GROUNDPO: POLICY OPTIMIZATION VIA INTERNALIZATION AND DISTILLATION FOR VISUAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual grounding becomes substantially more challenging when a query identifies its target through implicit comparison, spatial relations, and other forms of reasoning. Recent advances in on-policy distillation (OPD) provide a promising paradigm for such reasoning-intensive grounding. However, we find that distilling from a teacher with privileged information can cause training collapse, manifested as *Privilege-Induced Behavioral Shift* (PrivShift): privilege-specific artifacts and excessively long reasoning trajectories are inherited by the student. To address this issue, we propose GROUNDPO, a two-stage framework following the principle of *internalizing the knowledge before teaching the policy*. We first perform *Policy Internalization* to absorb privileged information into the teacher policy, encouraging reasoning that is free from privilege leakage and maintains an appropriate response length. *Policy Distillation* then uses the resulting internalized policy as the teacher to transfer its grounding capability through OPD to a student without privileged input. We further introduce DEEPGROUND, comprising a high-quality training set and a challenging benchmark spanning five tracks of reasoning-intensive grounding. Experiments show that GROUNDPO substantially mitigates PrivShift, while GROUNDPO-27B achieves state-of-the-art performance, surpassing Gemini-3.1-Pro. Further analysis shows that GROUNDPO improves fine-grained perception and relative-position reasoning, demonstrating benefits beyond visual grounding on V* Bench.

1 INTRODUCTION

The rapid development of multimodal large language models (MLLMs) (Qwen Team, 2026a; Google, 2026b; OpenAI, 2026) has substantially advanced joint vision–language understanding, enabling models to reason about visual content through natural language. Among their fundamental capabilities, visual grounding (Yu et al., 2016; Peng et al., 2024), also known as referring expression comprehension, aims to localize the image region corresponding to a given description. Accurate grounding provides a crucial interface between language understanding and fine-grained visual perception, and is therefore essential for downstream applications requiring precise language-conditioned interaction, such as robotic manipulation and image editing (Xiao et al., 2026; Jiang et al., 2026c). This capability becomes particularly challenging when identifying the target requires compositional, relational, or commonsense reasoning beyond direct visual matching.

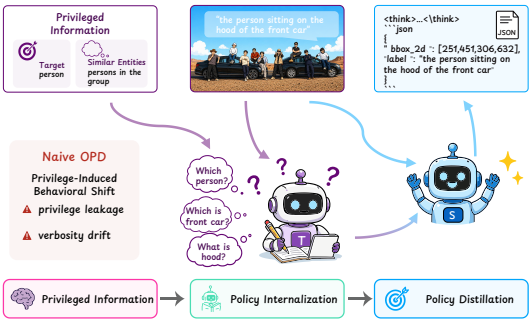


Figure 1: The central idea of GROUNDPO: internalize the knowledge before teaching the policy.

Studying such reasoning-intensive visual grounding is first constrained by the limited availability of high-quality data. Conventional grounding datasets (Yu et al., 2016; Gao et al., 2026b; Mao et al., 2016) largely emphasize relatively direct referring expressions, whereas recent challenging benchmarks (Dong et al., 2026; Li et al., 2026b) focus more on relational, comparative, and knowledge-based reasoning but are primarily designed for evaluation. To bridge this gap, we develop an automated data-construction pipeline and introduce DEEPGROUND, which provides both a dedi-

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.

cated training split and a carefully curated evaluation benchmark for reasoning-intensive grounding. DEEPGROUND-TEST spans five complementary reasoning tracks and remains challenging even for Qwen3.5-397B, which achieves only 38.18% F1@mIoU.

To improve the visual grounding capabilities of MLLMs, several post-training paradigms have been explored (Tie et al., 2025; Gao et al., 2026a; Huang et al., 2026a), primarily through supervised fine-tuning (SFT) and reinforcement learning (RL). SFT optimizes models on annotated image–query–region pairs and remains the dominant approach for acquiring grounding capabilities (Peng et al., 2024; You et al., 2024; Ma et al., 2024; Yang et al., 2026a), while recent RL-based methods further improve autoregressive grounding through task-level optimization (Jiang et al., 2026a; Bai et al., 2025b; Ren et al., 2026). In contrast, on-policy distillation (OPD) (Agarwal et al., 2024; Lu & Lab, 2025; Huang et al., 2026b) provides dense token-level supervision on student-generated trajectories, reducing the trajectory mismatch inherent in offline imitation. Recent work has further extended OPD to multimodal perception and reasoning by equipping teachers with privileged visual information, such as image crops or visual thoughts (Yuan et al., 2026; Li et al., 2026a). However, the applicability of such privileged-information-based OPD to visual grounding remains unexplored.

In this work, we systematically investigate OPD for visual grounding and identify a critical obstacle to its direct application. Following prior work (Yuan et al., 2026; Li et al., 2026a), a natural strategy is to equip the teacher with annotations of the ground-truth region. However, we find that such naive OPD can lead to training collapse, manifested in two characteristic behaviors. First, the student generates privilege-specific content that is available only to the teacher, which we refer to as *privilege leakage*. Second, its reasoning trajectories become progressively longer due to excessive reasoning and repetitive generation, which we term *verbosity drift*. We collectively refer to these two phenomena as *Privilege-Induced Behavioral Shift* (PrivShift). Our analysis in Fig. 2 further reveals the underlying cause of PrivShift. Privileged conditioning alters the teacher policy itself: the teacher produces both privilege-dependent content and substantially longer reasoning trajectories than the original policy. Direct OPD then transfers these altered behaviors to the student, despite the corresponding privileged information being unavailable at inference time. Consequently, the student inherits both privilege leakage and verbosity drift, exposing a mismatch between the privilege-conditioned policy being distilled and the policy desired at inference.

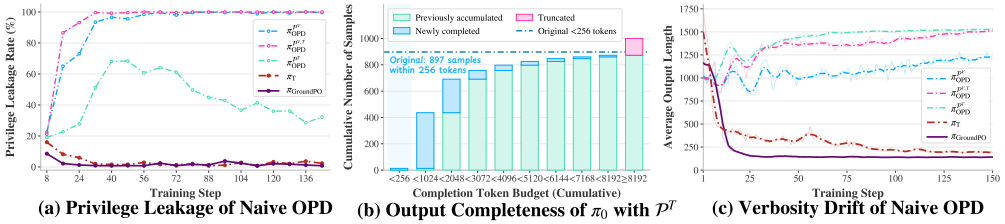
To address this issue, we propose GROUNDPO, a two-stage framework following the principle of *internalizing the knowledge before teaching the policy*. As illustrated in Fig. 1, we first perform *Policy Internalization* to absorb privileged information into the teacher policy, encouraging reasoning that is free from privilege leakage and maintains an appropriate response length. We then perform *Policy Distillation*, using the resulting internalized policy as the teacher to transfer its grounding capability through OPD to a student without privileged input, rather than directly distilling privilege-dependent trajectories. This design retains the benefits of privileged supervision while substantially mitigating PrivShift. Our contributions are summarized as follows:

- We introduce DEEPGROUND, comprising a high-quality training set and a carefully curated benchmark, which provides a challenging testbed for reasoning-intensive visual grounding.
- We systematically investigate OPD for visual grounding and identify PrivShift, a failure mode of privilege-conditioned OPD that manifests as privilege leakage and verbosity drift.
- We propose GROUNDPO, a two-stage framework that first internalizes knowledge from privileged information and subsequently transfers the resulting policy through OPD.
- Extensive experiments show that GROUNDPO substantially mitigates PrivShift, not only achieving state-of-the-art performance on DEEPGROUND-TEST but also extending its benefits to fine-grained perception and relative-position reasoning.

2 RELATED WORK

2.1 VISUAL GROUNDING

Visual grounding, or referring expression comprehension, aims to localize image regions specified by natural-language descriptions (Kazemzadeh et al., 2014; Yu et al., 2016; Mao et al., 2016). Early benchmarks primarily focus on direct referring expressions, while later datasets introduce compositional descriptions and harder distractors to better assess fine-grained language–vision alignment



(a) Privilege Leakage of Naive OPD **(b) Output Completeness of π_0 with $\mathcal{P}^T$** **(c) Verbosity Drift of Naive OPD**

Figure 2: Motivation for GROUNDPO on the 4B model. (a) OPD with privileged information induces privilege leakage, while textual privilege exhibits substantially less leakage than visual privilege. (b) Privilege-conditioned models tend to produce substantially more verbose reasoning. (c) Direct OPD from such teachers propagates this behavior to the student, resulting in verbosity drift.

(Chen et al., 2020). With the emergence of MLLMs, recent work has further extended grounding toward more challenging settings involving adversarial expressions, multimodal reasoning, and multi-dimensional evaluation (Dong et al., 2026; Gao et al., 2026b; Li et al., 2026b). Meanwhile, general-purpose MLLMs and grounding-oriented models have increasingly incorporated region-level supervision to strengthen localization capabilities (Peng et al., 2024; You et al., 2024; Ma et al., 2024; Jiang et al., 2026b; Wang et al., 2026; Su et al., 2026). Despite this progress, existing challenging benchmarks either lack corresponding training data or provide limited headroom for evaluating reasoning-intensive grounding. In contrast, DEEPGROUND jointly provides training and evaluation data covering five complementary forms of reasoning-intensive grounding.

2.2 POST-TRAINING FOR VISUAL GROUNDING

Post-training (Guo et al., 2025; Chen et al., 2024; 2026a;b; Liang et al., 2024) has become a central paradigm for improving the reasoning and structured prediction capabilities of foundation models. RL, particularly GRPO (Shao et al., 2024), has recently been extended to multimodal perception and grounding. UniVG-R1 applies RL to universal visual grounding (Bai et al., 2025b), VisionReasoner integrates reasoning with visual perception (Liu et al., 2026), and Rex-Omni introduces geometry-aware RL for generative localization (Jiang et al., 2026b). These methods optimize grounding policies primarily through outcome-level rewards. OPD (Gu et al., 2024; Lu & Lab, 2025) provides a complementary paradigm by offering token-level supervision on trajectories sampled from the student policy itself, thereby reducing the trajectory mismatch inherent to offline imitation (Agarwal et al., 2024). Recent multimodal OPD methods further strengthen the teacher with privileged information: Vision-OPD transfers crop-conditioned fine-grained perception to full-image inference (Yuan et al., 2026), while Visual-OPSD leverages privileged visual thoughts for multimodal reasoning (Li et al., 2026a). In our grounding experiments, however, directly distilling from privilege-conditioned teachers leads to privilege leakage and verbosity drift; we collectively refer to these behaviors as PrivShift. GROUNDPO addresses this issue by combining the two paradigms: RL first performs *Policy Internalization* to absorb privileged grounding knowledge into a stable teacher policy, after which OPD performs *Policy Distillation* to transfer this policy to a student without privileged input.

3 DEEPGROUND

3.1 CONSTRUCTION OF DEEPGROUND

To construct high-quality grounding data, we design a three-stage pipeline consisting of *Entity Tree Construction*, *Query Generation*, and *Quality Filtering*, as illustrated in Fig. 4. The hierarchical visual entity tree is adapted from (Yang et al., 2026b) and built through seven steps, producing a structured representation for each image. Each entity, such as a person or vehicle, is represented as a visual entity node annotated with a bounding box, description, and child nodes. Each child node recursively follows the same annotation structure. Based on these structured nodes,

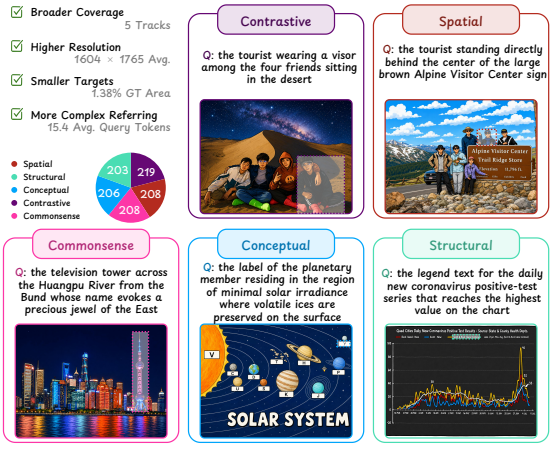


Figure 3: Visualizations of the five tracks in DEEPGROUND-TEST, a challenging benchmark for reasoning-intensive visual grounding.

