Improving Thinking Process in Visual Grounding via Free Thinking Rewards

Anonymous CVPR submission

Abstract

We study egocentric visual intention grounding, where an assistant must infer and localize the object implied by a first-person view and an intention sentence without explicit object naming. Existing approaches either use a two-stage reasoning–then-grounding pipeline or apply reinforcement learning (RL) to train think-then-answer VLMs, but they optimize only IoU-based rewards, which often leads to reward hacking, improving box accuracy while neglecting reasoning quality. We introduce a label-free thinking-process reward that needs no human Chain-of-Thought labels or teacher models; it scores each sampled reasoning trace by how much it increases the likelihood of the correct answer, thus favoring reasoning that truly supports prediction. We also propose a data filtering strategy that selects informative easy-to-medium samples for RL using rollout error rate and reward variance. Together, these form a general recipe for process-aware RL finetuning of vision-language assistants for egocentric intention grounding. Our method achieves new state-of-the-art results on EgoIntention, boosting Precision@0.5 by +3.2 (3B) and +2.1 (7B) over strong Qwen2.5-VL baselines, and generalizes zero-shot to RefEgo-Int with +10.2 (3B) and +7.5 (7B).

1. Introduction

Recent reinforcement learning (RL) algorithms [8, 10, 20, 26, 41] such as Group Relative Policy Optimization (GRPO [8]) enable think-then-answer vision-language models (VLMs), where the model first produces an explicit reasoning trace and then commits to a final answer. For AI assistance applications [27, 32, 37], this capability allows VLMs to serve as a bridge from ambiguous human requests to downstream embodied agents' with actionable steps. Given a natural-language request, the system must first infer what the user actually needs, then determine which object in the scene should be acted on and where it is, and finally decompose the goal into actionable substeps for interacting with the environment. For example, when shown a cluttered kitchen and told "I need something to wipe the counter," a useful system should infer the towel as the intended object, localize it in the scene, and only then plan how to pick it up. In this paper, we focus on this crucial upstream stage, which is visual intention grounding task [23, 27, 31]: given an intention sentence and a scene,

Figure 1. An example of a correct bounding box with incorrect reasoning. VLM trained with IoU reward only (colored in red) mistakenly identifies the food packaging bag in the user's hand as a handbag. The bounding box in the final answer correctly refers to the carrier paper box, but due to the flawed reasoning, the box is mislabeled as a handbag. The model with our free thinking reward corrects this mismatch (colored in green).

the model reasons about the intended target and localizes it as a foundation for subsequent action.

In visual grounding and related perception tasks [4, 13, 14, 17], RL is typically plugged in as a post-training stage that optimizes a verifiable reward, most often derived from the overlap (Intersection over Union, IoU) between the predicted and ground-truth bounding boxes. This setup is effective because it directly aligns training with the evaluation metric, but the policy is only told whether the final box is good, not whether the intermediate reasoning was helpful or misleading. Recent think-then-answer approaches [4] try to address this by supervising the reasoning process itself, for example, with human-written rationales [25] or chains of thought (CoT) [35] distilled from stronger teacher models [1, 8]; however, such labels are expensive to collect at scale, and can overfit models to narrow reasoning templates.

Our experiments further show that naïvely combining GRPO with IoU rewards only can lead to reward hacking [26], where the model improves box accuracy while its internal reasoning drifts away from semantically faithful explanations, as shown in Figure 1. Therefore, these observations motivate our goal: to use RL to shape both the final outcome and the intermediate reasoning without any extra reasoning labels, yielding a label-free, process-aware training signal. For each input, we generate multiple rollouts and assess how strongly each rollout supports the ground-truth answer by inspecting the model's token-level likelihoods conditioned on that rollout (detailed in Sec. 3.2). We then aggregate these signals to reward thoughts that raise probability on the ground truth answer tokens. This process-aware signal complements IoU reward, mitigates reward hacking that overfits to final boxes.

Beyond making the model's thinking more reliable, we also ask how to use RL under realistic compute budgets. Because each training step requires multiple rollouts per sample, naïvely running RL on the full supervised finetuning (SFT) corpus would multiply computation cost by the number of rollouts. We therefore adopt a static data filtering strategy that prunes the RL set before training, using the current model's rollout statistics to estimate sample difficulty *e.g.*, by rollout error rate and informativeness level *e.g.*, by the variance of rollout rewards across attempts. In contrast to prior data-selection practices [30, 39, 40] in reasoning tasks [21, 22, 24] that favor hard unsolved cases [33], we find that focusing RL on informative samples of easy-to-medium difficulty stabilizes optimization and improves generalization. In practice, we perform RL on only a small fraction (about 5%) of the supervised fine-tuning samples while still achieving clear gains over the SFT baseline.

