REWARD-GROUNDED VISUAL EVIDENCE: SHAPING TOKEN-LEVEL ATTENTION IN MULTIMODAL REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning (RL) enhances visual reasoning in multimodal large language models (MLLMs), yet existing methods rely primarily on holistic, outcome rewards. This creates a mismatch with visual grounding, which demands spatial and token specificity. We reveal that while outcome RL improves benchmark scores, target attention remains largely unaligned. Furthermore, visual sensitivity varies sharply across generation, surging on decisive tokens like coordinates while remaining negligible on auxiliary text, rendering coarse scalar rewards insufficient for critical steps. To bridge this gap, we propose **V**isual-credit **I**nduced **S**elf-attention **T**arget **A**lignment (**VISTA**) to provide selective, token-level visual supervision within policy optimization. Rather than intervening uniformly, **VISTA** derives gradient-based visual credits to construct targeted attention distributions on decisive tokens, regularizing cross-modal attention through intrinsically derived targets while focusing on decisive steps. Across grounding benchmarks, **VISTA** lifts 3B accuracy from 68.5% to 70.3% and 7B accuracy from 69.8% to 72.1%, with gains of +4.3 and +4.8 on LISA-G. Generalizing beyond grounding, it secures a top 54.7% average on multimodal reasoning among 7B architectures, rivaling the InternVL2.5-38B model. Mechanistically, **VISTA** concentrates internal cross-modal attention directly onto target regions while suppressing spurious visual noise, fostering stronger alignment between policy updates and decisive visual evidence.

1 INTRODUCTION

Multimodal large language models (MLLMs) have advanced rapidly in visual perception and reasoning through stronger vision-language backbones and instruction tuning (Bai et al., 2025). Recently, reinforcement learning (RL) (Schulman et al., 2017; Shao et al., 2024) has emerged as an effective post-training paradigm. Extending R1-style optimization from language models (Shao et al., 2024; DeepSeek-AI, 2025) to multimodal settings has yielded substantial gains in grounding and mathematical reasoning (Shen et al., 2025; Zhan et al., 2026; Wang et al., 2025a; Yang et al., 2025), demonstrating the power of outcome-based rewards in improving multimodal policies.

Despite their success, standard multimodal RL algorithms optimize policies purely for task correctness, rarely constraining how visual inputs are processed during generation. As Figure 1 illustrates, while Outcome RL lifts grounding IoU above the base model (left), its target visual attention distribution remains superimposed on the baseline across percentiles (right). This disconnect reveals that gains in task correctness do not inherently improve visual allocation toward target regions for decision tokens. Scalar rewards simply reinforce favorable outputs without guiding the policy where to look, leaving internal visual attention unguided.

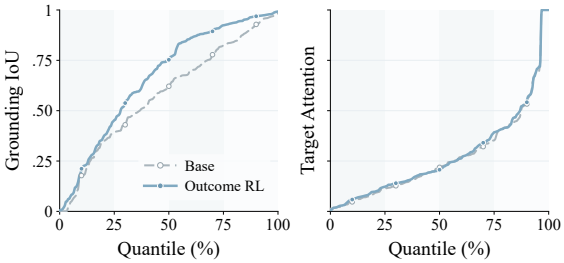


Figure 1: Outcome RL improves grounding IoU over the base model, while the target attention distribution remains largely unchanged.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

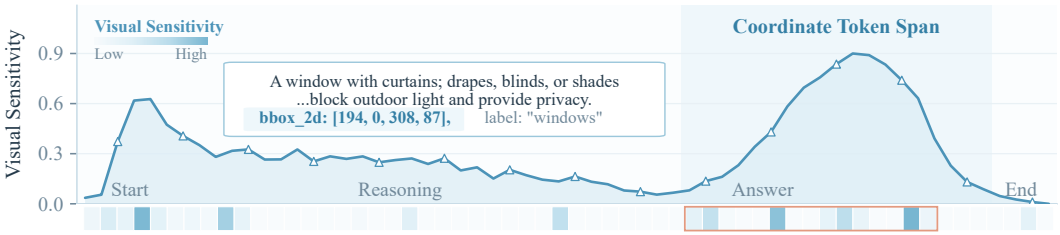


Figure 2: **Token-specific visual sensitivity across generation.** Visual reliance centers predominantly on spatial decision tokens over descriptive context. This disparity motivates **VISTA** to deploy token-level visual supervision selectively rather than uniformly throughout the sequence.