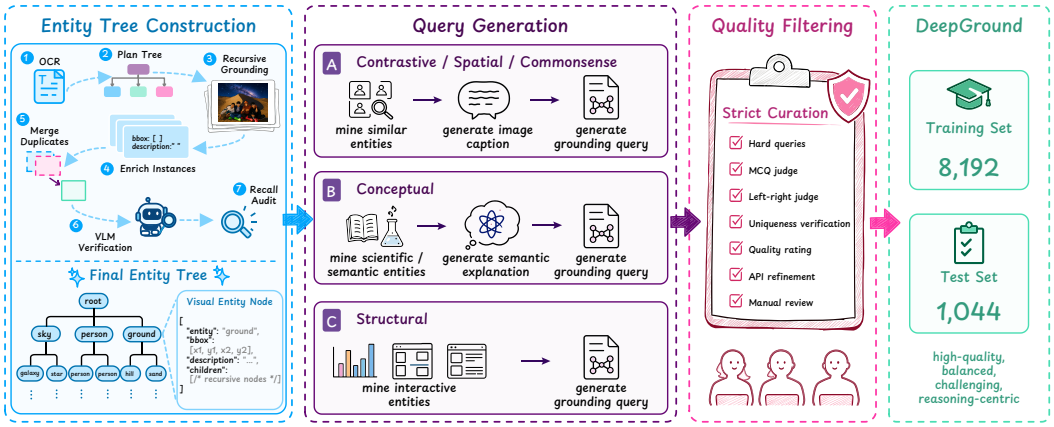


Figure 4: Construction of DEEPGROUND. The pipeline consists of three stages and ultimately produces the high-quality dataset, comprising DEEPGROUND-TRAIN and DEEPGROUND-TEST.

the Query Generation stage can directly identify similar entities, knowledge-bearing entities, and interaction-related entities, which are then used to synthesize challenging grounding queries. The queries are subsequently processed by a multi-stage filtering and verification pipeline to obtain the final DEEPGROUND. Further details are provided in the App. A.

3.2 ANALYSIS OF DEEPGROUND

DEEPGROUND is a high-quality and challenging dataset comprising two components: an 8,192-query training split, DEEPGROUND-TRAIN, and a 1,044-query test split, DEEPGROUND-TEST, with both splits cross-validated by human annotators. DEEPGROUND supports both training and evaluation, with training restricted to DEEPGROUND-TRAIN. As illustrated in Fig. 3, DEEPGROUND-TEST spans five tracks designed to evaluate whether a model can identify the target through implicit visual, relational, and knowledge-based reasoning rather than relying on a single salient attribute. Further quantitative analysis is provided in App. A.5.

4 GROUNDPO

GROUNDPO comprises two stages, as illustrated in Fig. 5. *Policy Internalization* optimizes a teacher with privileged information through RL, using grounding and format rewards. *Policy Distillation* then transfers the internalized policy to a student without privileged information, using token-level supervision on student-generated trajectories via OPD. Sec. 4.1 provides the analysis underlying the design of GROUNDPO, while Sec. 4.2 and Sec. 4.3 present the two stages in detail.

4.1 ANALYSIS OF OPD FOR VISUAL GROUNDING

Each visual grounding example is represented as a tuple $x = (I, q, B^\star)$, where I denotes an image, q a referring query, and $B^\star = \{b_j^\star\}_{j=1}^{n_{\text{gt}}}$ the set of ground-truth boxes. Given (I, q), a policy generates a response $y = (r, \widehat{B})$, where r is a reasoning trace and $\widehat{B} = \{\hat{b}_i\}_{i=1}^{n_{\text{pred}}}$ is the predicted box set. During training, the teacher may additionally observe privileged information, denoted by $\mathcal{P}$, whereas the final policy must generate y from (I, q) alone. We denote the base policy by π_0.

To analyze the effect of privileged information in OPD for visual grounding, we compare three privilege designs: $\mathcal{P}^V$, $\mathcal{P}^{V,T}$, and $\mathcal{P}^T$. Under $\mathcal{P}^V$, the teacher receives a visual annotation of the ground-truth region while using the same textual prompt as the student. $\mathcal{P}^{V,T}$ further augments the teacher prompt with a transition instruction, following (Zhao et al., 2026), to reduce the distributional discrepancy induced by the visual cue. In contrast, $\mathcal{P}^T$ removes visual annotations and provides only textual privileged information from the entity node corresponding to the ground-truth region, extracted from the hierarchical entity tree constructed in Sec. 3. Our analysis reveals two characteristic behavioral shifts under direct OPD. First, as shown in Fig. 2(a), students increasingly generate privilege-specific expressions such as "red bounding box" or "hierarchy node," despite such information being unavailable in the student input at inference time. We refer to this behavior as privilege leakage. Second, privileged conditioning substantially alters the response-length distribution of the teacher. As shown in Fig. 2(b), 89.7% of responses from the base model ter-

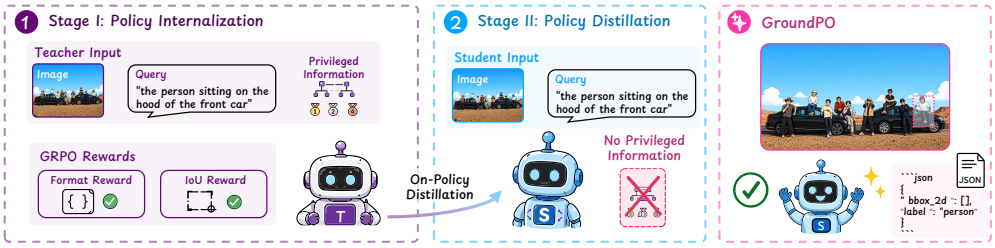


Figure 5: Overview of GROUNDPO. *Policy Internalization* embeds privileged information into the teacher through GRPO. *Policy Distillation* uses the internalized teacher to supervise student-generated trajectories while the student receives no privileged input via OPD.

minate within 256 tokens, whereas only 1.3% do so under privileged conditioning. Direct OPD subsequently transfers this shifted length distribution to the student, leading to progressively longer responses and, in severe cases, truncation before a valid localization result is produced, as shown in Fig. 2(c). We refer to this behavior as verbosity drift. Collectively, we term these two phenomena PrivShift, which exposes a mismatch between the privilege-conditioned teacher policy and the student policy desired at inference time. Representative responses are provided in App. C.4.

4.2 STAGE I: POLICY INTERNALIZATION

The above analysis suggests that an effective OPD teacher should provide additional grounding knowledge while maintaining stable generation behavior. Among the three privilege designs, $\mathcal{P}^T$ induces substantially less privilege leakage than visual privileges, and its leakage rate further decreases during training. We therefore adopt textual privileged information for teacher construction. Specifically, we encode the entity-tree descriptions and retrieve up to five entities most similar to the target, providing the descriptions of the target and its confusable entities without bounding-box coordinates or visual annotations. However, directly conditioning the teacher on $\mathcal{P}^T$ still causes severe verbosity drift. We therefore use GRPO to internalize the privileged knowledge before distillation. Rather than directly exposing the student to an unadapted privilege-conditioned teacher through token-level imitation, GRPO first optimizes the teacher for accurate grounding and complete, parseable outputs using grounding and format rewards. Under a fixed generation budget, excessively long trajectories that fail to produce a valid final prediction are disfavored, thereby mitigating verbosity drift without an explicit length penalty. Meanwhile, the reduction in privilege leakage is an empirical effect of this policy adaptation rather than the result of an explicit leakage penalty. The resulting teacher remains conditioned on $\mathcal{P}^T$ during Stage II, but exhibits substantially more stable behavior for subsequent OPD.

For each training example, the policy samples G rollouts $y_i \sim \pi_\phi(\cdot \mid I, q, \mathcal{P}^T)$. Each rollout is scored by a weighted combination of grounding accuracy and output-format correctness:

$$R_i = \lambda R_{\text{acc}}(y_i, B^\star) + (1-\lambda) R_{\text{fmt}}(y_i), \tag{1}$$

where $\lambda = 0.9$ in our experiments. R_{fmt} is a binary format reward that equals 1 when the response follows the required reasoning format with valid `<think>...</think>` delimiters, and 0 otherwise. R_{acc} measures the localization accuracy of the predicted boxes. To handle examples containing multiple targets, we compute R_{acc} using optimal one-to-one matching. Let $\mathcal{A}(\widehat{B}_i, B^\star)$ denote the set of all one-to-one matchings between predicted and ground-truth boxes. We obtain the maximum-weight Hungarian matching (Kuhn, 1955) as

$$\mathcal{M}_i^\star = \arg\max_{\mathcal{M}\in\mathcal{A}(\widehat{B}_i, B^\star)} \sum_{(k,j)\in\mathcal{M}} \text{IoU}(\hat{b}_{i,k}, b_j^\star), \; R_{\text{acc}}(y_i, B^\star) = \frac{\sum_{(k,j)\in\mathcal{M}_i^\star} \text{IoU}(\hat{b}_{i,k}, b_j^\star)}{\max(n_{\text{pred},i}, n_{\text{gt}})}. \tag{2}$$

We optimize the policy using the GRPO objective:

$$\mathcal{L}_{\text{PI}}(\phi) = -\frac{1}{\sum_{i,t} m_{i,t}} \sum_{i=1}^{G} \sum_{t} m_{i,t} \min\left(\frac{\pi_\phi(y_{i,t} \mid I, q, \mathcal{P}^T, y_{i,<t})}{\pi_{\phi_{\text{old}}}(y_{i,t} \mid I, q, \mathcal{P}^T, y_{i,<t})} \widehat{A}_i, \right.$$
$$\left. \text{clip}\left(\frac{\pi_\phi(y_{i,t} \mid I, q, \mathcal{P}^T, y_{i,<t})}{\pi_{\phi_{\text{old}}}(y_{i,t} \mid I, q, \mathcal{P}^T, y_{i,<t})}, 1-\epsilon_{\text{low}}, 1+\epsilon_{\text{high}} \right) \widehat{A}_i \right). \tag{3}$$

where $m_{i,t}$ masks valid response tokens and $\widehat{A}_i$ denotes the group-relative advantage,

$$\widehat{A}_i = R_i - \overline{R}, \qquad \overline{R} = \frac{1}{G}\sum_{j=1}^{G} R_j. \tag{4}$$

Through *Policy Internalization*, privileged information is absorbed into the model parameters rather than directly transferred through token-level imitation. The resulting policy serves as a behaviorally stable teacher for the subsequent distillation stage.

4.3 Stage II: Policy Distillation

Policy Internalization produces a teacher policy π_T with substantially less privilege leakage and a response-length distribution closer to that of π_0. We freeze π_T as the teacher and initialize the student π_θ from π_0. Following Yuan et al. (2026), the student generates an on-policy response

$$y = (y_1, \ldots, y_{|y|}) \sim \pi_\theta(\cdot \mid I, q) \tag{5}$$

without privileged information. Both policies then evaluate the same student-generated trajectory. At each position t, the student conditions on $(I, q, y_{<t})$, while the teacher additionally observes the textual privilege $\mathcal{P}^T$:

$$p_S(\cdot \mid I, q, y_{<t}) = \pi_\theta(\cdot \mid I, q, y_{<t}), \tag{6}$$

$$p_T(\cdot \mid I, q, \mathcal{P}^T, y_{<t}) = \pi_T(\cdot \mid I, q, \mathcal{P}^T, y_{<t}). \tag{7}$$

Thus, teacher and student are evaluated on the same student-generated prefixes and differ only in the availability of privileged information. We then minimize the token-level divergence between the teacher and student distributions along the student's on-policy trajectory:

$$\mathcal{D}_{\text{PD}} = \frac{1}{|y|}\sum_{t=1}^{|y|} \text{JSD}_\beta\left(p_T(\cdot \mid I, q, \mathcal{P}^T, y_{<t}) \,\|\, p_S(\cdot \mid I, q, y_{<t})\right), \tag{8}$$

where the generalized Jensen–Shannon divergence is defined as

$$m_t = \beta p_T + (1-\beta) p_S, \tag{9}$$

$$\text{JSD}_\beta(p_T \| p_S) = \beta\, \text{KL}(p_T \| m_t) + (1-\beta)\, \text{KL}(p_S \| m_t). \tag{10}$$

We set $\beta = 0.5$ in all experiments. For memory-efficient distillation, we follow the top-K implementation of Vision-OPD (Yuan et al., 2026): at each position, we retain the student's top-K tokens, compute the corresponding teacher probabilities on the same token set, and aggregate the remaining probability mass into a tail bucket. The *Policy Distillation* objective is then

$$\mathcal{L}_{\text{PD}}(\theta) = \mathbb{E}_{(I,q,\mathcal{P}^T)\sim\mathcal{D}} \mathbb{E}_{y\sim\pi_\theta(\cdot|I,q)}\left[\mathcal{D}_{\text{PD}}\right]. \tag{11}$$

Gradients are propagated only through the student, while π_T serves as a stop-gradient privileged target. The resulting policy π_{GroundPO} therefore learns from the internalized teacher during training while requiring only (I, q) at inference time. As shown in Fig. 2(a)(c), using π_T as the teacher substantially reduces both privilege leakage and verbosity drift, thereby mitigating PrivShift compared with direct privileged OPD.