Bringing these two components together, we obtain a general training framework for visual intention grounding. We evaluate this framework on benchmarks such as EgoIntention [27] and RefEgo-Int [14] using mainstream open-source backbones Qwen2.5-VL-3B/7B-Instruct [3]. Our method yields consistent gains in both reasoning accuracy and grounding accuracy: the former measures how often the model correctly predicts the intended target object category, while the latter is reported as Precision@0.5 based on IoU between predicted and ground-truth boxes, improving by 1.2 and 3.2 points respectively. More generally, the proposed label-free thinking reward and data filtering are generic and can be applied to other vision-language tasks that emphasize high-quality thinking processes.

Our contributions can be summarized as follow:

1. We develop a label-free reward that encourages high-quality intermediate reasoning without requiring human annotations of thought traces or distillation from larger models.
2. We propose a data filtering strategy that jointly considers sample difficulty and informativeness, and show that informative easy–medium samples benefits visual intention grounding, in contrast to prior reasoning tasks that favor unsolved hard examples.
3. We provide a training recipe for reliable and effective visual intention grounding, achieving consistent gains over Qwen2.5-VL-3B/7B baselines on the EgoIntention and RefEgo-Int benchmarks.

2. Related Work

2.1. RL for Vision Language Models

Recent industrial foundation models such as DeepSeek-R1 [8], Kimi k1.5 [29], and OpenAI's o1 system [10] 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 [12, 22, 24, 28, 34].

In contrast, RL on perception-centric tasks (*e.g.*, recognition, OCR, and visual grounding) remains comparatively under-explored, though recent works demonstrate that think-then-answer VLMs still benefit from RL even when the final objective is a perceptual metric [4, 21, 26, 38]. However, these methods typically treat the reasoning trace as a hidden byproduct and supervise only the final prediction, leaving the content and faithfulness of the thinking process unconstrained. Explicitly labeling human chains of thought is costly and can overfit models to narrow, human-written templates, while distilling thoughts from stronger models risks propagating teacher biases and harming generalization. Our work instead keeps the think-then-answer framework but introduces a label-free thinking reward that directly evaluates how much a rollout supports the ground-

truth answer, thereby regularizing the reasoning process itself for perception-heavy tasks such as visual intention grounding.

2.2. Data Filtering for RL

Early instruction-tuning studies [5, 11, 15, 16, 19, 36, 44] already demonstrate that carefully filtered, task-specialized subsets can outperform using the full corpus: LTD Instruction Tuning [5], DEITA [19], and IFD-style selection [43] all prune redundant or low-quality instructions and retain only a small fraction of the original data while matching or surpassing larger baselines [5, 16, 19]. As RL becomes the final post-training stage for large (vision-)language models, analogous data-filtering and curriculum strategies have been explored: Vision-G1 [39], QwenLong-L1 [30], and ThinkLite-VL [33] 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) [30, 33, 39].

Our experiments reveal that aggressively emphasizing hard, frequently unsolved samples can be harmful here, as persistent failure yields near-zero IoU rewards and noisy, uninformative gradients. Instead, we propose a selection scheme that favors informative samples of easy-to-medium difficulty, leading to more stable optimization and more effective RL fine-tuning for visual intention grounding than the "hard-unsolved-first" heuristics prevalent in reasoning-focused RL works.

3. Method

In this section, we present a practical recipe for training a reliable and effective visual grounding model. Section 3.1 introduces the visual intention grounding task and reviews GRPO, an RL fine-tuning method for vision–language models. Section 3.2 details our core contribution, the free thinking reward, which scores each sampled reasoning trace by the increase it induces in the likelihood of the ground-truth answer. Finally, Section 3.3 describes our data-filtering strategy: we estimate per-sample difficulty and informativeness, and, by performing RL updates on only 5% of the supervised fine-tuning data, we enable the model to handle visual intention grounding effectively.

3.1. Preliminary

3.1.1. Visual Intention Grounding

For AI assistant, VLMs bridge ambiguous human requests to embodied action; the first step is Visual Intention Grounding [23, 27], which maps an intention sentence to the specific target object and its location. Formally, given an input image I and a human intention query Q, the model must infer the underlying intent and answer the target object and its location by predicting its bounding box (x_1, y_1, x_2, y_2) in answer tokens A. Beyond predicting the box, recent RL-finetuned reasoning models additionally produce a thinking trace T that explains how the VLM links the intention query Q to the target object within answer A. Concretely, T should give an accurate justification that links the intended action to the affordances and visible attributes of target object.