This phenomenon is further complicated because visual reliance is non-uniform across the generated sequence. As Figure 2 shows, visual sensitivity spikes at grounding tokens like bounding box coordinates, but remains negligible on auxiliary text. A single response-level reward disperses across this trajectory, making it difficult to separate critical decisions from surrounding context. Consequently, standard multimodal RL introduces an architectural disparity between its holistic reward signal and the spatially, token-localized nature of visual supervision. Overcoming this limitation calls for attributing individual tokens to pertinent visual cues alongside outcome optimization.

Building on this insight, we introduce **V**isual-credit **I**nduced **S**elf-attention **T**arget **A**lignment (**VISTA**), a framework providing token-level visual attention supervision for multimodal RL. As illustrated in Figure 3, **VISTA** estimates gradient-based visual credits for critical tokens to construct target attention distributions, directly regularizing text-to-image attention without external region annotations. Supported by the trajectory patterns in Figure 2, this supervision avoids uniform intervention across the sequence and concentrates primarily on key terminal tokens, such as bounding box coordinates in spatial tasks. In this way, **VISTA** complements coarse rewards with fine-grained visual guidance while preserving computational efficiency.

We evaluate **VISTA** across 13 multimodal grounding and reasoning benchmarks. In grounding evaluations, we consistently outperform baselines across scales, elevating 3B average accuracy to 70.3% from 68.5% and advancing the 7B variant to 72.1% over 69.8%. These gains are particularly prominent on LISA, reaching 67.2% (+4.3) at 3B and 72.6% (+4.8) at 7B over the baseline. When applied to multimodal reasoning, **VISTA** similarly demonstrates strong effectiveness, attaining a top 54.7% average among 7B models and rivaling the much larger InternVL2.5-38B (54.5%). Mechanistically, it focuses cross-modal attention directly on target regions while reducing spurious visual noise, which grounds policy updates in task-relevant visual evidence.

Our contributions are as follows:

- We identify two limitations in multimodal RL, namely insufficient target visual allocation despite task gains and token-dependent visual reliance, motivating our visual credit formulation.
- We propose **VISTA**, which derives visual targets from reward credits to regularize token-conditioned attention within standard multimodal RL pipelines.
- Our evaluations across 13 benchmarks demonstrate consistent gains, while mechanistic analyses confirm heightened target attention and reduced spurious noise.

2 Related Work

Reinforcement learning for multimodal reasoning. Verifiable reward-based RL has become a cornerstone for complex reasoning in LLMs, utilizing group advantages (Shao et al., 2024) and outcome-level feedback for self-correction (DeepSeek-AI, 2025). Recent multimodal methods adapt this paradigm to vision-language domains, including Vision-R1 (Huang et al., 2026), VLM-R1 (Shen et al., 2025), R1-OneVision (Yang et al., 2025), VL-Rethinker (Wang et al., 2025a), ThinkLite-VL (Wang et al., 2025c), and DIVA-GRPO (Gao et al., 2026). Yet sequence-level rewards alone fail to decouple genuine grounding from language shortcuts. We introduce a visual-credit mechanism that links RL advantages to token-level visual utilization during policy optimization.