5 Experiments

5.1 Experimental Setup

Training and evaluation. We train GroundPO on Qwen3.5 models at 4B, 9B, and 27B scales (Qwen Team, 2026a), with detailed training settings provided in App. B. We evaluate grounding performance using P@mIoU, R@mIoU, and their harmonic mean, $\text{F1@mIoU} = 2PR/(P+R)$, with all metrics reported as percentages. In addition to DeepGround-Test, we evaluate GroundPO along four complementary dimensions: (1) reasoning-intensive visual grounding on Ref-Adv (Dong et al., 2026), GroundingME (Li et al., 2026b), and PR-Bench (Gao et al., 2026b); (2) fine-grained

Table 1: Results on DEEPGROUND-TEST. P, R, and F1 are percentages. ✓and ✗denote thinking and non-thinking inference; "–" denotes an unreported status. The **bold** entries indicate the highest performance, and the underlined entries indicate the next-highest performance.

Model	Size	Think	Contrastive			Spatial			Commonsense			Conceptual			Structural			Overall		
			P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Proprietary models																				
GPT-5.6-Sol	–	✓	34.91	35.39	35.15	40.45	34.62	37.31	34.21	28.13	30.87	49.34	51.02	50.17	46.97	48.13	47.54	41.41	39.35	40.35
Gemini-3.1-Pro	–	✓	36.43	37.26	36.84	44.18	44.18	44.18	49.91	52.07	50.96	56.44	59.17	57.77	51.20	52.46	51.82	47.55	48.87	48.20
Gemini-3.5-Flash	–	✓	**45.41**	**46.03**	**45.71**	**55.16**	**56.49**	**55.82**	**57.87**	**58.70**	**58.28**	**63.43**	**64.66**	**64.04**	**52.55**	**54.88**	**53.69**	**54.78**	**56.03**	**55.40**
Non-thinking models																				
LLaVA-OneVision-2	8B	✗	4.77	4.70	4.74	3.87	3.80	3.83	4.81	4.81	4.81	1.06	0.73	0.86	1.35	1.13	1.23	3.40	3.07	3.23
InternVL3.5	8B	✗	2.88	2.88	2.88	2.26	2.26	2.26	3.65	3.65	3.65	4.37	4.37	4.37	4.58	4.58	4.58	3.53	3.53	3.53
InternVL3.5	38B	✗	4.57	4.57	4.57	3.41	3.41	3.41	5.36	5.34	5.35	10.49	10.39	10.44	5.95	5.71	5.83	5.92	5.86	5.89
PaDT-Pro	7B	✗	3.29	3.29	3.29	3.12	3.12	3.12	3.03	3.03	3.03	8.33	8.20	8.26	6.91	6.60	6.75	4.87	4.82	4.85
Rex-Omni	3B	✗	2.31	36.62	4.34	4.06	40.72	7.38	1.19	31.49	2.29	**20.97**	**26.17**	**23.28**	**21.17**	**21.38**	**21.27**	2.84	31.39	5.21
LocateAnything	3B	✗	19.86	19.86	19.86	21.50	21.39	21.45	17.77	17.60	17.68	8.16	8.16	8.16	13.79	13.79	13.79	16.27	16.23	16.25
Qwen3-VL	4B	✗	13.59	13.65	13.62	24.52	24.52	24.52	29.52	29.52	29.52	7.91	7.91	7.91	8.03	8.03	8.03	16.74	16.75	16.74
Qwen3-VL	8B	✗	24.98	24.98	24.98	26.56	28.61	27.55	37.60	37.60	37.60	7.86	7.86	7.86	11.87	11.87	11.87	21.95	22.29	22.12
Qwen3-VL	32B	✗	37.26	37.26	37.26	**41.20**	**41.20**	**41.20**	43.51	43.51	43.51	17.67	17.67	17.67	13.50	13.50	13.50	30.80	30.80	30.80
Qwen3-VL	235B	✗	**43.92**	**43.52**	**43.72**	39.95	39.95	39.95	**48.51**	**48.51**	**48.51**	18.11	18.11	18.11	17.93	17.93	17.93	**33.88**	**33.81**	**33.84**
Thinking models																				
UniVG-R1	7B	✓	5.66	5.66	5.66	5.14	5.14	5.14	6.68	6.68	6.68	5.67	5.58	5.62	3.09	2.96	3.02	5.28	5.22	5.25
VisionReasoner	7B	✓	6.84	15.80	9.54	7.14	15.38	9.76	10.34	22.02	14.07	8.59	12.67	10.24	4.75	8.52	6.10	7.54	14.92	10.02
MiMo-VL	7B	✓	12.92	12.74	12.83	12.93	11.87	12.38	15.41	15.19	15.30	0.20	0.19	0.20	4.95	4.63	4.78	9.36	9.00	9.18
Keye-VL-1.5	8B	✓	1.72	1.87	1.79	0.77	0.82	0.79	0.89	2.36	1.30	0.84	0.92	0.88	1.00	1.13	1.06	1.02	1.43	1.19
GLM-4.1V	9B	✓	15.45	15.02	15.23	15.76	15.53	15.64	22.21	21.78	21.99	20.53	20.63	20.58	11.40	11.23	11.32	17.08	16.84	16.96
GLM-4.6V	106B	✓	31.78	31.78	31.78	34.42	34.42	34.42	41.11	41.11	41.11	25.97	25.97	25.97	20.00	19.70	19.85	30.76	30.67	30.71
Qwen3.5	4B	✓	28.38	27.99	28.18	35.66	33.94	34.78	37.77	37.40	37.58	6.27	6.21	6.24	13.20	13.40	13.30	24.24	23.92	24.08
Qwen3.5	9B	✓	34.47	33.84	34.15	42.70	40.24	41.44	42.16	41.35	41.75	9.53	9.76	9.64	18.28	17.83	18.05	29.31	28.75	29.02
Qwen3.5	27B	✓	44.86	44.25	44.55	50.05	48.13	49.07	50.19	49.71	49.95	20.97	20.97	20.97	23.49	22.56	23.02	38.06	37.30	37.68
Qwen3.5	397B	✓	41.69	41.69	41.69	46.90	45.10	45.98	46.01	46.01	46.01	32.72	32.72	32.72	24.07	24.19	24.13	38.31	38.06	38.18
GROUNDPO	4B	✓	48.02	47.58	47.80	47.74	47.74	47.74	51.88	51.63	51.76	11.60	11.60	11.60	26.31	26.31	26.31	37.29	37.18	37.24
GROUNDPO	9B	✓	52.88	52.88	52.88	52.60	52.60	52.60	57.69	57.69	57.69	12.09	12.09	12.09	24.04	24.04	24.04	40.12	40.12	40.12
GROUNDPO	27B	✓	**57.81**	**57.81**	**57.81**	**58.41**	**58.41**	**58.41**	**66.71**	**66.39**	**66.55**	**55.05**	**55.05**	**55.05**	**44.68**	**44.68**	**44.68**	**56.60**	**56.54**	**56.57**

visual grounding on RefCOCOm (Wang et al., 2024); (3) conventional visual grounding on RefCOCO (Yu et al., 2016) and RefCOCOg (Mao et al., 2016); and (4) general visual understanding on V* Bench (Wu & Xie, 2024). Detailed descriptions of these datasets for visual grounding and the corresponding experimental results are provided in App. C.

Baselines. We compare against three groups of baselines. *Proprietary models* include GPT-5.6-Sol (OpenAI, 2026), Gemini-3.1-Pro (Google, 2026a), and Gemini-3.5-Flash (Google, 2026b). *Non-thinking models* include LLaVA-OneVision-2 (An et al., 2026), InternVL3.5-8B/38B (Wang et al., 2025), PaDT-Pro (Su et al., 2026), Rex-Omni (Jiang et al., 2026b), LocateAnything (Wang et al., 2026), and Qwen3-VL-4B/8B/32B/235B-Instruct (Bai et al., 2025a). *Thinking models* include Keye-VL-1.5 (Yang et al., 2025), UniVG-R1 (Bai et al., 2025b), VisionReasoner (Liu et al., 2026), MiMo-VL (Yue et al., 2025), GLM-4.1V/4.6V (Hong et al., 2025), and Qwen3.5-4B/9B/27B/397B.

5.2 EXPERIMENTAL RESULTS

Evaluation on DEEPGROUND-TEST. As shown in Tab. 1, the results reveal several consistent trends. Among proprietary models, Gemini-3.5-Flash achieves the strongest overall performance, while the Qwen3-VL family remains competitive among non-thinking models. Specialist grounding models exhibit uneven behavior: Rex-Omni performs strongly on the conceptual and structural tracks, whereas UniVG-R1 and VisionReasoner achieve substantially lower overall performance on DEEPGROUND-TEST. Among thinking models, GROUNDPO-27B delivers the best overall result and surpasses Gemini-3.5-Flash. The results also highlight the difficulty of reasoning-intensive grounding. Scaling the Qwen3.5 family does not yield monotonic improvements, with Qwen3.5-397B underperforming Qwen3.5-27B on some tracks. Moreover, Qwen3.5 models consistently achieve lower performance on the conceptual and structural tracks than on the contrastive, spatial, and commonsense tracks, indicating that knowledge-intensive reasoning and fine-grained structural understanding remain particularly challenging.

Evaluation on Other Grounding Benchmarks. Tab. 2 reports the performance of GROUNDPO on six grounding benchmarks. Since GROUNDPO is trained on our reasoning-intensive visual grounding split, DEEPGROUND-TRAIN, it yields consistent improvements over the corresponding base models across GroundingME, PR-Bench, and Ref-Adv at different model scales. On the fine-grained grounding benchmark RefCOCOm, GROUNDPO also achieves substantial gains, indicating improved sensitivity to fine-grained visual semantics and localization granularity. On conventional referring expression benchmarks, GROUNDPO primarily improves precision while maintaining recall close to that of the base models. This suggests that reasoning-intensive post-training

Table 2: Results on other grounding benchmarks with thinking enabled. P@mIoU, R@mIoU, and F1@mIoU are reported as percentages; Δ denotes the absolute change from the base policy.

	GroundingME			PR-Bench			Ref-Adv			RefCOCOm			RefCOCO			RefCOCOg		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Qwen3.5-4B	40.09	40.24	40.16	58.82	59.57	59.19	45.42	44.98	45.20	33.32	33.76	33.53	68.87	70.78	69.81	72.49	73.74	73.11
GROUNDPO-4B	44.09	43.87	43.98	59.30	59.27	59.28	49.39	49.34	49.36	39.22	39.13	39.18	70.21	70.15	70.18	73.04	72.57	72.80
Δ	+4.00	+3.63	+3.82	+0.48	-0.30	+0.09	+3.97	+4.36	+4.16	+5.90	+5.37	+5.65	+1.34	-0.63	+0.37	+0.55	-1.17	-0.31
Qwen3.5-9B	47.31	46.07	46.68	58.83	60.59	59.70	50.16	49.33	49.74	35.28	35.62	35.45	71.22	72.76	71.98	74.33	75.33	74.82
GROUNDPO-9B	49.09	49.09	49.09	60.74	60.72	60.73	51.06	50.97	51.02	41.16	41.13	41.15	71.71	71.81	71.76	74.36	74.28	74.32
Δ	+1.78	+3.02	+2.41	+1.91	+0.13	+1.03	+0.90	+1.64	+1.28	+5.88	+5.51	+5.70	+0.49	-0.95	-0.22	+0.03	-1.05	-0.50
Qwen3.5-27B	52.32	51.85	52.08	59.48	62.20	60.81	52.16	52.12	52.14	36.24	37.26	36.75	72.70	75.09	73.87	75.73	77.56	76.64
GROUNDPO-27B	58.26	58.26	58.26	62.72	62.71	62.71	53.42	53.42	53.42	44.06	44.05	44.05	73.33	73.27	73.30	76.37	76.36	76.37
Δ	+5.94	+6.41	+6.18	+3.24	+0.51	+1.90	+1.26	+1.30	+1.28	+7.82	+6.79	+7.30	+0.63	-1.82	-0.57	+0.64	-1.20	-0.27

enhances target discrimination and suppresses spurious predictions without substantially sacrificing the model's original grounding coverage.

5.3 ABLATION STUDY

Choice of Privileged Information. As described in Sec. 4.2, our default privileged information $\mathcal{P}^T$ consists of textual descriptions of the ground-truth target and entities similar to it. To examine this design choice, we further compare target-only textual descriptions $\mathcal{P}_b^T$, derived from the Entity Tree in Sec. 3, and visual ground-truth cues $\mathcal{P}^V$. After *Policy Internalization*, these variants yield teacher policies $\pi_T^{\mathcal{P}^T}$, $\pi_T^{\mathcal{P}_b^T}$, and $\pi_T^{\mathcal{P}^V}$, respectively. All models are evaluated on DEEPGROUND-TEST using only the original image and query. As shown in Tab. 3, $\pi_T^{\mathcal{P}^T}$ performs best among the teacher variants, indicating that descriptions of confusable entities provide more effective supervision for internalizing target-discriminative knowledge than target-only or visual privileges. It also outperforms π_{GRPO}, which is trained without privileged information, confirming the benefit of textual privilege during *Policy Internalization*. Finally, π_{GroundPO} achieves the best overall performance, surpassing both $\pi_T^{\mathcal{P}^T}$ and π_{GRPO}. This result shows that *Policy Distillation* further transfers and strengthens the internalized grounding capability in a privilege-free student policy.