3.1.2. GRPO for Vision Language Models

Compared with traditional visual grounding, Visual Intention Grounding (VIG) first requires the model to parse the user's intention and reason about the object that satisfies it. Earlier approaches [1, 18, 27, 31] often adopt a two-stage pipeline, infer the target object name and then localize it. With the recent adoption of GRPO-style reinforcement learning, VLMs can think before answering, allowing the model to transform an implicit intention into an explicit target and its location in a single pass.

Training objective. We fine-tune the VLM with Group Relative Policy Optimization (GRPO) to obtain a reliable and effective AI assistant for intention grounding. 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 N candidate outputs $\{o_i\}_{i=1}^{N}$. Their rewards $\{r_i\}_{i=1}^{N}$ are normalized to form advantages

$$A_i = \frac{r_i - \text{mean}\{r_1, \ldots, r_N\}}{\text{std}\{r_1, \ldots, r_N\}}, \tag{1}$$

which measure each candidate's quality relative to its siblings. The GRPO objective is

$$\begin{aligned}
\mathcal{J}_{\text{GRPO}}(\theta) &= \mathbb{E}_{\{o_i\}\sim\pi_{\theta_{\text{old}}}(q)}\Bigg[\frac{1}{N}\sum_{i=1}^{N}\Big\{\min\big[s_1\cdot A_i,\ s_2\cdot A_i\big] \\
&\quad -\beta\,\mathbb{D}_{\text{KL}}\big[\pi_\theta \,\|\, \pi_{\text{ref}}\big]\Big\}\Bigg], \\
s_1 &= \frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)}, \\
s_2 &= \text{clip}\left(\frac{\pi_\theta(o_i \mid q)}{\pi_{\theta_{\text{old}}}(o_i \mid q)},\, 1-\epsilon,\, 1+\epsilon\right).
\end{aligned} \tag{2}$$

Rewards. We next detail the reward design for visual intention grounding task. Following prior RL-for-reasoning setups, we use a verifiable IoU accuracy reward and a format reward that enforces valid output structure. Beyond these, we introduce a new free thinking reward that improves intention reasoning without any reasoning trace annotations, this component is described in Section 3.2.

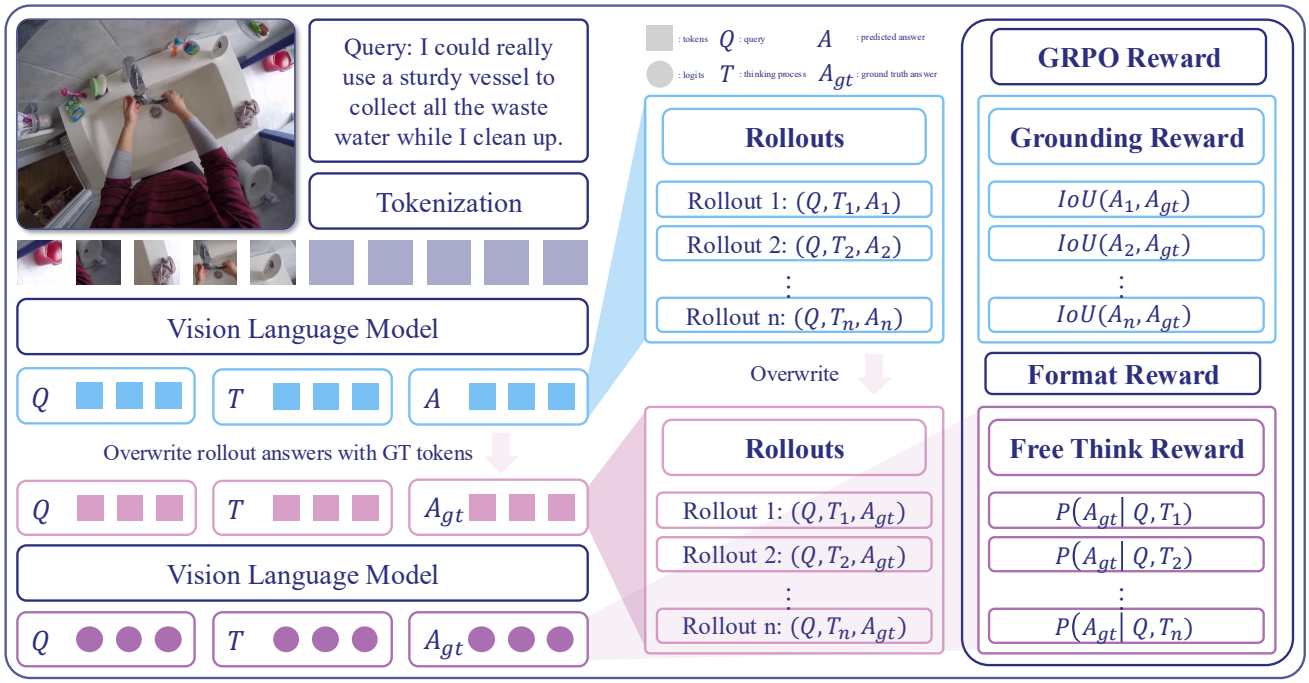

Figure 2. Overview of free thinking reward computation. Given an input image and human intention query, the VLM samples multiple think-then-answer rollouts (Q, T_k, A_k), from which we derive a grounding reward via IoU with the ground-truth box and a format reward. We then overwrite A_k with the ground-truth answer A_{gt} and run the VLM in teacher-forced mode to obtain logits for the answer tokens, which define the free thinking reward $P(A_{\text{gt}} \mid Q, T_k)$. The grounding, free thinking, and format rewards are combined into the GRPO training signal.