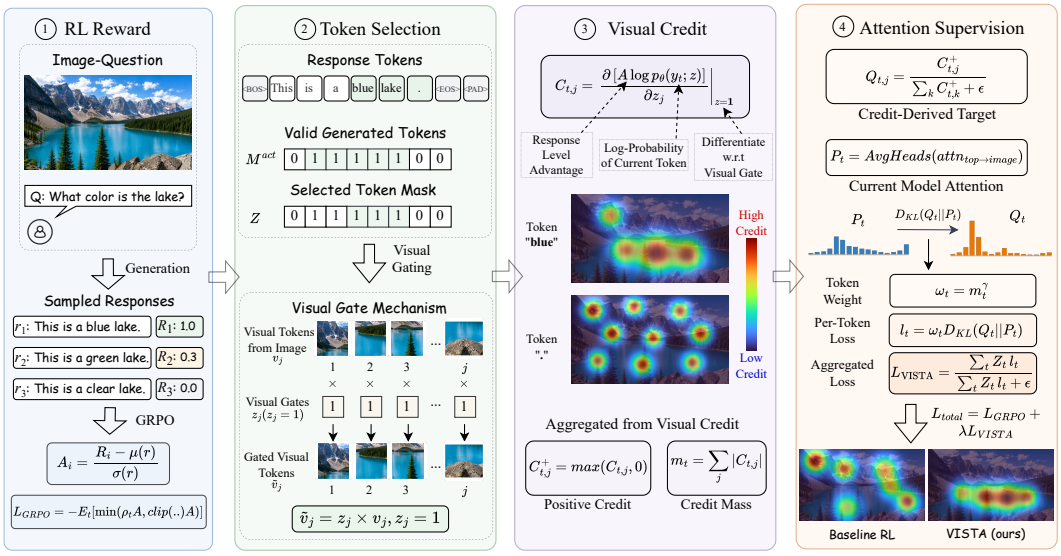


Figure 3: Overview of **VISTA**. **VISTA** derives token-level visual targets from reinforcement feedback to regularize internal attention during policy optimization.

Fine-grained rewards and process supervision. To refine coarse outcome feedback, prior studies introduce step-level multimodal rewards (Zhang et al., 2025; Wang et al., 2025b; Du et al., 2025) or perception PRMs to penalize errors (Xiao et al., 2026; Min et al., 2026), while others use preference signals to curb hallucinations (Sun et al., 2024; Yu et al., 2024). However, these dense schemes require external reward models, judges, or annotations. In contrast, our approach derives visual credit directly from the policy's first-order sensitivities, eliminating auxiliary scoring pipelines.

Visual grounding, attention routing, and attribution. Grounding reasoning in visual evidence rather than language priors is vital for VLM reliability. Recent works inspect internal attention patterns (Kang et al., 2025a;b) or steer attention toward informative visual regions (Yin et al., 2025; An et al., 2025; Zhao et al., 2026; Wu et al., 2026). Nevertheless, raw attention alone can reflect diffuse activations rather than direct task relevance (Selvaraju et al., 2017; Jain & Wallace, 2019). Rather than relying on uncalibrated attention, we leverage gradient-based attribution to derive reward-aware visual credits that regularize cross-modal attention during GRPO (Shao et al., 2024).

3 PRELIMINARIES

Multimodal autoregressive generation. For the b-th image–text input $x_b = (I_b, q_b)$, a vision-language model first represents the image as a sequence of visual tokens:

$$V_b = (v_{b,1}, \ldots, v_{b,N_b}), \qquad v_{b,j} \in \mathbb{R}^d, \tag{1}$$

where N_b is the number of visual tokens and $\mathcal{V}_b = \{1, \ldots, N_b\}$ denotes their valid positions. The model then autoregressively generates a response $y_b = (y_{b,1}, \ldots, y_{b,T_b})$ according to:

$$y_{b,t} \sim \pi_\theta(\cdot \mid I_b, q_b, y_{b,<t}), \tag{2}$$

where π_θ denotes the policy model, T_b is the response length, and $y_{b,<t}$ denotes the previously generated tokens.