Table 3: Privileged information ablation on Qwen3.5-4B with thinking enabled. P, R, and F1 are percentages.

Policy	**P@mIoU**	**R@mIoU**	**F1@mIoU**
π_0	24.24	23.92	24.08
$\pi_T^{\mathcal{P}^T}$	36.88	36.88	36.88
$\pi_T^{\mathcal{P}_b^T}$	35.00	35.04	35.02
$\pi_T^{\mathcal{P}^V}$	33.19	33.03	33.11
π_{GRPO}	35.12	35.19	35.16
π_{GroundPO}	**37.29**	**37.18**	**37.24**

Privilege leakage and verbosity drift. To further quantify the effectiveness of GROUNDPO in mitigating PrivShift, we sample 1,000 queries from DEEPGROUND-TRAIN and analyze the behavior of both teacher and student models. We compare the base policy π_0; privilege-conditioned variants of π_0 equipped with $\mathcal{P}^V$, $\mathcal{P}^{V,T}$, and $\mathcal{P}^T$, corresponding to visual privilege, visual privilege with a transition prompt, and textual privilege, respectively; their directly distilled students π_{OPD}^V, $\pi_{\text{OPD}}^{V,T}$, and π_{OPD}^T; and our internalized teacher π_T and final model π_{GroundPO}. As shown in Tab. 4, directly privilege-conditioned teachers exhibit substantial privilege leakage and significantly longer responses. In particular, the average response length of π_0 with $\mathcal{P}^T$ is approximately $20\times$ that of π_0. Directly distilling these teachers further propagates the shifted behavior to the student, resulting in both higher artifact rates and increased response lengths. In contrast, *Policy Internalization* substantially mitigates PrivShift: π_T exhibits only a small fraction of privilege leakage while maintaining a much shorter response distribution. Consequently, π_{GroundPO} preserves stable generation behavior after *Policy Distillation*, with low privilege leakage and limited verbosity drift.

Table 4: Privilege leakage and verbosity drift on 4B models, measured by leakage rate and average output length, respectively. Evaluation details are provided in App. B.4.

Metric	π_0	π_0 w. $\mathcal{P}^V$	π_0 w. $\mathcal{P}^{V,T}$	π_0 w. $\mathcal{P}^T$	$\pi_{\text{OPD}}^{\mathcal{P}^V}$	$\pi_{\text{OPD}}^{\mathcal{P}^{V,T}}$	$\pi_{\text{OPD}}^{\mathcal{P}^T}$	π_T	π_{GroundPO}
Leakage (%)	2.27	92.18	98.08	40.72	98.03	100.00	30.91	0.21	0.21
Avg. length	204.7	174.1	1232.6	4089.4	333.8	8108.9	8192.0	133.7	134.3

5.4 DISCUSSION

5.4.1 GENERALIZATION TO VISUAL UNDERSTANDING

Although GROUNDPO is trained exclusively on visual grounding, it introduces no task-specific architectural modifications and preserves general multimodal understanding capabilities. Tab. 5 reports the results on V* Bench. Across both 4B and 9B models, π_{GroundPO} consistently improves relative-position reasoning, while direct-attribute accuracy decreases slightly, leaving the overall performance unchanged from the corresponding base models. A similar trend is observed for π_T, with more pronounced gains in relative-position reasoning. These results suggest that the privileged information describing similar entities in $\mathcal{P}^T$ primarily strengthens relational reasoning among visual entities while preserving overall visual understanding.

Table 5: Accuracy (%) on V* Bench. Parenthesized values are absolute changes from the base policy.

Policy	Size	Direct attributes	Relative position	Overall
π_0	4B	84.35	80.26	82.72
π_T	4B	82.61 (-1.74)	86.84 (+6.58)	84.29 (+1.57)
π_{GroundPO}	4B	82.61 (-1.74)	82.89 (+2.63)	82.72 (+0.00)
π_0	9B	85.22	86.84	85.86
π_T	9B	83.48 (-1.74)	89.47 (+2.63)	85.86 (+0.00)
π_{GroundPO}	9B	84.35 (-0.87)	88.16 (+1.32)	85.86 (+0.00)

5.4.2 CASE STUDY

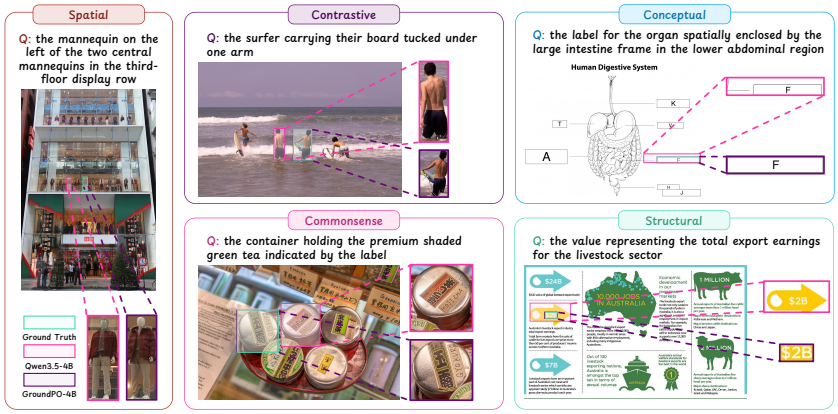


Figure 6: Qualitative case studies on DEEPGROUND-TEST. GROUNDPO demonstrates stronger reasoning and more fine-grained grounding capabilities.

Fig. 6 presents representative qualitative examples that further illustrate the strengths of GROUNDPO. On commonsense examples, GROUNDPO exhibits stronger reasoning ability, consistent with the quantitative gains observed on reasoning-intensive benchmarks in Tab. 1 and Tab. 2. On spatial and contrastive examples, it more reliably distinguishes the target from visually similar candidates through relative and comparative cues, supporting the role of $\mathcal{P}^T$ in improving inter-entity comparison and aligning with the gains on V* Bench reported in Tab. 5. On conceptual and structural examples, GROUNDPO produces more fine-grained localization, whereas the Qwen3.5 baselines frequently suffer from granularity mismatch; this observation is consistent with the improvements on RefCOCOm in Tab. 2. Together, these cases provide qualitative evidence for the improvements observed across the five reasoning-intensive grounding tracks.

6 CONCLUSION

We study OPD for reasoning-intensive visual grounding and identify PrivShift as a key obstacle to directly distilling from privilege-conditioned teachers. Privileged inputs can induce both privilege leakage and verbosity drift, causing the student to inherit generation behaviors unavailable at inference time. To address this issue, we propose GROUNDPO, which first internalizes textual privileged information through GRPO and then distills the resulting teacher policy through OPD. We further introduce DEEPGROUND, providing both training and evaluation data across five forms of reasoning-intensive grounding. Extensive experiments show that GROUNDPO substantially mitigates PrivShift, achieves state-of-the-art performance on DEEPGROUND-TEST, and further improves fine-grained perception and relative-position reasoning. These results support a simple principle for privileged OPD: *internalizing the knowledge before teaching the policy*.

AI USE STATEMENT

In this work, we used generative AI tools to assist literature discovery, manuscript drafting, language editing, generate synthetic data sets, interpret results, assist with translation, create or modify scientific figures or images, suggest experimental parameters, create or edit software code, edit a research paper to improve readability, and identify relevant literature.

We did not use generative AI tools to suggest a structure for a research paper, propose a title or keywords for a research paper, design or provide feedback on research methodology or experiments, implement methods, clean and reformat dataset, support qualitative and thematic data analysis, or help develop theoretical models or conceptual frameworks. The categories formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, and propose or refine hypotheses are not applicable to this work.

All AI-assisted content was reviewed by the authors. In particular, literature and reference suggestions provided by generative AI were manually checked and verified against the original sources, and LLM-generated code was reviewed, executed, and tested for correctness before use. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We follow the ICLR Code of Ethics. Our work involves the construction and evaluation of a visual grounding benchmark using publicly available images and model-assisted data generation. Human annotators were asked only to review and verify grounding examples derived from public images and were free to skip any example. The annotation process did not require annotators to provide personal information or make judgments about sensitive personal attributes.

DEEPGROUND is intended for research on visual grounding, fine-grained perception, and multimodal reasoning. Although some source images may contain people, the benchmark is not designed for surveillance, biometric identification, identity inference, or sensitive-attribute prediction. We release only referring expressions, target regions, and dataset identifiers, and provide usage guidance discouraging applications involving identification of individuals or inference of sensitive attributes. As with other benchmarks constructed from public image collections and model-generated annotations, DEEPGROUND may inherit biases from its source data and the MLLMs used during data construction; therefore, performance on the benchmark should not be interpreted as evidence of fairness or robustness across demographic groups or deployment settings. We encourage users to follow the licenses and usage terms of the original data sources and to consider potential privacy and misuse risks when adapting the benchmark to downstream applications.

All automatically generated or model-assisted annotations used in the benchmark were subjected to additional filtering and human verification as described in the paper. We are not aware of any conflicts of interest related to this work, and we take responsibility for the ethical use, documentation, and release of the resulting dataset and models.

REPRODUCIBILITY STATEMENT

We provide materials to facilitate reproduction of both GROUNDPO and DEEPGROUND. The submitted supplementary materials include the core implementation of the GROUNDPO algorithm. Representative examples from DEEPGROUND-TEST are provided in Figure 3, while Appendix A documents the data sources and construction pipeline of DEEPGROUND in detail, including the procedures used to generate and curate DEEPGROUND-TRAIN and DEEPGROUND-TEST. Together with the implementation and experimental details reported in the main paper and appendix, these materials are intended to support reproduction of our method, dataset construction, and evaluation results. We will publicly release the full code and dataset upon acceptance.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Kaichen Zhang, Wenkang Zhang, Zheng Cheng, Nansen Zhang, Chunsheng Wu, Chunjiang Ge, Zimin Ran, Dehua Song, Chunyuan Li, Shikun Feng, Ming Hu, Zhangquan Chen, Junbo Niu, Bo Li, Ziyong Feng, Ziwei Liu, Zongyuan Ge, and Jiankang Deng. Llava-onevision-2: Towards next-generation perceptual intelligence, 2026. URL https://arxiv.org/abs/2605.25979.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 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.

Xiao Chen, Qihui Zhang, and Yan Wang. Cliff: Leveraging ambiguous samples for enhanced test-time adaptation. In *ECAI 2024: 27th European Conference on Artificial Intelligence, 19–24 October 2024, Santiago de Compostela, Spain–Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024)*, pp. 642–649. SAGE Publications Pvt. Ltd 1 Oliver's Yard, 55 City Road, London, EC1Y 1SP, 2024.

Xiao Chen, Zhongjing Du, Jiazhen Huang, Xu Jiang, Li Lu, Jingyan Jiang, and Zhi Wang. Neural collapse in test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10567–10576, June 2026a.

Xiao Chen, Jiazhen Huang, Zhiming Liu, Qinting Jiang, Fanding Huang, Jingyan Jiang, and Zhi Wang. Test-time distillation for continual model adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 7593–7604, June 2026b.

Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee K. Wong, and Qi Wu. Cops-ref: A new dataset and task on compositional referring expression comprehension. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10083–10092, 2020. doi: 10.1109/CVPR42600.2020.01010.

Qihua Dong, Kuo Yang, Lin Ju, Handong Zhao, Yitian Zhang, Yizhou Wang, Huimin Zeng, Jianglin Lu, and Yun Fu. Ref-adv: Exploring MLLM visual reasoning in referring expression tasks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=iEBgrepR9i.

Lei Gao, Zhuoming Li, Mengxi Jia, Jiakang Yuan, Hongbo Sun, Hao Sun, and Xuelong Li. Segment-aligned policy optimization for multi-modal reasoning. *arXiv preprint arXiv:2605.01327*, 2026a.

Tianyi Gao, Han Fang, Tianyi Ding, Hao Li, Xin Wei, Hongbo Sun, Xiaodong Dong, Ye Yuan, Jinglin Xu, Kongming Liang, Hao Sun, and Jingmin Xin. Mixture-of-thought-tokens: Unifying perception and reasoning for free-form multimodal grounding. *arXiv preprint arXiv:2607.24407*, 2026b.