IoU reward. We directly optimize spatial overlap:

$$R_{\text{iou}}(I, Q, T) = \text{IoU}\big(\hat{b}, b^{\star}\big) \in [0, 1], \tag{3}$$

where $\hat{b}$ is the predicted box extracted from answer A, $b^{\star}$ is the ground truth bounding box.

Format reward. We require a well-formed output containing a `<think>...</think>` trace T and a JSON box in `<answer>...</answer>`. Denote by $\mathcal{F}$ the set of outputs that satisfy the schema and yield a valid $\hat{b}$; then

$$R_{\text{fmt}}(I, Q, T) = \mathbb{I}\{(T, \hat{b}) \in \mathcal{F}\}. \tag{4}$$

Overall reward. Training uses a weighted sum

$$R(I, Q, T) = \alpha\, R_{\text{iou}} + \beta\, R_{\text{fmt}} + \gamma\, R_{\text{think}}, \tag{5}$$

where R_{think} is our label-free thinking reward (Sec. 3.2) that evaluates the usefulness of the trace T for predicting the correct box, and $\alpha, \beta, \gamma \geq 0$ are fixed weights, all set as 1 in our framework.

3.2. Free Thinking Reward for VIG

Early RL approaches for VLM perception tasks predominantly optimize verifiable rewards that directly align with evaluation metrics (*e.g.*, IoU-based grounding accuracy), which avoids relying on learned reward models but still leaves the model at risk of reward hacking, where the thinking process becomes mismatched with the final answer. However, the content of the reasoning is critical: given an implicit intention query, both the inferred object semantics and the predicted box must be intention consistent, a correct box paired with a wrong object interpretation is unreliable. We therefore introduce a free thinking reward that improves the internal reasoning without any thinking trace annotations. It scores each sampled reasoning trace based on how much it raises the likelihood of the ground-truth answer, thereby rewarding reasoning paths that strengthen correct predictions. Given the input image I and intention query Q, the VLM first samples a trace T_{pred} and then an answer sequence A_{pred} that decodes to a box (x_1, y_1, x_2, y_2). Under supervised finetuning setting, the standard answer cross-entropy is

$$\mathcal{L}_{\theta}(A \mid Q) = -\frac{1}{N} \sum_{i=1}^{N} \log P_{\theta}\big(w_i^A \mid Q,\, w_1^A, \ldots, w_{i-1}^A\big)\,, \tag{6}$$

where $A = (w_1^A, \ldots, w_N^A)$ is the ground-truth answer sequence of length N, w_i^A denotes its i-th token, and $P_{\theta}(\cdot)$ is the VLM's conditional next-token distribution parame-

terized by θ given the query Q and previously generated answer tokens. While with RL-style generation the answer tokens are conditioned on the sampled thinking trace,

$$P_\theta\Big(w_i^{A_{\text{pred}}} \mid Q,\, T_{\text{pred}},\, w_1^{A_{\text{pred}}}, \ldots, w_{i-1}^{A_{\text{pred}}}\Big)\,. \tag{7}$$

Hypothesis. *For a given sample, better (more intention-faithful) traces make it easier for the model to produce the correct answer.*

Therefore, as shown in Figure 2, the likelihood of A_{GT} conditioned on the sampled trace can serve as an implicit measure of trace quality, without any thinking process labels. Thus, we evaluate the teacher-forced likelihood of the ground-truth answer A_{GT} conditioned on the sampled trace and define

$$\mathcal{L}_\theta\left(A_{\text{GT}} \mid Q, T_{\text{pred}}\right) = -\frac{1}{N}\sum_{i=1}^{N} \log P_\theta\Big(w_i^{A_{\text{GT}}} \mid Q,\, T_{\text{pred}},\, w_1^{A_{\text{GT}}}, \ldots, w_{i-1}^{A_{\text{GT}}}\Big)\,, \tag{8}$$

which we convert to a scalar reward via $r_{\text{think}} = -\mathcal{L}_\theta(A_{\text{GT}} \mid Q, T_{\text{pred}})$. Practically, we form the token sequence $(Q_{\text{tok}}, T_{\text{tok}}, A_{\text{GT,tok}})$ by replacing the generated answer with ground-truth tokens and perform a teacher-forced forward pass to read logits; because T_{pred} is the only varying factor per input, (8) acts as an implicit supervision signal on the thinking process. Finally, we combine r_{think} with verifiable signals, which are IoU accuracy and output-format validity, and optimize the policy using GRPO with group-relative advantages, preferring traces that raises the likelihood of ground truth answer without any thinking process annotations.