Reinforcement learning with response-level feedback. During policy optimization, each sampled response y_b receives a scalar reward R_b and an advantage A_b. In GRPO (Shao et al., 2024), responses are grouped by input, and the advantage for the g-th response of the i-th input is:

$$A_{i,g} = \frac{R_{i,g} - \mu_i}{\sigma_i + \epsilon}, \qquad \mu_i = \frac{1}{G} \sum_{g=1}^{G} R_{i,g}, \tag{3}$$

where σ_i is the reward standard deviation within the group, G is the group size, and ϵ prevents division by zero. We denote this response-level advantage as A_b hereafter, which weights token log-probabilities in the RL objective.

Cross-modal attention from text to image. For the b-th sample, let $h_{b,t-1}$ denote the hidden state used to predict token $y_{b,t}$, and let $h_{b,j}^{\text{img}}$ denote the hidden state at visual token position $j \in \mathcal{V}_b$. Using the final language layer, we form an isolated visual attention distribution by normalizing query and key affinities over valid visual token positions. For attention head h (with projection weights $W_Q^{(h)}, W_K^{(h)}$), the attention from the generated token to visual token j is:

$$a_{b,t,j}^{(h)} = \text{softmax}_{j \in \mathcal{V}_b} \left(\frac{\left\langle W_Q^{(h)} h_{b,t-1}, W_K^{(h)} h_{b,j}^{\text{img}} \right\rangle}{\sqrt{d_h}} \right). \tag{4}$$

Averaging over the H attention heads yields the visual-only attention probability:

$$P_{b,t,j} = \frac{1}{H} \sum_{h=1}^{H} a_{b,t,j}^{(h)}, \tag{5}$$

forming the complete visual attention distribution $P_{b,t} = (P_{b,t,1}, \ldots, P_{b,t,N_b})$ over all visual tokens for $y_{b,t}$. These response-level learning signals and token-level visual attention distributions provide the necessary foundation for **VISTA**.

4 METHOD

As shown in Figure 3, **VISTA** introduces a token-level visual objective for multimodal RL. Guided by response-level advantages, we compute visual credits derived from policy sensitivities for output tokens to supervise cross-modal attention, transforming coarse rewards into fine-grained guidance that grounds policy updates in relevant visual evidence.

4.1 TOKEN-SPECIFIC VISUAL CREDIT

For the b-th image–text input $x_b = (I_b, q_b)$, the policy π_θ generates a response:

$$y_b = (y_{b,1}, \ldots, y_{b,T_b}), \qquad y_{b,t} \sim \pi_\theta(\cdot \mid I_b, q_b, y_{b,<t}), \tag{6}$$

where I_b and q_b denote the input image and textual instruction, respectively, and T_b denotes the total response sequence length.

Let the visual encoder and projector represent I_b as:

$$V_b = (v_{b,1}, \ldots, v_{b,N_b}), \qquad v_{b,j} \in \mathbb{R}^d, \tag{7}$$

where N_b is the number of visual tokens and d is their hidden dimension.

To estimate the contribution of each visual token, we introduce a temporary differentiable gate:

$$\widetilde{v}_{b,j} = z_{b,j} v_{b,j}, \qquad z_{b,j} = 1. \tag{8}$$

The gate $z_{b,j}$ is applied to the projected visual embedding solely for credit estimation, without serving as a trainable parameter.

Let A_b denote the response-level advantage assigned to y_b. The visual credit of visual token $v_{b,j}$ for generating token $y_{b,t}$ is defined as:

$$C_{b,t,j} = \left. \frac{\partial \left[\text{sg}(A_b) \log \pi_\theta \left(y_{b,t} \mid I_b, q_b, y_{b,<t}; z \right) \right]}{\partial z_{b,j}} \right|_{z=\mathbf{1}}, \tag{9}$$

where $\text{sg}(\cdot)$ denotes stop-gradient. Thus, $C_{b,t,j}$ is the signed first-order sensitivity of the token log-probability weighted by the advantage to a multiplicative perturbation of visual token j.