Google. Gemini 3.1 Pro. https://deepmind.google/models/model-cards/gemini-3-1-pro/, 2026a.

Google. Gemini 3.5 Flash. https://deepmind.google/models/model-cards/gemini-3-5-flash/, 2026b.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=5h0qf7IBZZ.

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

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025.

Yu-Chung Hsiao, Fedir Zubach, Gilles Baechler, Srinivas Sunkara, Victor Carbune, Jason Lin, Maria Wang, Yun Zhu, and Jindong Chen. ScreenQA: Large-scale question-answer pairs over mobile app screenshots. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9427–9452, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.477. URL https://aclanthology.org/2025.naacl-long.477/.

Jiazhen Huang, Xiao Chen, Zhiming Liu, Yaru Sun, Jingyan Jiang, and Zhi Wang. What drives test-time adaptation for clip? a controlled empirical study from an update perspective. *arXiv preprint arXiv:2606.14299*, 2026a.

Jiazhen Huang, Xiao Chen, Xiao Luo, Yong Dai, Senkang Hu, and Yuzhi Zhao. Skill-conditioned gated self-distillation for llm reasoning. *arXiv preprint arXiv:2605.28791*, 2026b.

Qing Jiang, Xingyu Chen, Zhaoyang Zeng, Junzhi Yu, and Lei Zhang. Rex-thinker: Grounded object referring via chain-of-thought reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=btWHQoSZZ1.

Qing Jiang, Junan Huo, Xingyu Chen, Yuda Xiong, Zhaoyang Zeng, Yihao Chen, Tianhe Ren, Junzhi Yu, and Lei Zhang. Detect anything via next point prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25472–25483, June 2026b.

Xu Jiang, Bin Chen, Gehui Li, Yule Duan, Ronggang Wang, and Jian Zhang. Octot2i: A self-evolving agentic text-to-image router. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31628–31638, June 2026c.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In Alessandro Moschitti, Bo Pang, and Walter Daelemans (eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 787–798, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1086. URL https://aclanthology.org/D14-1086/.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pp. 235–251. Springer, 2016.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are You Smarter Than a Sixth Grader? Textbook Question Answering for Multimodal Machine Comprehension . In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5376–5384, Los Alamitos, CA, USA, July 2017. IEEE Computer Society. doi: 10.1109/CVPR.2017.571. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.571.

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

H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. doi: https://doi.org/10.1002/nav.3800020109. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nav.3800020109.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Hugo Laurençon, Léo Tronchon, and Victor Sanh. Unlocking the conversion of web screenshots into html code with the websight dataset. *arXiv preprint arXiv:2403.09029*, 2024.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14369–14387, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.775. URL https://aclanthology.org/2024.acl-long.775/.

Pengyu Li, Zhitao Gao, Lingling Zhang, Muye Huang, Yuanming Li, Fangzhi Xu, and Jun Liu. Visual-opsd: Cross-modal on-policy self-distillation for efficient unified multimodal reasoning. *arXiv preprint arXiv:2606.18974*, 2026a.

Rang Li, Lei Li, Shuhuai Ren, Hao Tian, Shuhao Gu, Shicheng Li, Zihao Yue, Yudong Wang, Wenhan Ma, Zhe Yang, Jingyuan Ma, Zhifang Sui, and Fuli Luo. Groundingme: Exposing the visual grounding gap in mllms through multi-dimensional evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2412–2422, June 2026b.

Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodel large language models. In *Proceedings of the 3rd international conference on computer, artificial intelligence and control engineering*, pp. 405–409, 2024.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. MMC: Advancing multimodal chart understanding with large-scale instruction tuning. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1287–1310, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.70. URL https://aclanthology.org/2024.naacl-long.70/.

Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao PENG, Shu Liu, Bei Yu, and Jiaya Jia. Visionreasoner: Unified reasoning-integrated visual perception via reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=QoDOwjsbAq.

Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20251026. https://thinkingmachines.ai/blog/on-policy-distillation.

Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *European Conference on Computer Vision*, pp. 417–435. Springer, 2024.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. UniChart: A universal vision-language pretrained model for chart comprehension and reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14662–14684, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.906. URL https://aclanthology.org/2023.emnlp-main.906/.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1697–1706, January 2022.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. https://openai.com/index/gpt-5-6/, 2026. Accessed: 2026-09-11.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=lLmqxkfSIw.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-27b.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026c. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Xiangxuan Ren, Zhongdao Wang, Liping Hou, Pin Tang, Guoqing Wang, and Chao Ma. Grounding everything in tokens for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 41171–41181, June 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Yongyi Su, Haojie Zhang, Shijie Li, Nanqing Liu, Jingyi Liao, Junyi Pan, Yuan Liu, Xiaofen Xing, Chong Sun, Chen Li, Nancy F. Chen, Shuicheng YAN, Xulei Yang, and Xun Xu. Patch-as-decodable-token: Towards unified multi-modal vision tasks in MLLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=xF0Dcmvsl0.

Guiyao Tie, Zeli Zhao, Dingjie Song, Fuyang Wei, Rong Zhou, Yurou Dai, Wen Yin, Zhejian Yang, Jiangyue Yan, Yao Su, et al. A survey on post-training of large language models. *arXiv preprint arXiv:2503.06072*, 2025.

Shihao Wang, Shilong Liu, Yuanguo Kuang, Xinyu Wei, Yangzhou Liu, Zhiqi Li, Yunze Man, Guo Chen, Andrew Tao, Guilin Liu, et al. Locateanything: Fast and high-quality vision-language grounding with parallel box decoding. *arXiv preprint arXiv:2605.27365*, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Wenxuan Wang, Tongtian Yue, Yisi Zhang, Longteng Guo, Xingjian He, Xinlong Wang, and Jing Liu. Unveiling parts beyond objects: Towards finer-granularity referring expression segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12998–13008, 2024. doi: 10.1109/CVPR52733.2024.01235.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13084–13094. IEEE, 2024.

Linhui Xiao, Xiaoshan Yang, Xiangyuan Lan, Yaowei Wang, and Changsheng Xu. Toward Visual Grounding: A Survey . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 48(03): 2749–2771, March 2026. ISSN 1939-3539. doi: 10.1109/TPAMI.2025.3630635. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2025.3630635.

Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, et al. Kwai keye-vl 1.5 technical report. *arXiv preprint arXiv:2509.01563*, 2025.

Guangheng Yang, Zhenliang Ni, Zhenkai Wu, Han Shu, Juan Feng, Wenming Yang, and Jie Hu. Mcpo: Modality-contrastive preference optimization for multimodal chain-of-thought compression. *arXiv preprint arXiv:2609.04947*, 2026a.

Zhiwei Yang, Yuanchen Wu, NAN ZHANG, Yucong Meng, Ke Yan, and Shouhong Ding. Scene graph thinking: Reinforcing structured visual reasoning for multimodal large language models. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=hRSLIreHlx.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2msbbX3ydD.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European conference on computer vision*, pp. 69–85. Springer, 2016.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, et al. Mimo-vl technical report. *arXiv preprint arXiv:2506.03569*, 2025.

Liang Zhang, Anwen Hu, Haiyang Xu, Ming Yan, Yichen Xu, Qin Jin, Ji Zhang, and Fei Huang. TinyChart: Efficient chart understanding with program-of-thoughts learning and visual token merging. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1882–1898, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.112. URL https://aclanthology.org/2024.emnlp-main.112/.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Jpxfof0EaS.

A CONSTRUCTION OF DEEPGROUND

A.1 DATA SOURCE

DEEPGROUND integrates diverse visual sources, including natural images, scientific textbook diagrams, screenshots, and other visually structured content. The images are collected from SA-1B (Kirillov et al., 2023), TQA (Kembhavi et al., 2017), ScreenQA (Hsiao et al., 2025), AI2D (Kembhavi et al., 2016), InfoVQA (Mathew et al., 2022), UniChart (Masry et al., 2023), FigureQA (Kahou et al., 2017), TinyChart (Zhang et al., 2024), MMC_Instruction (Liu et al., 2024), ArxivQA (Li et al., 2024), and WebSight (Laurençon et al., 2024). DEEPGROUND-TRAIN contains 8,192 queries over 5,934 images, while DEEPGROUND-TEST contains 1,044 queries over 904 images and spans five reasoning-intensive grounding tracks: contrastive, spatial, commonsense, conceptual, and structural grounding. The detailed source distributions of DEEPGROUND-TRAIN and DEEPGROUND-TEST are reported in Tab. 6 and Tab. 7, respectively. In addition to the above public sources, DEEPGROUND-TEST includes a small self-constructed subset consisting primarily of images synthesized with Python scripts and images generated by GPT-5.6-Sol (OpenAI, 2026). This subset contains 33 images and 38 queries.

Table 6: Source distribution of DEEPGROUND-TRAIN. Shares are computed over the 8,192 queries in the training split.

Source	Queries	Images	Share (%)
SA-1B	6,612	4,805	80.7
TQA	792	565	9.7
InfoVQA	254	183	3.1
ScreenQA	192	141	2.3
AI2D	90	69	1.1
UniChart	66	42	0.8
FigureQA	55	36	0.7
TinyChart	51	33	0.6
MMC_Instruction	34	24	0.4
ArxivQA	30	22	0.4
WebSight	16	14	0.2
Total	**8,192**	**5,934**	**100.0**

A.2 ENTITY TREE CONSTRUCTION

To obtain structured and fine-grained visual representations, we design a seven-stage Entity Tree Construction pipeline that converts each image into a hierarchical visual entity tree as shown in Fig. 4. Each node stores an entity name, pixel-space bounding box, visual description, and recursively defined child nodes. The pipeline begins with OCR and whole-image planning to infer entity categories, instance counts, and parent–child relations. It then recursively grounds each planned entity to instance-level regions and descriptions. To improve precision, highly overlapping entities are merged using geometric criteria, while an MLLM further validates whether each localized region is semantically consistent with its entity description, correcting or removing invalid nodes when necessary. Since these cleaning operations may disrupt aggregation structures, group nodes are subsequently rebuilt from their surviving children.

The final stage performs complementary recall audits to recover missed entities. At the image level, uncovered regions are identified from the spatial coverage of existing boxes and inspected for missing objects; at the parent level, entities with insufficient child coverage are re-examined for omitted sub-entities. Newly discovered entities are grounded, deduplicated against the existing tree, semantically validated, and inserted at the appropriate hierarchy level. The resulting entity tree therefore combines global planning, recursive localization, spatial deduplication, semantic verification, structural repair, and recall completion, providing a structured representation for subsequent query construction.

A.3 QUERY CONSTRUCTION

We construct grounding queries for five reasoning tracks: *Spatial*, *Contrastive*, *Commonsense*, *Conceptual*, and *Structural*. Spatial, Contrastive, and Commonsense queries are generated mainly from

Table 7: Source distribution of DEEPGROUND-TEST. "Self" denotes the self-constructed benchmark subset. Shares are computed over 1,044 queries.

Source	Queries	Images	Share (%)
SA-1B	624	522	59.8
TQA_train	179	165	17.1
InfoVQA	60	52	5.7
ScreenQA	48	45	4.6
AI2D	26	24	2.5
FigureQA	15	14	1.4
UniChart	15	14	1.4
ArxivQA	15	12	1.4
TinyChart	13	12	1.2
MMC_Instruction	9	9	0.9
WebSight	2	2	0.2
Self	38	33	3.6
Total	**1,044**	**904**	**100.0**

natural images and share a common pipeline. We first encode the name and description of each visual entity node using Qwen3-Embedding-8B (Zhang et al., 2025) and cluster nodes by semantic similarity to identify confusable entities. We then generate an image-level caption from the entity tree using Qwen3.6-27B (Qwen Team, 2026b). Given the caption and the retrieved similar entities, Qwen3.5-397B-A17B (Qwen Team, 2026a) determines the appropriate reasoning track and synthesizes a grounding query that uniquely refers to the target.

Baseline Prompt

```
Locate every object that matches the description \"{query}\" in the image. Report bbox coordinates in JSON format.
```

Figure 7: Conventional baseline prompt.

Conceptual and Structural queries use specialized generation pipelines. For Conceptual grounding, Qwen3.5-397B-A17B first identifies entities associated with scientific or domain-specific concepts, generates an explanation of the underlying concept, and then converts this knowledge into an implicit grounding query. For Structural grounding, the model identifies entities involved in structured or interactive visual relations and directly generates queries based on these relations. Together, these track-specific pipelines transform the structured entity tree into diverse reasoning-intensive referring expressions rather than relying on direct appearance descriptions.

