule Bai, Mingxing Li, Yong Liu, Jing Tang, Haoji Zhang, Lei Sun, Xiangxiang Chu, and Yansong Tang. UniVG-R1: Reasoning Guided Universal Visual Grounding with Reinforcement Learning. *arXiv preprint arXiv:2505.14231*, 2025b.

Hao Chen, Yiming Zhang, Qi Zhang, Hantao Yang, Xiaomeng Hu, Xuetao Ma, Yifan Yanggong, and Junbo Zhao. Maybe Only 0.5% Data is Needed: A Preliminary Exploration of Low Training Data Instruction Tuning. *arXiv preprint arXiv:2305.09246*, 2023a.

Hongzhan Chen, Siyue Wu, Xiaojun Quan, Rui Wang, Ming Yan, and Ji Zhang. MCC-KD: Multi-CoT consistent knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 6805–6820, 2023b.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. MiniGPT-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023c.

Abhishek Dutta and Andrew Zisserman. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, 2019.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the World in 3,000 Hours of Egocentric Video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18973–18990, 2022.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645:633–638, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Naman Jain, Tianjun Zhang, Wei-Lin Chiang, Joseph E. Gonzalez, Koushik Sen, and Ion Stoica. LLM-Assisted Code Cleaning For Training Accurate Code Generators. In *International Conference on Learning Representations*, 2024.

Chaoya Jiang, Yongrui Heng, Wei Ye, Han Yang, Haiyang Xu, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. VLM-R^3: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Shuhei Kurita, Naoki Katsura, and Eri Onami. RefEgo: Referring Expression Comprehension Dataset from First-Person Perception of Ego4D. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15168–15178, 2023.

Ming Li, Yong Zhang, Shwai He, Zhitao Li, Hongyu Zhao, Jianzong Wang, Ning Cheng, and Tianyi Zhou. Superfiltering: Weak-to-Strong Data Filtering for Fast Instruction-Tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14255–14273, 2024a.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From Quantity to Quality: Boosting LLM Performance with Self-Guided Data Selection for Instruction Tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7602–7635, 2024b.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What Makes Good Data for Alignment? A Comprehensive Study of Automatic Data Selection in Instruction Tuning. In *International Conference on Learning Representations*, 2024.

Wei Liu, Siya Qi, Xinyu Wang, Chen Qian, Yali Du, and Yulan He. NOVER: Incentive Training for Language Models via Verifier-Free Reinforcement Learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025a.

Yuqi Liu, Bohao Peng, Zhisheng Zhong, Zihao Yue, Fanbin Lu, Bei Yu, and Jiaya Jia. Seg-Zero: Reasoning-chain guided segmentation via cognitive reinforcement. *arXiv preprint arXiv:2503.06520*, 2025b.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-RFT: Visual Reinforcement Fine-Tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2034–2044, 2025c.

Romy Luo, Zihui (Sherry) Xue, Alex Dimakis, and Kristen Grauman. When thinking drifts: Evidential grounding for robust video reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yan Ma, Linge Du, Xuyang Shen, Shaoxiang Chen, Pengfei Li, Qibing Ren, Lizhuang Ma, Yuchao Dai, Pengfei Liu, and Junjie Yan. One RL to See Them All: Visual Triple Unified Reinforcement Learning. *arXiv preprint arXiv:2505.18129*, 2025.

José Matos, Catarina Silva, and Hugo Gonçalo Oliveira. Cognitive flow: An LLM-automated framework for quantifying reasoning distillation. In *Proceedings of the 18th International Natural Language Generation Conference*, pp. 596–616, 2025.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Tiancheng Han, Daocheng Fu, Kaipeng Zhang, Ping Luo, Yu Qiao, Jiaheng Zhang, Michael Qizhe Shieh, Qiaosheng Zhang, and Wenqi Shao. MM-Eureka: Toward stable multimodal reasoning via rule-based reinforcement learning with policy drift control. *Transactions on Machine Learning Research*, 2025.

Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. VLM-R1: A Stable and Generalizable R1-style Large Vision-Language Model. *arXiv preprint arXiv:2504.07615*, 2025.

Pengzhan Sun, Junbin Xiao, Tze Ho Elden Tse, Yicong Li, Arjun Akula, and Angela Yao. Visual intention grounding for egocentric assistants. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2512–2522, 2025.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Xiansheng Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-RFT: Reinforcement fine-tuning for visual reasoning of vision language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yunhao Tang, Sid Wang, Lovish Madaan, and Rémi Munos. Beyond verifiable rewards: Scaling reinforcement learning in language models to unverifiable data. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1.5: Scaling reinforcement learning with LLMs. *arXiv preprint arXiv:2501.12599*, 2025.