4.2 CREDIT-DERIVED VISUAL TARGET

We retain the positive component of the visual credit:

$$C^{+}_{b,t,j} = \max(C_{b,t,j}, 0), \tag{10}$$

and let $\mathcal{S}^{K_v}_{b,t}$ denote the indices of the K_v largest positive credits among valid visual tokens. We construct a sparse, scale-invariant token-specific visual target as:

$$Q_{b,t,j} = \frac{\mathbb{I}[j \in \mathcal{S}^{K_v}_{b,t}]\, C^{+}_{b,t,j}}{\sum_{\ell=1}^{N_b} \mathbb{I}[\ell \in \mathcal{S}^{K_v}_{b,t}]\, C^{+}_{b,t,\ell} + \epsilon}. \tag{11}$$

The top-K_v operator selects the target support, while the linear normalization preserves the relative magnitude of positive credit within that support. We stop gradients through $Q_{b,t}$ during attention alignment. We further define the visual credit mass of token $y_{b,t}$ as:

$$m_{b,t} = \sum_{j=1}^{N_b} |C_{b,t,j}|\,, \tag{12}$$

which measures the overall sensitivity of the generated token to the visual input.

4.3 SELECTIVE VISUAL SUPERVISION OBJECTIVE

Let $P_{b,t}$ denote the final-layer visual-only attention distribution for predicted token $y_{b,t}$, averaged over all attention heads. To guide the policy toward relevant visual cues, we regularize $P_{b,t}$ toward the credit-derived target $Q_{b,t}$ via KL divergence (Kullback & Leibler, 1951), $D_{\mathrm{KL}}(Q_{b,t} \,\|\, P_{b,t})$. While computing token-level visual credits involves an auxiliary backward pass per supervised token, applying this regularization uniformly across the entire response incurs redundant computation on non-visual tokens without yielding proportional performance gains.

Crucially, as Figure 2 illustrates, visual sensitivity concentrates heavily on spatial decision tokens rather than descriptive context. Guided by this insight, **VISTA** bypasses uniform sequence intervention and restricts visual supervision strictly to the decisive coordinate-token span, denoted by $\mathcal{T}_b$. For generations lacking valid coordinate spans, $\mathcal{T}_b$ naturally defaults to the final K_t output tokens. The visual regularization loss over these target tokens is formulated as:

$$\mathcal{L}_{\mathrm{VISTA}} = \frac{\sum_{b=1}^{B} \sum_{t \in \mathcal{T}_b} m_{b,t}\, D_{\mathrm{KL}}\left(Q_{b,t} \,\|\, P_{b,t}\right)}{\sum_{b=1}^{B} \sum_{t \in \mathcal{T}_b} m_{b,t} + \epsilon}, \tag{13}$$

where B is the batch size, $\epsilon > 0$ prevents division by zero, and the visual credit mass $m_{b,t}$ scales each token contribution so that visually sensitive tokens provide stronger guidance. Finally, **VISTA** jointly optimizes the primary policy objective alongside this fine-grained visual regularization:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{GRPO}} + \lambda_{\mathrm{VISTA}} \mathcal{L}_{\mathrm{VISTA}}, \tag{14}$$

where λ_{VISTA} balances the two terms. By concentrating on decisive tokens, **VISTA** delivers targeted visual guidance while preserving training efficiency.

Overall, **VISTA** converts coarse response-level feedback into token-specific visual supervision via credit estimation and attention regularization, deriving targets from policy objectives without extra spatial annotations. Crucially, it integrates into multimodal RL pipelines to regularize coordinate-token attention alongside policy optimization, grounding updates in localized visual evidence.

5 EXPERIMENTS

We evaluate **VISTA** across 13 multimodal reasoning and visual grounding benchmarks, comparing representative open-source MLLMs against frontier proprietary and task-specific baselines.