Teacher Prompt used in Policy Internalization

```
Locate every object that matches the description \"{query}\" in the image.\nThe following entities describe the intended target together with other semantically similar, easily-confused entities in the same scene. \"target\" is the entity the description refers to; \"similar\" entries are distractors that resemble it. Each node carries an entity name and a short description. Use this contrast only to appreciate the distinguishing attributes that set the target apart from the distractors; do not copy these names verbatim into your answer.\n```json\n{\n  \"nodes\": [\n    {\n      \"entity\": \"...\",\n      \"description\": \"...\",\n      \"role\": \"...\"\n    },\n    {\n      \"entity\": \"...\",\n      \"description\": \"...\",\n      \"role\": \"...\"\n    }, ...\n  ]\n}\n```\n\nUse the similar-entity context above to contrast the target with the easily-confused distractors: figure out which distinguishing attributes (appearance, text, spatial relation, structure, or function) set the target apart from them. Decide which one the description refers to from that semantic contrast, and ground it from the image.\nFirst think about the reasoning process step by step as an internal monologue enclosed within <think> </think> tags, then output the localization result as exactly one JSON code block with the following format:\n```json\n[\n  {{\"bbox_2d\": [x1, y1, x2, y2], \"label\": \"a concise description of the target\"}}\n]\n```
```

Figure 8: Prompt in Policy Internalization.

A.4 QUALITY FILTERING AND VERIFICATION

The generated queries undergo a multi-stage filtering and verification process. For Spatial, Contrastive, and Commonsense samples, we first retain queries on which both Qwen3-VL-4B-Instruct

Table 8: Comparison of grounding datasets. Target area is the average ground-truth region share, query length is measured in tokens, and the final column reports Qwen3.5-397B F1@mIoU (%).

Dataset	Train	Target area (%)	Query tokens	F1@mIoU
RefCOCO	✓	19.86	4.0	74.34
RefCOCOg	✓	20.64	8.8	77.61
RefCOCO+	✓	20.24	4.4	69.89
Ref-Adv	✗	8.80	12.8	51.17
GroundingME	✗	1.89	49.3	50.77
PR-Bench	✓	6.80	13.3	58.23
DEEPGROUND	✓	**1.38**	**15.4**	**38.18**

and Qwen3-VL-8B-Instruct (Qwen Team, 2025) fail to match the ground-truth region, thereby removing overly easy examples. Queries without left–right semantics are converted into a multiple-choice verification task containing the predictions of the two models and the ground-truth region; Qwen3.5-397B-A17B is then required to identify the region referred to by the query, and only samples consistent with the ground truth are retained. Queries involving left–right relations are separately checked and corrected by Qwen3.6-35B-A3B (Qwen Team, 2026c) with thinking enabled. The resulting samples are further subjected to uniqueness verification, difficulty scoring, and quality assessment with Qwen3.5-397B-A17B. High-quality samples receiving the highest score are additionally reviewed and refined by Gemini-3.5-Pro before final manual inspection.

Student Prompt used in Policy Distillation

```
Locate every object that matches the description \"{query}\" in the image. First think about the reasoning process step by step as an internal monologue enclosed within <think> </think> tags, then output the localization result as exactly one JSON code block with the following format:\n```json\n[\n {\"bbox_2d\": [x1, y1, x2, y2], \"label\": \"a concise description of the target\"}\n]\n```
```

Figure 9: Prompt of student used in Policy Distillation.

Visual Prompt with Transition

```
Locate every object that matches the description \"{query}\" in the image.\nThe target region(s) are already highlighted with red rectangle bounding box(es) in the image. Use the highlighted region(s) only to identify which object(s) the description refers to, but do not mention the rectangles, highlights, annotations, or GT boxes in your reasoning; reason independently in your own words.\n First think about the reasoning process step by step as an internal monologue enclosed within <think> </think> tags, then output the localization result as exactly one JSON code block with the following format:\n ```json\n [\n {{"bbox_2d": [x1, y1, x2, y2], "label": "a concise description of the target"}}\n ]\n ```
```

Figure 10: Prompt for Visual Privilege and Transition Instructions.

Conceptual and Structural queries follow analogous verification procedures tailored to their generation processes. Conceptual samples that remain challenging for Qwen3-VL-4B-Instruct and Qwen3-VL-8B-Instruct are judged by Qwen3.5-397B-A17B with thinking enabled, while Structural samples are directly evaluated by the same model for semantic correctness and target consistency. All retained samples are finally manually reviewed to remove ambiguous, incorrect, or low-quality queries. This automated-to-manual curation pipeline ensures that the resulting data are both challenging and reliably grounded to the intended target.

A.5 ANALYSIS OF DEEPGROUND-TEST

DEEPGROUND-TEST contains 1,044 queries, including 219 contrastive, 208 spatial, 208 commonsense, 206 conceptual, and 203 structural examples, with no overlapping with DEEPGROUND-TRAIN. The five tracks capture complementary sources of reasoning difficulty. *Contrastive* grounding requires distinguishing the target from visually similar candidates through comparative cues; *Spatial* grounding relies on relative position, ordering, direction, or topology; *Commonsense* grounding requires everyday knowledge about object functions, affordances, body functions, or social conventions; *Conceptual* grounding involves domain-specific concepts, scientific knowledge, or technical semantics; and *Structural* grounding requires reasoning over structured visual information, such

as part–whole relations, layouts, or visual organization. Together, these tracks evaluate whether a model can resolve the intended target through implicit visual, relational, and knowledge-based reasoning rather than simply matching a salient attribute.

Tab. 8 compares DEEPGROUND-TEST with both conventional and recent reasoning-oriented grounding datasets. Beyond covering a broader range of reasoning patterns, DEEPGROUND is accompanied by a dedicated training split, enabling systematic post-training studies on reasoning-intensive grounding. In contrast, challenging benchmarks such as Ref-Adv and GroundingME are primarily designed for evaluation and do not provide corresponding training data. PR-Bench provides both training and evaluation data, but offers relatively limited headroom for current MLLMs. DEEPGROUND-TEST also differs substantially in its visual and linguistic characteristics: the ground-truth target occupies only 1.38% of the image on average, making localization more sensitive to fine-grained spatial precision, while the referring expressions contain an average of 15.4 tokens, reflecting their greater semantic and compositional complexity.

These properties translate into substantially greater empirical difficulty. While current MLLMs achieve strong performance on conventional grounding benchmarks, even the large-scale Qwen3.5-397B obtains only 38.18 F1@mIoU on DEEPGROUND-TEST. Moreover, the difficulty is not concentrated in a single type of query but spans visual disambiguation, spatial reasoning, external knowledge, and structural understanding. DEEPGROUND-TEST therefore provides a challenging and diverse evaluation setting for studying whether advances in visual grounding arise from genuine reasoning and fine-grained perception rather than improved matching of explicit appearance cues.

Table 9: Results on DEEPGROUND-TEST. P@IoU, R@IoU, and F1@IoU are percentages. ✓and ✗denote thinking and non-thinking inference; "–" denotes an unreported status. The **bold** entries indicate the highest performance, and the underlined entries indicate the next-highest performance.

Model	Size	Think	IoU@0.5			IoU@0.7			IoU@0.9			mIoU		
			P	R	F1	P	R	F1	P	R	F1	P	R	F1
Closed models														
GPT-5.6-Sol	–	✓	67.14	63.79	65.42	48.89	46.46	47.64	15.02	14.27	14.64	41.41	39.35	40.35
Gemini-3.1-Pro	–	✓	66.82	68.68	67.74	56.01	57.57	56.78	23.30	23.95	23.62	47.55	48.87	48.20
Gemini-3.5-Flash	–	✓	**73.50**	**75.19**	**74.34**	**65.64**	**67.15**	**66.38**	**28.00**	**28.64**	**28.31**	**54.78**	**56.03**	**55.40**
Non-thinking models														
LLaVA-OneVision-2	8B	✗	10.21	9.20	9.68	2.66	2.39	2.52	0.32	0.29	0.30	3.40	3.07	3.23
InternVL3.5	8B	✗	8.81	8.81	8.81	3.45	3.45	3.45	0.38	0.38	0.38	3.53	3.53	3.53
InternVL3.5	38B	✗	12.20	12.07	12.13	6.00	5.94	5.97	2.23	2.20	2.21	5.92	5.86	5.89
PaDT-Pro	7B	✗	7.85	7.76	7.80	6.01	5.94	5.97	1.94	1.92	1.93	4.87	4.82	4.85
Rex-Omni	3B	✗	4.05	44.83	7.44	3.30	36.49	6.05	1.46	16.19	2.69	2.84	31.39	5.21
LocateAnything	3B	✗	22.29	22.22	22.25	18.73	18.68	18.71	9.41	9.39	9.40	16.27	16.23	16.25
Qwen3-VL	4B	✗	23.54	23.56	23.55	18.85	18.87	18.86	9.67	9.67	9.67	16.74	16.75	16.74
Qwen3-VL	8B	✗	30.94	31.42	31.18	25.09	25.48	25.29	12.74	12.93	12.83	21.95	22.29	22.12
Qwen3-VL	32B	✗	44.44	44.44	44.44	36.21	36.21	36.21	16.09	16.09	16.09	30.80	30.80	30.80
Qwen3-VL	235B	✗	**47.50**	**47.41**	**47.46**	**39.44**	**39.37**	**39.41**	**17.66**	**17.62**	**17.64**	**33.88**	**33.81**	**33.84**
Thinking models														
UniVG-R1	7B	✓	10.47	10.34	10.40	5.81	5.75	5.78	1.45	1.44	1.45	5.28	5.22	5.25
VisionReasoner	7B	✓	12.98	25.67	17.24	8.77	17.34	11.64	2.71	5.36	3.60	7.54	14.92	10.02
MiMo-VL	7B	✓	16.63	16.00	16.31	10.26	9.87	10.06	3.88	3.74	3.81	9.36	9.00	9.18
Keye-VL-1.5	8B	✓	3.48	4.89	4.06	0.61	0.86	0.72	0.07	0.10	0.08	1.02	1.43	1.19
GLM-4.1V	9B	✓	29.15	28.74	28.94	19.34	19.06	19.20	6.22	6.13	6.17	17.08	16.84	16.96
GLM-4.6V	106B	✓	47.84	47.70	47.77	35.93	35.82	35.88	12.87	12.84	12.85	30.76	30.67	30.71
Qwen3.5	4B	✓	36.50	36.02	36.26	28.83	28.45	28.64	9.71	9.58	9.64	24.24	23.92	24.08
Qwen3.5	9B	✓	43.85	43.01	43.42	34.57	33.91	34.24	13.67	13.41	13.54	29.31	28.75	29.02
Qwen3.5	27B	✓	56.21	55.08	55.64	45.45	44.54	44.99	17.01	16.67	16.84	38.06	37.30	37.68
Qwen3.5	397B	✓	57.38	56.99	57.18	46.48	46.17	46.32	16.30	16.19	16.24	38.31	38.06	38.18
GROUNDPO	4B	✓	51.20	51.05	51.13	43.52	43.39	43.45	19.21	19.16	19.18	37.29	37.18	37.24
GROUNDPO	9B	✓	53.54	53.54	53.54	46.17	46.17	46.17	23.95	23.95	23.95	40.12	40.12	40.12
GROUNDPO	27B	✓	**75.07**	**75.00**	**75.04**	**67.40**	**67.34**	**67.37**	**30.58**	**30.56**	**30.57**	**56.60**	**56.54**	**56.57**

B Implementation and Experimental Details

B.1 Compute and Software

We train GroundPO on Qwen3.5 models at 4B, 9B, and 27B scales, using only the training split of DeepGround. We use the repository's verl-based training stack (Sheng et al., 2024) for both GRPO and on-policy distillation, with vLLM (Kwon et al., 2023) for rollout generation. Throughout training, we retain the native Qwen3.5 chat template with thinking enabled and introduce no architectural modifications. The final method does not rely on an SFT cold-start checkpoint.

B.2 Training Configuration

Table 10: Results on GroundingME with thinking enabled. P@IoU, R@IoU, and F1@IoU are percentages. The **bold** entries indicate the highest performance, and the underlined entries indicate the next-highest performance.