3.3. Difficulty- and Variance-Aware Data Filtering

To efficiently train a VLM for egocentric visual intention grounding, we identify training signals that are both learnable for the model and informative for policy improvement. Concretely, we employ a think–then–answer VLM and configure its exploration hyperparameters to stochastically sample multiple reasoning rollouts for each training sample before RL finetuning.

Specifically, for each training sample i we execute R=8 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 $d_i = \sum_{k=1}^{R} c_i^{(k)}$ (smaller d_i means harder), and partition the dataset into three buckets: hard $B_h = \{i \,|\, d_i \in \{0,1\}\}$, medium $B_m = \{i \,|\, d_i \in \{2,3,4,5,6\}\}$, and easy $B_e = \{i \,|\, d_i \in \{7,8\}\}$. In addition to difficulty, we quantify a sample's informativeness by the reward spread $\sigma_i = \text{Std}\Big(\{r_i^{(k)}\}_{k=1}^{R}\Big)$, which upper-bounds the per-sample advantage dispersion since for any baseline b_i the rollout advantages $A_i^{(k)} = r_i^{(k)} - b_i$ satisfy $\text{Std}(\{A_i^{(k)}\}) = \text{Std}(\{r_i^{(k)}\}) = \sigma_i$. Intuitively, 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 draws a portion vector $\boldsymbol{\pi} = (\pi_h, \pi_m, \pi_e)$ with $\pi_h + \pi_m + \pi_e = 1$, and samples $N\pi_b$ items from bucket B_b according to a variance-biased distribution

$$P(i \mid B_b) = \frac{\sigma_i^\alpha}{\sum_{j \in B_b} \sigma_j^\alpha}, \quad \alpha > 0, \tag{9}$$

where α is a temperature controlling within-bucket sharpness. To avoid misleading updates from unlearnable cases, we discard degenerate samples with $d_i = 0$ regardless of high σ_i; this prevents the optimizer from chasing high-variance but consistently incorrect rollouts. In practice, a balanced choice of $\boldsymbol{\pi}$ emphasizes B_m for learnability, supplements with B_h for exploration and robustness, and includes B_e for stabilization, while the within-bucket σ_i^α weighting prioritizes samples whose rollouts provide larger advantage gaps and thus more effective learning signals.

4. Experiments

4.1. Implementation Details

Datasets. Considering the most practical application scenario, for example headmount camera, or smart glasses, we focus on egocentric vision datasets [7]. We evaluate the effectiveness of our frameworks on EgoIntention [27], a visual grounding benchmark pairing Ego4D images with intention sentences. It contains two splits: *context* (typical intents, e.g., "I need to sit down" → chair) and *uncommon* (atypical intents, *e.g.*, "I need a boost to change the bulb" → chair). To assess generalization, we construct RefEgo-Int from RefEgo [14] 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. These challenges effectively extend the diversity of our testing data, which in turn better assess model generalizability.

Models. Due to the strong vision-language understanding performance of Qwen2.5VL models [3], we use Qwen2.5-VL-3B-Instruct as the main backbone and report scalability with Qwen2.5-VL-7B-Instruct. GRPO settings follow VLM-R1 [26]: N=8 rollouts, temperature 0.9, iterations 1, KL ratio 0.04, and learning rate 1e−6. The grounding template and thinking prompts also match VLM-R1. For the SFT baseline, we perform full finetuning with LLaMA-Factory [42] on the EgoIntention training set for 1 epoch, whereas for RL finetuning we apply Low-Rank Adaptation (LoRA) [9] for 500 update steps.

Metrics. We report *reasoning accuracy* (correct target object category given image and intention) and *grounding*

Table 1. EgoIntention results split by Context / Uncommon and Overall. Metrics are *P@0.5* (↑) and *mIoU* (↑).

Method	Context		Uncommon		Overall	
	P@0.5	mIoU	P@0.5	mIoU	P@0.5	mIoU
Qwen-VL [2]	32.1	0.350	26.1	0.298	29.1	0.324
MiniGPT-v2 [6]	46.0	0.429	40.9	0.373	43.4	0.401
Qwen2.5-VL-3B-Instruct [3]	64.8	0.581	56.1	0.483	60.4	0.532
Ours-3B	**66.6**	**0.594**	**60.7**	**0.507**	**63.6**	**0.551**
Qwen2.5-VL-7B-Instruct [3]	67.2	0.594	61.8	0.514	64.5	0.554
Ours-7B	**68.9**	**0.605**	**64.3**	**0.527**	**66.6**	**0.566**

accuracy as Precision@0.5 using IoU between predicted and ground-truth boxes.