Fanqi Wan, Weizhou Shen, Shengyi Liao, Yingcheng Shi, Chenliang Li, Ziyi Yang, Ji Zhang, Fei Huang, Jingren Zhou, and Ming Yan. QwenLong-L1: Towards Long-Context Large Reasoning Models with Reinforcement Learning. *arXiv preprint arXiv:2505.17667*, 2025.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. SoTA with Less: MCTS-Guided Sample Selection for Data-Efficient Visual Reasoning Self-Improvement. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yuanfu Wang, Zhixuan Liu, Xiangtian Li, Chaochao Lu, and Chao Yang. Native Reasoning Models: Training Language Models to Reason on Unverifiable Data. In *International Conference on Learning Representations*, 2026a.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, et al. Perception-aware policy optimization for multimodal reasoning. In *International Conference on Learning Representations*, 2026b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Robert Welch, Emir Konuk, and Kevin Smith. The cost of reasoning: Chain-of-thought induces overconfidence in vision-language models. In *European Conference on Computer Vision (ECCV)*, pp. 498–515, 2026.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting Influential Data for Targeted Instruction Tuning. In *International Conference on Machine Learning*, 2024.

En Yu, Kangheng Lin, Liang Zhao, Jisheng Yin, Yana Wei, Yuang Peng, Haoran Wei, Jianjian Sun, Chunrui Han, Zheng Ge, Xiangyu Zhang, Daxin Jiang, Jingyu Wang, and Wenbing Tao. Perception-R1: Pioneering Perception Policy with Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025a.

Tianyu Yu, Bo Ji, Shouli Wang, Shu Yao, Zefan Wang, Ganqu Cui, Lifan Yuan, Ning Ding, Yuan Yao, Zhiyuan Liu, Maosong Sun, and Tat-Seng Chua. RLPR: Extrapolating RLVR to General Domains without Verifiers. *arXiv preprint arXiv:2506.18254*, 2025b.

Yuheng Zha, Kun Zhou, Yujia Wu, Yushu Wang, Jie Feng, Zhi Xu, Shibo Hao, Zhengzhong Liu, Eric P. Xing, and Zhiting Hu. Vision-G1: Towards general reasoning vision-language models via reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28131–28139, 2026.

Kesen Zhao, Beier Zhu, Qianru Sun, and Hanwang Zhang. Unsupervised visual chain-of-thought reasoning via preference optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 2024.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. LIMA: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.

Xiangxin Zhou, Zichen Liu, Anya Sims, Haonan Wang, Tianyu Pang, Chongxuan Li, Liang Wang, Min Lin, and Chao Du. Reinforcing General Reasoning without Verifiers. In *International Conference on Learning Representations*, 2026.

This supplementary material provides the construction of the RefEgo-Int dataset (Appendix A), rollout statistics of the difficulty bins that motivate our data filtering (Appendix B), the thinking-drift analysis behind the statistics in the Introduction (Appendix C), ablations on the IoU margin of the consistency reward and on reference-answer overwrite (Appendix D), the prompt template (Appendix E), and a discussion of limitations and future work (Appendix F).

A REFEGO-INT DATASET CONSTRUCTION

RefEgo-Int is an image-based counterpart of RefEgo (Kurita et al., 2023) that both preserves the original dataset's challenge distribution and augments each sample with a human intention sentence expressing the need for the target object, hence the suffix "Int". In total, we extract 1,317 images from RefEgo to construct RefEgo-Int.

RefEgo→RefEgo-Int (one-frame-per-clip). To create an image version that faithfully reflects the video dataset's difficulty profile, we select exactly one annotated frame per clip while preserving the original $2\times2\times2$ challenge distribution over camera motion (stationary or moved), the number of referred objects (unique or multiple), and target visibility (visible or not). Concretely, we compute the annotation-level proportions of these 8 bins in RefEgo and perform stratified one-per-clip selection to match them in RefEgo-Int; when multiple frames qualify within a clip, we prefer the earliest frame. This yields an image dataset that closely mirrors the video set's challenge mix (Table 3) while retaining a single, clean frame per clip for image-only evaluation. We further annotate each RefEgo-Int sample with an intention sentence, following the annotation pipeline of EgoIntention (Sun et al., 2025). Fig. 6 illustrates the procedure using VGG Image Annotator (VIA) (Dutta & Zisserman, 2019) for checking intention sentences and collecting and correcting bounding box annotations.