	IoU@0.5			IoU@0.7			IoU@0.9			mIoU		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Qwen3.5-4B	51.30	51.49	51.40	44.86	45.02	44.94	29.74	29.85	29.80	40.09	40.24	40.16
Qwen3.5-9B	60.28	58.71	59.48	53.26	51.87	52.55	33.21	32.34	32.77	47.31	46.07	46.68
Qwen3.5-27B	65.81	65.23	65.52	59.03	58.50	58.76	37.52	37.18	37.35	52.32	51.85	52.08
Qwen3.5-397B-A17B	64.60	63.56	64.08	57.65	56.72	57.18	36.66	36.07	36.36	51.19	50.36	50.77
GroundPO-4B	54.87	54.60	54.74	49.00	48.76	48.88	32.00	31.84	31.92	44.09	43.87	43.98
GroundPO-9B	61.32	61.32	61.32	54.35	54.35	54.35	36.07	36.07	36.07	49.09	49.09	49.09
GroundPO-27B	**71.14**	**71.14**	**71.14**	**63.61**	**63.61**	**63.61**	**45.29**	**45.29**	**45.29**	**58.26**	**58.26**	**58.26**

Stage I: Policy Internalization. Stage I uses all 8,192 queries from DeepGround-Train with $G = 8$ rollouts per prompt, temperature 1.0, top-$p = 1.0$, and unrestricted top-k sampling. GRPO uses group-centered advantages without standard-deviation normalization, one PPO epoch per batch, and token-mean loss aggregation. Prompts exceeding 8,192 tokens are filtered, and on-policy generation is capped at 2,048 tokens. For the 4B and 9B models, we use a batch size of 64 and a PPO mini-batch size of 64, while for the 27B model, we use a batch size of 32 and a PPO mini-batch size of 8. All models are trained for one epoch with a learning rate of 1×10^{-6}. The clipping coefficients are set to $\epsilon_{\text{low}} = 0.2$ and $\epsilon_{\text{high}} = 0.3$.

Table 11: Results on PR-Bench with thinking enabled. P@IoU, R@IoU, and F1@IoU are percentages. The **bold** entries indicate the highest performance, and the underlined entries indicate the next-highest performance.

	IoU@0.5			IoU@0.7			IoU@0.9			mIoU		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Qwen3.5-4B	75.32	76.28	75.79	65.76	66.60	66.18	39.97	40.48	40.22	58.82	59.57	59.19
Qwen3.5-9B	75.12	77.36	76.22	65.95	67.92	66.92	40.09	41.28	40.67	58.83	60.59	59.70
Qwen3.5-27B	76.32	79.82	78.03	67.03	70.10	68.53	40.01	41.84	40.90	59.48	62.20	60.81
Qwen3.5-397B-A17B	69.06	79.90	74.08	60.93	70.50	65.37	36.84	42.62	39.52	54.28	62.80	58.23
GroundPO-4B	76.01	75.96	75.98	66.48	66.44	66.46	40.16	40.14	40.15	59.30	59.27	59.28
GroundPO-9B	77.09	77.06	77.08	67.61	67.58	67.59	42.06	42.04	42.05	60.74	60.72	60.73
GroundPO-27B	**78.92**	**78.90**	**78.91**	**69.15**	**69.14**	**69.15**	**44.47**	**44.46**	**44.46**	**62.72**	**62.71**	**62.71**

Stage II: Policy Distillation. Stage II initializes the student from the corresponding Qwen3.5 base model and freezes the same-scale Stage I checkpoint as the teacher. For the 4B and 27B models, we use 4,800 hard examples selected from DeepGround-Train based on Qwen3.5-4B predictions with IoU < 0.75, with both the batch size and PPO mini-batch size set to 32. For the 9B model, we use all 8,192 examples from DeepGround-Train, with both sizes set to 64. The student receives the original image and unprivileged prompt, while the teacher receives the same image, the same student-generated response prefix, and textual information about similar visual entity nodes. Following Vision-OPD (Yuan et al., 2026), we set $K = 100$ for top-K distillation. The teacher remains fixed throughout training, while reward computation and the GRPO fallback are disabled, leaving the top-100-plus-tail Jensen–Shannon distillation loss described in Sec. 4.3 as the sole optimization signal. As in Stage I, the maximum on-policy generation length is 2,048 tokens and training lasts for one epoch. Other settings follow Stage I, with 10 warmup steps for the learning-rate schedule.

Table 12: Results on Ref-Adv with thinking enabled. P@IoU, R@IoU, and F1@IoU are percentages. The **bold** entries indicate the highest performance, and the underlined entries indicate the next-highest performance.

Model	IoU@0.5			IoU@0.7			IoU@0.9			mIoU		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Qwen3.5-4B	60.12	59.54	59.83	51.72	51.23	51.47	29.35	29.07	29.21	45.42	44.98	45.20
Qwen3.5-9B	65.72	64.62	65.17	56.54	55.60	56.07	33.75	33.19	33.47	50.16	49.33	49.74
Qwen3.5-27B	67.40	67.34	67.37	59.33	59.28	59.31	34.27	34.24	34.25	52.16	52.12	52.14
Qwen3.5-397B-A17B	65.04	67.78	66.38	58.07	60.51	59.26	32.94	34.33	33.62	50.13	52.24	51.17
GROUNDPO-4B	64.77	64.71	64.74	56.88	56.83	56.86	31.11	31.09	31.10	49.39	49.34	49.36
GROUNDPO-9B	66.05	65.94	65.99	57.63	57.53	57.58	**34.82**	**34.76**	**34.79**	51.06	50.97	51.02
GROUNDPO-27B	**69.44**	**69.44**	**69.44**	**61.30**	**61.30**	**61.30**	34.59	34.59	34.59	**53.42**	**53.42**	**53.42**

B.3 PROMPT TEMPLATES

Conventional baseline prompt. For the Qwen3.5 base models, we follow the original grounding instruction provided in the Qwen3.5 technical report (Qwen Team, 2026a), as shown in Fig. 7. For all reproduced baselines, we use the official prompts or inference templates recommended by the corresponding papers or repositories, rather than enforcing a unified prompt across models. This avoids introducing prompt-format mismatch when comparing heterogeneous grounding systems.

Table 13: Results on RefCOCO with thinking enabled. P@IoU, R@IoU, and F1@IoU are percentages. The **bold** entries indicate the highest performance, and the underlined entries indicate the next-highest performance.

Model	IoU@0.5			IoU@0.7			IoU@0.9			mIoU		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Qwen3.5-4B	85.37	87.73	86.54	76.38	78.49	77.42	51.41	52.84	52.11	68.87	70.78	69.81
Qwen3.5-9B	86.65	88.52	87.57	78.90	80.61	79.75	55.35	56.55	55.94	71.22	72.76	71.98
Qwen3.5-27B	87.61	90.48	**89.02**	79.93	82.55	81.22	57.81	59.71	58.74	72.70	75.09	73.87
Qwen3.5-397B-A17B	84.58	**91.95**	88.11	78.17	**84.99**	**81.44**	57.81	**62.85**	**60.22**	71.36	**77.58**	**74.34**
GROUNDPO-4B	86.09	86.03	86.06	77.25	77.19	77.22	53.21	53.17	53.19	70.21	70.15	70.18
GROUNDPO-9B	86.50	86.61	86.55	78.91	79.02	78.96	56.29	56.37	56.33	71.71	71.81	71.76
GROUNDPO-27B	**88.16**	88.09	88.12	**80.79**	80.73	80.76	**58.34**	58.29	58.32	**73.33**	73.27	73.30

Prompt in Policy Internalization. The Stage-I prompt is shown in Fig. 8. In addition to the original image and grounding query, it provides textual privileged information derived from similar visual entity nodes. Each privileged node contains only the entity name, a short description, and its target/similar role; no coordinates or visual annotations are included. The target-only privilege variants evaluated in Sec. 5.3 use the same prompt structure, with only the node content and corresponding descriptive text replaced. When no similar-node record is available, the privileged context is omitted and the input reduces to the unprivileged prompt. This occurs for 2 of the 8,192 training examples and 112 of the 1,044 benchmark examples.

Table 14: Results on RefCOCOg with thinking enabled. P@IoU, R@IoU, and F1@IoU are percentages. The **bold** entries indicate the highest performance, and the underlined entries indicate the next-highest performance.

Model	IoU@0.5			IoU@0.7			IoU@0.9			mIoU		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Qwen3.5-4B	87.25	88.76	88.00	80.65	82.05	81.34	56.27	57.24	56.75	72.49	73.74	73.11
Qwen3.5-9B	88.48	89.67	89.07	82.13	83.23	82.68	59.00	59.79	59.39	74.33	75.33	74.82
Qwen3.5-27B	89.24	91.40	**90.31**	83.00	85.00	83.99	61.54	63.03	62.28	75.73	77.56	76.64
Qwen3.5-397B-A17B	87.87	**92.41**	90.08	82.48	**86.74**	**84.56**	**62.96**	**66.22**	**64.55**	75.70	**79.61**	**77.61**
GROUNDPO-4B	87.26	86.70	86.98	80.71	80.20	80.46	57.90	57.53	57.71	73.04	72.57	72.80
GROUNDPO-9B	88.26	88.17	88.22	81.63	81.55	81.59	59.29	59.23	59.26	74.36	74.28	74.32
GROUNDPO-27B	**89.63**	89.62	89.62	**83.70**	83.69	83.70	62.59	62.58	62.58	**76.37**	76.36	76.37

Prompt in Policy Distillation. During Stage II, the teacher uses the same privileged prompt as in Fig. 8, while the student receives no privileged information and follows the prompt shown in Fig. 9. This design ensures that teacher and student differ only in the availability of privileged context, while sharing the same image, query, and response format. Unless otherwise specified, the student prompt is also used for standard evaluation and ablation settings, including the privilege-free evaluations in Tab. 3. This keeps the student-side training and inference conditions consistent.

Prompts for Visual Privilege and Transition Instructions. We additionally investigate visually privileged teachers. Under $\mathcal{P}^V$, the ground-truth region is visually marked in the image, while the textual prompt remains identical to the student prompt in Fig. 9. Since such visual cues can induce privilege-specific expressions such as "red bounding box," we further construct $\mathcal{P}^{V,T}$ by augmenting the textual prompt with a transition instruction, following (Zhao et al., 2026), to reduce the discrepancy between privileged and unprivileged generation behavior. The corresponding prompt is shown in Fig. 10. These variants are used only to analyze the effect of different privilege designs and are not part of the final GROUNDPO configuration.

B.4 EVALUATION DETAILS OF PRIVSHIFT

In Fig. 2 and Tab. 4, we evaluate the two manifestations of PrivShift: privilege leakage and verbosity drift. Privilege leakage is evaluated using Qwen3.5-397B-A17B as the judge, with the prompt shown in Fig. 11. Given the original image, grounding query, and model response, the judge determines whether the response contains information leaked from the privileged context, outputting 1 if leakage is detected and 0 otherwise. For verbosity drift, we measure the generated response length in tokens. To allow models to complete their responses whenever possible and minimize truncation caused by the generation budget, we set the maximum generation length to 8,192 tokens for the experiments in Tab. 4. Notably, all responses from $\pi_{\text{OPD}}^{\mathcal{P}^T}$ reach this limit and are truncated, resulting in an average output length of 8,192 tokens.

Judging Prompt for Privilege Leakage

```
You are given the ORIGINAL clean image (no annotations drawn on it), a grounding QUERY, and a
model's RAW OUTPUT (its chain-of-thought inside <think></think> plus a final JSON answer).\n\n
Decide whether the RAW OUTPUT hallucinates / leaks PRIVILEGED information that is NOT present in
the original image, such as:\n  - referring to \"the red bounding box\", \"the red rectangle\", \"the
highlighted/marked region\", \"the annotation/overlay\";\n  - claiming a target is \"already highlighted
/ boxed / marked\";\n  - explicitly copying or citing provided ground-truth coordinates as if they were
given.\n\nNote: mentioning a genuinely red-colored object that truly exists in the scene (e.g. a red car)
is NOT leakage. Only flag references to annotation artifacts (boxes/highlights/marks/given coordinates)
that are not real image content.\n\nQUERY:\n{query}\n\nRAW OUTPUT:\n{raw_output}\n\nAnswer
with a SINGLE character only: 1 if it hallucinates/leaks privileged annotations, 0 if it is a normal
output. Output just 0 or 1, nothing else.
```

Figure 11: Judging prompt for privilege leakage.

C DETAILED EXPERIMENTAL RESULTS

C.1 BENCHMARKS

We provide detailed grounding results on seven benchmarks covering reasoning-intensive, conventional, and fine-grained referring expression grounding. DEEPGROUND-TEST (Tab. 9) is our reasoning-intensive benchmark and contains five complementary tracks: contrastive, spatial, commonsense, conceptual, and structural grounding. These tracks evaluate whether a model can identify the intended target through implicit comparison, spatial relations, commonsense knowledge, domain-specific concepts, and structured visual cues, rather than relying only on salient appearance attributes.

We further evaluate on three recent reasoning-oriented grounding benchmarks. GroundingME (Li et al., 2026b) (Tab. 10) evaluates challenging grounding along discriminative, spatial, limited-visibility, and rejection dimensions. Ref-Adv (Dong et al., 2026) (Tab. 12) focuses on adversarial referring expressions involving hard visual distractors, complex descriptions, and frequent negation, placing greater emphasis on visual disambiguation and reasoning. PR-Bench (Gao et al., 2026b) (Tab. 11) studies the perception–reasoning gap through both fine-grained visual-cue perception and compositional reasoning, covering attributes, positions, interactions, relations, commonsense, and rejection cases. Together with DEEPGROUND-TEST, these benchmarks provide a broad evaluation of grounding under diverse visual and semantic reasoning requirements.