4.2. Main Results

We evaluated our method on EgoIntention context split and uncommon split, reporting overall accuracy as shown in Table 1. Compared with previous methods Qwen-VL and MiniGPT-v2, our base model Qwen2.5-VL-3B-Instruct achieves substantial improvements. Building on supervised finetuning models, for the 3B model, our method grounding precision on the context split increases by 1.8 points, while the uncommon split improves by 4.6 points. For 7B models, the context split increases by 1.7 points and the uncommon split by 2.5 points. These results demonstrate that our framework is effective even when compared to strong supervised fine-tuning baselines at both 3B and 7B scales. The mIoU metric follows the same trend observed in precision@0.5.

Free thinking reward. Here we use the 3B model with randomly selected (unfiltered) data for RL tuning to illustrate the effectiveness of our thinking reward. As shown in Figure 3: First, RL finetuning with only the IoU reward causes a sharp early drop in performance due to reward hacking—object reasoning accuracy falls to 82.5% at step 200. Second, adding our free-thinking reward stabilizes object reasoning accuracy at the start of RL tuning. Finally, the combined IoU + thinking reward yields consistent gains throughout training, reaching 86.4% at step 500, a 1.2% improvement over the IoU-only model.

4.3. Ablation Study

Does a logit-space reward alone help? To understand whether a reward defined in logit space itself is helpful, we conduct an ablation where we compute a logit-based reward between the predicted answer distribution and a one-hot ground-truth label, without overwriting the answer tokens. Baseline model is trained with IoU and format rewards only. As shown in Table 2, adding this logit reward alone leads to a 0.6% drop in reasoning performance at 500 steps compared with the baseline. In contrast, our free thinking reward replaces the predicted answer tokens with the ground-truth answer and measures how likely the model is to generate the correct answer given the sampled thinking trace, *i.e.*, it constrains the probability $P(A_{\text{GT}} \mid Q, T_{\text{pred}})$. This explicitly ties the quality of the thinking process to answer correctness and yields a 1.2% improvement over the baseline model, demonstrating highly effectiveness.

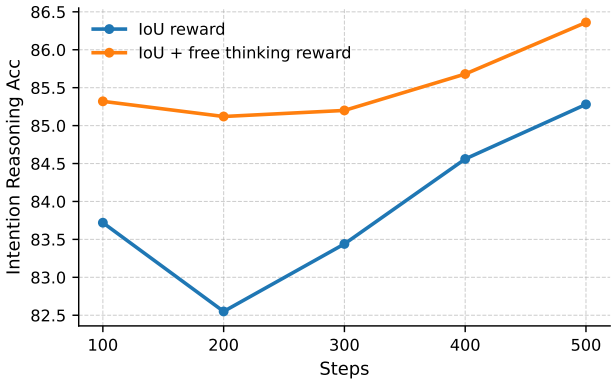

Figure 3. Intention reasoning accuracy (%) on EgoIntention context split. We show the intention reasoning accuracy trend during 500 steps RL finetuning.

Table 2. Intention reasoning accuracy (%) at different RL update steps on the EgoIntention context split. Step 0 corresponds to the supervised finetuned model before RL.

Method	0	100	200	300	400	500
IoU + format reward	86.6	83.7	82.5	83.4	84.5	85.2
+logit reward	86.6	85.2	84.6	85.2	83.7	84.6
+free thinking reward	86.6	85.3	85.1	85.2	85.7	**86.4**

Data Filtering. We broadly explore multiple data selection methods for RL on visual intention grounding tasks.

- Random. The baseline randomly samples the RL training set without any data filtering.
- Unsolved. Following prior RL-for-reasoning works, we keep only samples that the SFT model fails on, these form

Table 3. Overall grounding accuracy (Precision@0.5, %) across different training update steps on the EgoIntention dataset using five data filtering strategies.

Method	0	100	200	300	400	500
Random	60.4	62.7	62.8	62.9	63.0	63.1
Unsolved	60.4	60.0	60.3	60.1	59.2	59.5
Near threshold	60.4	62.8	62.5	62.6	62.4	62.9
Normal distribution	60.4	62.2	63.1	63.1	63.4	63.5
Ours	60.4	61.8	62.6	63.2	63.8	**64.0**

Table 4. Visual grounding accuracy (Precision@0.5, %) at different training update steps on the RefEgo-Int dataset.