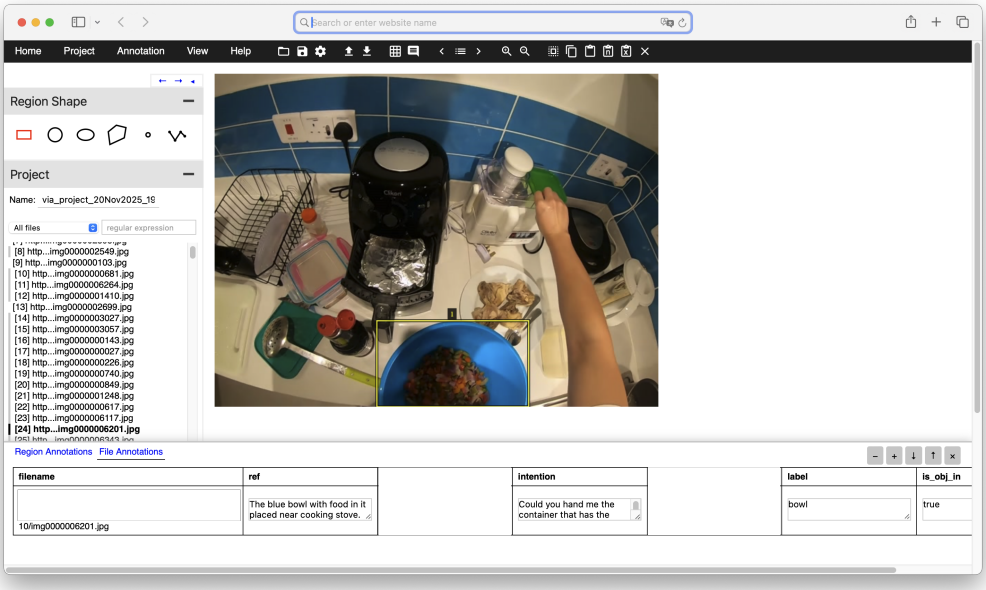


Figure 6: Screenshot of the VGG Image Annotator interface used to verify intention sentences and correct bounding boxes.

B STATISTICAL CHARACTERISTICS OF SAMPLE DIFFICULTY BINS

We analyze the rollout statistics on the context split of the EgoIntention training set to clarify why our data selection favors easy-to-medium samples for grounding RL (see Table 4). Easy samples provide clean but partially saturated supervision: they have high $\mathbb{E}[\mu_{\text{IoU}}]$ (0.851) and a very low zero-IoU rate (0.023), yet their small within-sample variance $\mathbb{E}[\sigma^2_{\text{IoU}}]$ (0.027) suggests limited outcome

Table 3: Challenge distribution of RefEgo (annotation level) and RefEgo-Int (one frame per clip) over camera motion, number of referred objects, and target visibility.

Camera	Referred objects	Target visible	RefEgo	RefEgo-Int
Moved	Multiple	No	1264 (2.00%)	27 (2.05%)
Moved	Multiple	Yes	3796 (6.01%)	85 (6.45%)
Moved	Unique	No	1290 (2.04%)	26 (1.97%)
Moved	Unique	Yes	4470 (7.08%)	79 (6.00%)
Stationary	Multiple	No	6858 (10.87%)	154 (11.69%)
Stationary	Multiple	Yes	20762 (32.89%)	489 (37.13%)
Stationary	Unique	No	6080 (9.63%)	113 (8.58%)
Stationary	Unique	Yes	18600 (29.47%)	344 (26.12%)

Table 4: **IoU rollout statistics across difficulty bins.** We bin samples by the number of correct rollouts (n_{correct} out of 8; IoU$\geq$ 0.5). For each bin, we report the number of samples, the mean n_{correct}, the mean of per-sample mean IoU, the mean of per-sample IoU variance, and the mean fraction of zero-IoU rollouts (averaged per sample).