We additionally report results on the widely used RefCOCO family to examine whether improvements on reasoning-intensive grounding preserve performance on conventional referring expression

Table 15: Results on RefCOCOm with thinking enabled. P@IoU, R@IoU, and F1@IoU are percentages. The **bold** entries indicate the highest performance, and the underlined entries indicate the next-highest performance.

Model	IoU@0.5			IoU@0.7			IoU@0.9			mIoU		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Qwen3.5-4B	46.66	47.27	46.96	36.60	37.08	36.84	21.55	21.84	21.69	33.32	33.76	33.53
Qwen3.5-9B	48.83	49.29	49.06	38.71	39.07	38.89	23.48	23.70	23.59	35.28	35.62	35.45
Qwen3.5-27B	49.77	51.17	50.46	39.54	40.65	40.09	24.96	25.67	25.31	36.24	37.26	36.75
Qwen3.5-397B-A17B	43.62	47.30	45.39	35.70	38.71	37.14	23.72	25.72	24.68	32.83	35.60	34.16
GROUNDPO-4B	56.13	56.00	56.06	43.20	43.10	43.15	24.34	24.28	24.31	39.22	39.13	39.18
GROUNDPO-9B	57.88	57.84	57.86	45.45	45.42	45.43	26.13	26.11	26.12	41.16	41.13	41.15
GROUNDPO-27B	**61.00**	**60.99**	**60.99**	**49.06**	**49.06**	**49.06**	**28.43**	**28.43**	**28.43**	**44.06**	**44.05**	**44.05**

comprehension. RefCOCO (Yu et al., 2016) (Tab. 13) contains relatively concise referring expressions collected in an interactive setting, while RefCOCOg (Mao et al., 2016) (Tab. 14) contains longer and more compositional free-form expressions collected in a non-interactive setting. We further evaluate on RefCOCOm (Wang et al., 2024) (Tab. 15), a multi-granularity extension that introduces part-level references and jointly evaluates object- and part-level grounding, making it particularly suitable for assessing fine-grained localization.

Since our study focuses on localization rather than rejection capability, we exclude rejection-related examples from GroundingME and PR-Bench. We use the testB split for RefCOCO and RefCOCOm, and the test split for RefCOCOg.

C.2 EVALUATION PROTOCOL

Evaluation is conducted with vLLM servers using bfloat16 precision. Inference is greedy with temperature 0 and a maximum generation budget of 4,096 tokens. Except for the non-thinking results reported in Tab. 1, Tab. 9 and Tab. 5, all experiments are conducted with thinking enabled.

For each grounding benchmark, we report precision, recall, and F1 at IoU thresholds of 0.5, 0.7, and 0.9, together with the corresponding mIoU metrics. Predicted and ground-truth boxes are greedily matched one-to-one in descending IoU order before micro-aggregating precision, recall, and F1. The reported mIoU metrics average results over ten IoU thresholds from 0.50 to 0.95 in increments of 0.05, providing an aggregate measure of localization quality across different levels of spatial precision.

The parser supports normalized, grid-1000, and pixel-space coordinates, with the coordinate format determined from the predicted coordinate magnitude. For thinking-mode outputs, bounding boxes are parsed only from content appearing after the final `</think>` tag. Responses truncated before this tag are treated as containing no valid prediction. This prevents intermediate coordinates mentioned in incomplete reasoning traces from being incorrectly counted as final localization results and ensures consistent evaluation across models with different reasoning lengths.

C.3 DETAILED RESULTS

Tab. 9 shows that GROUNDPO consistently improves the corresponding base models across different IoU thresholds, indicating gains in both target identification and precise localization. Improvements at IoU@0.5 and IoU@0.7 reflect a stronger ability to identify and localize the intended entity, while the gains remain pronounced under the substantially stricter IoU@0.9 criterion. High-IoU evaluation is particularly sensitive to localization granularity: identifying the correct semantic object is insufficient if the predicted box is overly coarse, includes irrelevant surrounding regions, or corresponds to the whole object when the query refers to a finer-grained part. The consistent improvement at IoU@0.9 therefore suggests that GROUNDPO enhances not only reasoning about *which* entity should be grounded, but also understanding of *which visual extent* precisely corresponds to the query.

The results on GroundingME, Ref-Adv, and PR-Bench further demonstrate that these gains are not specific to the construction or distribution of DEEPGROUND-TEST. These benchmarks introduce different sources of difficulty, including visually similar distractors, complex relational descriptions, negation, fine-grained perceptual cues, and compositional reasoning. Consistent improve-

Representative Outputs



Q: Locate every object that matches the description \"the tourist wearing a visor among the four friends sitting in the desert\" in the image. First think about the reasoning process step by step as an internal monologue enclosed within <think> </think> tags, then output the localization result as exactly one JSON code block with the following format:\n```json\n[\n {\"bbox_2d\": [x1, y1, x2, y2], \"label\": \"a concise description of the target\"}\n]\n```

Qwen3.5-4B: <think>\nGot it, let's find the tourist with the visor. Looking at the image, the fourth person from the left is wearing a visor. Let's check the bounding box. The coordinates for that person would be [617, 490, 945, 997].\n</think>\n\n```json\n[\n\t{\"bbox_2d\": [617, 490, 945, 997], \"label\": \"the tourist wearing a visor among the four friends sitting in the desert\"}\n]\n```

GroundPO-4B: <think>\nThe user wants to identify a specific object based on the description \"a person wearing a visor\".\nThe visual content shows four young men sitting in a desert landscape under a starry sky. The description explicitly mentions a person wearing a visor.\nThe subject fits well within the description.\nThe object itself is the person fitting the description.\n\n</think>\n\n```json\n[\n {\n \"bbox_2d\": [\n 642, 490, 949, 995],\n \"label\": \"person wearing visor\"}\n]\n```

π_T: <think>\nThe user wants to identify a specific object based on the description \"a tourist wearing a visor among the four friends sitting in the desert\".\nThe subject fits well within the description of a person wearing a visor/headband in a desert setting.\nThe visual content is clearly visible.\n\n</think>\n\n```json\n[\n {\n \"bbox_2d\": [\n 645, 490, 949, 995],\n \"label\": \"tourist wearing visor\"}\n]\n```

$\pi_{OPD}^{\mathcal{P}^V}$: <think>\nThe user wants me to identify the object within the red bounding box(es) based on the provided image and description.\n\n1. **Analyze the Request:**\n * **Description:** \"the tourist wearing a visor among the four friends sitting in the desert\"\n * **Visual Cue:** \"The target region(s) are already highlighted with red rectangle bounding box(es) in the image.\"\n * **Image Content:** The image shows a group of people sitting in a desert-like setting. There are four people visible. One person is wearing a visor.\n * [Verbosity Drift] \n</think>\n\n```json\n[\n {\"bbox_2d\": [610, 480, 940, 980], \"label\": \"the tourist wearing a visor among the four friends sitting in the desert\"}\n]\n```

$\pi_{OPD}^{\mathcal{P}^{V,T}}$: <think>\nThe user wants me to identify the object described as \"the tourist wearing a visor among the four friends sitting in the desert\".\n1. **Analyze the image:** I see four young men sitting in a desert landscape at night under a starry sky.\n2. **Identify the specific person:**\n * Person 1 (far left): Wearing a white and blue jacket, no visor.\n * Person 2 (second from left): Wearing a dark jacket, no visor.\n * Person 3 (second from right): Wearing an orange hoodie with a smiley face, no visor.\n * Person 4 (far right): Wearing a black jacket and a dark visor/headband. This matches the description \"wearing a visor\".\n3. **Check for highlighting:** The prompt states that the target region is already highlighted with a red rectangle. Looking at the image, there is indeed a red rectangle around the person on the far right.\n [Verbosity Drift] \n</think>\n\n```json\n[\n {\"bbox_2d\": [610, 480, 940, 980], \"label\": \"the tourist wearing a visor\"}\n]\n```

$\pi_{OPD}^{\mathcal{P}^T}$: <think>\nThe user wants to identify a target entity from the image based on a provided hierarchical entity-tree context and then ground its location from the image.\n\n**1. Analyze the Input:**\n* **Image:** Shows four young men sitting in a desert at night under a starry sky.\n * Person 1 (Left): White jacket with blue stripes, barefoot.\n * Person 2 (Middle-Left): Blue jacket, blue jeans, bright green socks.\n * Person 3 (Middle-Right): Orange hoodie with a smiley face logo, red socks.\n * Person 4 (Right): Black jacket, black pants with white stripes, red socks, wearing a headband/visor.\n* **Context:** The prompt provides a specific JSON structure describing the target's granularity.\n * [Verbosity Drift] \n</think>\n\n```json\n[\n {\"bbox_2d\": [610, 480, 940, 980], \"label\": \"the tourist wearing a visor\"}\n]\n```

Figure 12: Representative outputs on Qwen3.5-4B. *[Verbosity Drift]* denotes excessively long generations that are omitted from the figure. Naive OPD with $\pi_{\text{OPD}}^{\mathcal{P}^V}$, $\pi_{\text{OPD}}^{\mathcal{P}^{V,T}}$, or $\pi_{\text{OPD}}^{\mathcal{P}^T}$ exhibits both privilege leakage and severe verbosity drift. In contrast, the internalized teacher π_T and GROUNDPO-4B produce outputs much closer to those of the original Qwen3.5-4B, with substantially less privilege-specific content and more stable response lengths.

ments across them indicate that the grounding policy learned by GROUNDPO transfers to heterogeneous forms of reasoning-intensive localization rather than overfitting to a single query-generation pipeline.

On conventional referring expression benchmarks, RefCOCO and RefCOCOg, GROUNDPO primarily improves precision while maintaining recall close to that of the corresponding base mod-

els. This indicates that reasoning-intensive post-training improves target discrimination and reduces spurious predictions without substantially compromising the original grounding coverage. Thus, GROUNDPO largely preserves conventional grounding performance while improving precision.

The improvement in fine-grained localization is further supported by RefCOCOm. Unlike conventional grounding datasets that predominantly refer to complete objects, RefCOCOm additionally contains part-level references and therefore requires the model to distinguish between different semantic granularities of the same visual entity. The strong improvement on RefCOCOm is consistent with the gains observed under the strict IoU@0.9 criterion on DEEPGROUND-TEST: both results indicate that GROUNDPO improves sensitivity to the precise visual extent implied by a referring expression. Taken together, the results across reasoning-intensive, conventional, and multi-granularity benchmarks demonstrate improvements in target discrimination and fine-grained spatial grounding.

C.4 QUALITATIVE OUTPUT ANALYSIS

Fig. 12 provides a representative comparison of the output behaviors of Qwen3.5-4B, naive OPD variants, the internalized teacher π_T, and GROUNDPO-4B. The original Qwen3.5-4B produces a concise reasoning trace and directly identifies the target. In contrast, naive OPD exhibits clear PrivShift under all three privilege designs, showing that directly distilling from a privilege-conditioned teacher transfers not only useful grounding knowledge but also generation behaviors induced by the privileged input.

With visual privilege $\mathcal{P}^V$, the student explicitly refers to the "red bounding box," directly revealing information that is available only to the privileged teacher. Introducing the transition instruction in $\mathcal{P}^{V,T}$ does not fully eliminate this behavior: the student still states that the target is "already highlighted with a red rectangle," indicating persistent privilege leakage. Textual privilege $\mathcal{P}^T$ avoids references to visual annotations but exhibits the same failure mode in another form, with the student explicitly mentioning the provided hierarchical entity-tree context and JSON structure. These examples show that privilege leakage is not specific to a particular privilege modality; rather, direct OPD can cause the student to imitate privilege-dependent generation patterns that are unavailable at inference time.

Naive OPD also exhibits pronounced verbosity drift. All three directly distilled students generate substantially longer reasoning traces than the original policy, with omitted portions marked as *[Verbosity Drift]* in Fig. 12. This suggests that the response-length characteristics of the privilege-conditioned teacher are also inherited during distillation. Such excessively long reasoning not only introduces unnecessary generation overhead but can also exhaust the generation budget, causing the response to be truncated before the final structured prediction is produced.

In contrast, after *Policy Internalization*, π_T produces a compact reasoning trace without explicitly referring to privileged information while still accurately localizing the target. The final GROUNDPO-4B student exhibits a similarly stable output pattern and remains much closer to the original Qwen3.5-4B in both reasoning style and response length. Importantly, this behavioral stabilization is accompanied by improved grounding performance rather than merely shorter responses. These qualitative examples therefore illustrate the role of *Policy Internalization* in mitigating PrivShift: privileged knowledge is first internalized into the teacher policy instead of being directly exposed as a generation condition, allowing subsequent OPD to transfer the improved grounding policy while substantially suppressing both privilege leakage and verbosity drift.