Method	0	100	200	300	400	500
Naïve RL-3B	14.2	22.2	22.3	22.6	23.0	23.4
Ours-3B	14.2	23.2	23.3	23.5	24.5	**24.4**
Ours-7B	18.0	23.5	24.9	24.7	24.9	**25.5**

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).
- Ours. 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 Table 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 60.4 at step 0 to 59.5 at step 500. Compared with the baseline method without any data filtering strategy, we see a clear gap at step 500, dropping from 63.1 to 59.5. 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 0.4. 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.

4.4. Visualization

As shown in Figure 4 , we visualize paired outputs to assess how the free thinking reward shapes reasoning and grounding. Human intention queries are shown in yellow. For clarity, the system prompt is omitted. Red denotes the thinking process and answer from the model trained only with the IoU reward, while green shows the output from the model trained with both IoU and free thinking rewards. Both models are finetuned on the same randomly sampled data

The six examples on the left illustrate cases where the IoU-only model predicts correct bounding boxes but incorrect object reasoning. Models trained solely with IoU reward often confuse visually or functionally similar objects even when localization is correct. For instance, in the first sample of the first row, the user asks for a container to hold leftover paint; the model correctly localizes the paint can but labels it as a bottle. In the second sample of the second row, the user asks for a container for drinking coffee; the model's trace suggests it searches for a cup or mug, and it correctly bounds the red and white cup on the counter, but mislabels it as a mug. Other typical confusions include bucket vs. tray, bucket vs. bowl, knife vs. scissors, and sponge vs. tissue paper.

The two examples on the right show failures in both reasoning and grounding. In the first row, the model mentions a ladder in its thought process, suggesting the user needs one to reach the top shelf. However, it is a hallucination that no ladder appears in the image. The model marks the area around the stool as a ladder. In the second row, the model's reasoning fixates on the television's prominent visual features. The thinking process doesn't mention finding a communication device like a cell phone or telephone. Our model explicitly reasons that a phone is needed for making calls or confirming plans, and successfully locates the telephone in the image.

4.5. Zero-shot Evaluation

To verify that our RL framework boosts the model's generalization ability, we test our models on RefEgo-Int, a challenging egocentric visual grounding dataset. First, we evaluate the base models, which are supervised fine-tuned on the EgoIntention dataset. The results are low, the 3B model achieves 14.2% precision@0.5, and the 7B model achieves 18.0%. This is because the models are zero-shot tested without training on RefEgo dataset, and because of challenges like blurry objects caused by view motion, long-tail object categories, and cases where no ground truth bounding boxes exist in the image. The model should be aware of how to reject a query when there is no target object in the image. However, the model trained on EgoIntention doesn't have such abilities. After applying our RL finetuning framework on the EgoIntention dataset, we zero-shot test our RL models on RefEgo-Int. We observe significant improvements compared with the SFT baseline at step 0 for both 3B (+10.2% P@0.5) and 7B (+7.5% P@0.5) models.

Figure 4. Visualization comparing the baseline VLM trained with IoU reward only and our model trained with both IoU and free thinking rewards. The human intention query is colored in yellow. The baseline model's reasoning and answers are shown in red, while ours are shown in green, with key reasoning phrases underlined. For the six examples on the left, both methods produce the same grounding, shown with gray boxes. Please zoom in for better view.

We also compare our RL framework with the IoU reward only RL method, and we achieve better performance, boosting approximately 1.0% P@0.5.

5. Conclusion

In this paper, we identify a mismatch between incorrect thinking processes and correct bounding boxes in visual intention grounding. To address this, we propose a free thinking reward that scores reasoning traces by how much they boost the likelihood of the correct answer. Beyond this thinking reward, we tailor a difficulty- and informativeness-aware data filtering strategy for visual intention grounding, which focuses GRPO training on informative easy-to-medium samples. Extensive experiments on EgoIntention and RefEgo-Int demonstrate that our approach consistently improves both grounding accuracy and intention reasoning, achieving state-of-the-art results over strong Qwen2.5-VL baselines, and we hope our framework will inspire more think–answer consistent VLMs in the community.

Future Work. Our label-free thinking reward is task-agnostic, the final output format is not limited to bounding box coordinates. We plan to extend it to reasoning tasks such as visual question answering and vision-language-action models. Moreover, beyond checking the correctness of the inferred target object, we plan to design dedicated metrics to evaluate the quality of the thinking process. For example, we will check thinking process consistency across rollouts during training and whether overthinking occurs.