Bin	**#Samples**	$\mathbb{E}[\boldsymbol{n}_{\text{correct}}]$	$\mathbb{E}[\boldsymbol{\mu}_{\text{IoU}}]$	$\mathbb{E}[\boldsymbol{\sigma}^{\boldsymbol{2}}_{\text{IoU}}]$	$\mathbb{E}[\text{Frac}(\text{IoU}{=}\mathbf{0})]$
Easy	9757	7.583	0.851	0.027	0.023
Medium	1988	4.118	0.479	0.133	0.240
Hard	1536	1.411	0.213	0.078	0.452
XHard	2346	0.000	0.054	0.004	0.676

diversity across rollouts, which reduces the relative comparison signal exploited by GRPO updates. Medium samples strike a better balance between correctness and informativeness, achieving non-trivial localization quality ($\mathbb{E}[\mu_{\text{IoU}}] = 0.479$) while exhibiting the largest variance (0.133), meaning that some rollouts succeed and others fail for the same input, yielding a stronger and more stable learning signal. In contrast, hard and extra-hard samples are dominated by failure rollouts: mean IoU drops to 0.213/0.054 and the fraction of zero-IoU rollouts rises to 0.452/0.676, making advantages weak or unstable; notably, XHard also has extremely low variance (0.004), indicating uniformly poor rollouts with little contrast to guide improvement. Overall, these statistics support training on easy-to-medium samples while filtering out very hard cases that mostly produce near-zero rewards. The XHard bin contains exactly the samples without any correct rollout ($d_i = 0$ in Section 3.3), which our filtering discards.

C Analysis of Thinking Drift under IoU-only RL

This section quantifies the thinking drift that IoU-only RL introduces, as reported in the Introduction. We analyze the UniVG-R1 (Bai et al., 2025b) re-implementation described in Section 4. Two details differ from the original recipe: the Qwen2.5-VL-72B-Instruct teacher is conditioned on the ground-truth box and category, and it generates a single trace per sample instead of selecting the best of three candidates. We compare the model after the chain-of-thought cold start with the model after 500 RL steps, sample by sample, on the 10,000 test samples of the *Context* and *Uncommon* splits. A sample drifts if its predicted box is correct (IoU $\geq$ 0.5) but its predicted object category differs from the ground truth. As for $\text{Acc}_{\text{think}}$, the predicted category serves as a proxy for the target of the reasoning; we do not judge the `<think>` text itself.

As shown in Table 5, IoU-only RL raises P@0.5 by 1.4 points, while overall drift rises from 4.01% to 4.98% of samples. The paired analysis locates this increase: RL newly corrects 539 boxes (and breaks 402), and 85 of the 539 newly corrected boxes (15.8%) come with a wrong category. This rate is 2.4$\times$ the drift rate among correct boxes before RL (6.54%), so the grounding gains of RL are disproportionately unsupported by a correct target. The increase is statistically significant: 229 samples drift only after RL and 132 only before RL (exact McNemar test, $p = 3.7 \times 10^{-7}$). Because it is measured after the chain-of-thought cold start, it also shows that supervised reasoning traces do not prevent drift from emerging during RL.

Table 5: **Thinking drift of UniVG-R1 before and after IoU-only RL** on EgoIntention (5,000 test samples per split). Drift is the percentage of samples whose box is correct (IoU ≥ 0.5) but whose predicted category is wrong; the last row normalizes drift by the number of correct boxes.

	Before RL (CoT cold start)	After IoU-only RL (step 500)
P@0.5 (%), Context / Uncommon / Overall	64.90 / 57.82 / 61.36	65.86 / 59.60 / 62.73
Drift (% of samples), Context / Uncommon / Overall	3.54 / 4.48 / 4.01	5.04 / 4.92 / 4.98
Drift among correct boxes (%), Overall	6.54	7.94

D Additional Ablation Study

D.1 IoU Margin of the Consistency Reward

Table 6: **Ablation on the IoU margin δ of the consistency reward.** We report P@0.5 / $\text{Acc}_{\text{think}}$ / $\text{Acc}_{\text{align}}$ (%) on the EgoIntention *Context* and *Uncommon* splits at different RL update steps. Both runs use Qwen2.5-VL-3B-Instruct and the full Rita reward; only δ differs. The main paper uses $\delta = 0.1$.

Step	Context		Uncommon	
	$\delta = 0.1$	$\delta = 0.5$	$\delta = 0.1$	$\delta = 0.5$
100	66.3 / 83.8 / 61.2	66.0 / 84.6 / 61.5	60.0 / 71.1 / 53.1	60.9 / 73.3 / 54.9
200	66.6 / 84.1 / 61.6	66.6 / 84.4 / 61.7	61.2 / 73.3 / 55.2	60.5 / 72.7 / 54.3
300	66.8 / 84.0 / 61.7	67.4 / 86.2 / 62.8	61.2 / 72.7 / 54.8	62.2 / 74.2 / 55.7
400	67.6 / 86.7 / 63.6	67.4 / 85.7 / 62.5	62.4 / 75.2 / 56.8	62.2 / 74.1 / 55.9
500	**67.9 / 86.6 / 63.7**	67.4 / 85.6 / 62.6	**62.4 / 75.0 / 56.5**	**62.4** / 74.5 / 56.0

The consistency reward in Eq. 5 rewards a rollout whose predicted box overlaps the target by more than δ and penalizes it otherwise, in proportion to its thinking score. Table 6 compares our default margin $\delta = 0.1$ with the standard grounding threshold $\delta = 0.5$. With $\delta = 0.5$, the model is ahead on most metrics at steps 100 and 300. With $\delta = 0.1$, the model continues to improve and reaches the best final results: at step 500 it outperforms $\delta = 0.5$ on all three metrics of the *Context* split (+0.5 P@0.5, +1.0 $\text{Acc}_{\text{think}}$, +1.1 $\text{Acc}_{\text{align}}$) and on both reasoning metrics of the *Uncommon* split (+0.5 each), with equal P@0.5. A plausible explanation is that $\delta = 0.5$ also penalizes rollouts whose boxes partially overlap the target, even when their traces point to the correct object, whereas $\delta = 0.1$ restricts the penalty to boxes that largely miss the target, which better matches the thinking-drift failure the consistency reward is designed to suppress.

D.2 Necessity of Reference-Answer Overwrite

Table 7: **Ablation on the effect of reference-answer overwrite for thinking reward.** We report $\text{Acc}_{\text{think}}$ (%) on the EgoIntention *Context* split at different RL update steps. *Base RFT* uses only the IoU and format rewards. Without overwrite, the thinking reward does not improve over Base RFT.

Method	100	200	300	400	500
Base RFT ($R_{\text{IoU}} + R_{\text{fmt}}$)	82.2	81.0	81.9	83.0	83.7
+ R_{think} w/o reference-answer overwrite	83.7	83.1	83.7	82.2	83.1
+ R_{think} w/ reference-answer overwrite	83.8	83.6	83.7	84.2	**85.1**

We compare two variants of the thinking reward: one scores the model's own sampled answer (*w/o reference-answer overwrite*), and the other replaces the answer tokens with the reference answer $A^\star$ as in Eq. 3 (*w/ reference-answer overwrite*). As shown in Table 7, at step 500 scoring the sampled answer lowers $\text{Acc}_{\text{think}}$ by 0.6 points relative to Base RFT, whereas scoring the reference answer raises it by 1.4 points, a 2.0-point gain over the variant without overwrite. This matches

the explanation in Section 3.2: after a reasoning trace, the likelihood of the model's own answer mainly reflects its consistency with the trace rather than its correctness (Welch et al., 2026), so it also rewards confident incorrect predictions, whereas a fixed reference answer ties the reward to whether the trace supports the correct target.

E PROMPT TEMPLATE

Our RL training and evaluation use the referring-expression prompt of VLM-R1 (Shen et al., 2025) without modification. The system turn is the Qwen default, and the intention sentence is inserted verbatim.

System: `You are a helpful assistant.`
User: `<image> Please provide the bounding box coordinate of the region this sentence describes: {intention sentence} First output the thinking process in <think> </think> tags and then output the final answer in <answer> </answer> tags. Output the final answer in JSON format.`

The prompt does not request an object label. The label in the answer is inherited from the supervised finetuning stage, whose targets follow the native Qwen2.5-VL grounding format, `[{"bbox_2d": [x1, y1, x2, y2], "label": "telephone"}]`, inside a fenced JSON block. $\text{Acc}_{\text{think}}$ and $\text{Acc}_{\text{align}}$ are scored on this label.

F LIMITATIONS AND FUTURE WORK

Limitations. While *Rita* encourages thought–answer consistency without reasoning labels, it introduces additional training overhead. Specifically, calculating R_{think} requires a secondary forward pass where predicted answers are overwritten with reference-answer tokens ($A^{\star}$) to extract logits. This reprocessing step increases computational cost and memory usage compared to standard RL pipelines.

Future Work. We aim to extend our reward into a multimodal version that explicitly verifies visual grounding. Current rewards implicitly consider images through joint embeddings, but do not ensure the thinking trace actually leverages specific scene cues. To achieve a vision-dependent reward, we propose a thinking-guided selection mechanism: conditioning the reward on an attended or masked version of the image derived from the reasoning content. By filtering visual features based on the objects or relations mentioned in the trace, the reward would better reflect if the model's logic is grounded in actual visual evidence rather than dataset biases, further enhancing robustness and interpretability.