References

[1] 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. 2, 3

[2] 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. arXiv:2308.12966 [cs]. 6

[3] 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*, 2025. 2, 5, 6

[4] 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, 2025. arXiv:2505.14231 [cs]. 2

[5] 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, 2023. arXiv:2305.09246 [cs]. 3

[6] 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, 2023. arXiv:2310.09478 [cs]. 6

[7] 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)*, pages 18973–18990, New Orleans, LA, USA, 2022. IEEE. 5

[8] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 1, 2

[9] 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. *ICLR*, 1(2):3, 2022. 5

[10] 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. 1, 2

[11] Naman Jain, Tianjun Zhang, Wei-Lin Chiang, Joseph E. Gonzalez, Koushik Sen, and Ion Stoica. LLM-Assisted Code Cleaning For Training Accurate Code Generators, 2023. arXiv:2311.14904 [cs]. 3

[12] 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, 2025. arXiv:2505.16192 [cs]. 2

[13] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, 2014. Association for Computational Linguistics. 2

[14] 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)*, pages 15168–15178, Paris, France, 2023. IEEE. 2, 5

[15] 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, 2024. arXiv:2402.00530 [cs]. 3

[16] 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, 2024. arXiv:2308.12032 [cs]. 3

[17] You Li, Heyu Huang, Chi Chen, Kaiyu Huang, Chao Huang, Zonghao Guo, Zhiyuan Liu, Jinan Xu, Yuhua Li, Ruixuan Li, and Maosong Sun. Migician: Revealing the Magic of Free-Form Multi-Image Grounding in Multimodal Large Language Models, 2025. arXiv:2501.05767 [cs]. 2

[18] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 3

[19] Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What Makes Good Data for Alignment? A Com-

prehensive Study of Automatic Data Selection in Instruction Tuning, 2024. arXiv:2312.15685 [cs]. 3

[20] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-RFT: Visual Reinforcement Fine-Tuning, 2025. arXiv:2503.01785 [cs]. 1

[21] 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, 2025. arXiv:2505.18129 [cs]. 2

[22] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. MM-Eureka: Exploring the Frontiers of Multimodal Reasoning with Rule-based Reinforcement Learning, 2025. arXiv:2503.07365 [cs]. 2

[23] Mengxue Qu, Yu Wu, Wu Liu, Xiaodan Liang, Jingkuan Song, Yao Zhao, and Yunchao Wei. RIO: A Benchmark for Reasoning Intention-Oriented Objects in Open Environments, 2023. arXiv:2310.17290 [cs]. 1, 3

[24] Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael J Tarr, Aviral Kumar, and Katerina Fragkiadaki. Grounded reinforcement learning for visual reasoning. *arXiv preprint arXiv:2505.23678*, 2025. 2

[25] 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. 2

[26] 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, 2025. arXiv:2504.07615 [cs]. 1, 2, 5

[27] 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)*, pages 2512–2522, 2025. 1, 2, 3, 5

[28] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-RFT: Reinforcement Fine-Tuning for Visual Reasoning, 2025. arXiv:2503.20752 [cs]. 2

[29] 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. 2

[30] 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, 2025. arXiv:2505.17667 [cs]. 2, 3

[31] Wenxuan Wang, Yisi Zhang, Xingjian He, Yichen Yan, Zijia Zhao, Xinlong Wang, and Jing Liu. Beyond Literal Descriptions: Understanding and Locating Open-World Objects Aligned with Human Intentions, 2024. arXiv:2402.11265 [cs]. 1, 3

[32] Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, et al. Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20270–20281, 2023. 1

[33] 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, 2025. arXiv:2504.07934 [cs]. 2, 3

[34] 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. *arXiv preprint arXiv:2507.06448*, 2025. 2

[35] 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. 2

[36] Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting Influential Data for Targeted Instruction Tuning, 2024. arXiv:2402.04333 [cs]. 3

[37] Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, et al. Egolife: Towards egocentric life assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28885–28900, 2025. 1

[38] 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, 2025. arXiv:2504.07954 [cs]. 2

[39] 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 Vision Language Reasoning with Multi-Domain Data Curation, 2025. arXiv:2508.12680 [cs]. 2, 3

[40] Bolin Zhang, Jiahao Wang, Qianlong Du, Jiajun Zhang, Zhiying Tu, and Dianhui Chu. A Survey on Data Selection for LLM Instruction Tuning, 2025. arXiv:2402.05123 [cs]. 2

[41] Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative Verifiers: Reward Modeling as Next-Token Prediction, 2025. arXiv:2408.15240 [cs]. 1

[42] 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), Bangkok, Thailand, 2024. Association for Computational Linguistics. 5

[43] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less Is More for Alignment. 3

[44] Terry Yue Zhuo, Armel Zebaze, Nitchakarn Suppattarachai, Leandro von Werra, Harm de Vries, Qian Liu, and Niklas Muennighoff. Astraios: Parameter-Efficient Instruction Tuning Code Large Language Models, 2024. arXiv:2401.00788 [cs]. 3