5.1 Experiment Setup

Models and Evaluation Datasets. We train and evaluate our models using two representative open-source MLLM families, Qwen2.5-VL (Bai et al., 2025) and InternVL2.5 (Chen et al., 2024b). We evaluate multimodal reasoning on MathVista (Lu et al., 2024), MathVision (Wang et al., 2024a), We-Math (Qiao et al., 2025), MMMU-Pro (Yue et al., 2025), MMStar (Chen et al., 2024a), and OCRBench (Liu et al., 2024). Multimodal grounding is evaluated on RefCOCO and RefCOCO+ (Yu et al., 2016), RefCOCOg (Mao et al., 2016), LISA-Grounding (LISA-G) (Lai et al., 2024), Flickr30k Entities (Plummer et al., 2015), RefCLEF (Kazemzadeh et al., 2014), and RIO (Qu et al., 2023).

Baselines. For multimodal reasoning, we compare with GPT-5 (Singh et al., 2025), GPT-4o (OpenAI, 2024), Gemini 2.5 Pro (Comanici et al., 2025), InternVL2.5 (Chen et al., 2024b), InternVL2.5-MPO (Wang et al., 2024b), Qwen2.5-VL (Bai et al., 2025), ThinkLite-VL (Wang et al., 2025c), NoisyRollout (Liu et al., 2025b), R1-OneVision (Yang et al., 2025), Vision-R1 (Huang et al., 2026), X-Reasoner (Liu et al., 2025a), MM-Eureka (Meng et al., 2025), R1-ShareVL (Yao et al., 2025), and VL-Rethinker (Wang et al., 2025a). For multimodal grounding, we compare with Qwen2.5-VL (Bai et al., 2025), InternVL2.5 (Chen et al., 2024b), InternVL2.5-MPO (Wang et al., 2024b), Groma (Ma et al., 2024), GroundingGPT (Li et al., 2024), LLaVA-Grounding (Zhang et al., 2024), VisionReasoner (Liu et al., 2026), Dr. Seg (Sun et al., 2026a), RegionReasoner (Sun et al., 2026b), Vision-R1 (Zhan et al., 2026), and VLM-R1 (Shen et al., 2025) (baseline details in Appendix B.6).

Training Details and Hyperparameters. For visual grounding, we train on merged RefCOCO/+/g splits (Yu et al., 2016; Mao et al., 2016) via VLM-R1 (Shen et al., 2025); for multimodal reasoning, we train on ViRL39K using VL-Rethinker (Wang et al., 2025a). Across both settings, we sample 8 rollouts per prompt with $\lambda_{\text{VISTA}} = 0.05$. In grounding tasks, guided by the localized sensitivity in Figure 2, **VISTA** targets the predicted coordinate-token span; for reasoning and understanding tasks without explicit bounding spans, it regularizes the final K valid output tokens; both share the identical core formulation. In all cases, **VISTA** acts strictly as an auxiliary objective without altering the base RL pipeline. Training runs on a cluster with 8× NVIDIA H100 GPUs, selecting checkpoints via training reward dynamics (see Appendix B.2 and Appendix C).

Evaluation Metrics. For multimodal grounding, we report $\text{Acc@IoU} > 0.5$ (%), where a prediction is considered correct if its maximum IoU with any valid ground-truth box is strictly greater than 0.5. For multimodal reasoning, we report Pass@1 (%) under our unified answer-verification pipeline. For more information about our evaluation, please refer to Appendix B.7.

5.2 Main Results

Grounding Performance. As shown in Table 1, **VISTA** consistently outperforms the VLM-R1 baseline under $\text{Acc@IoU} > 0.5$, lifting the macro-average from 68.5% to 70.3% at 3B and from 69.8% to 72.1% at 7B. These gains span REC, segmentation, and OOD grounding tasks. Improvements are particularly notable on LISA-G (+4.3 at 3B; +4.8 at 7B), RefCOCO+ (+1.0 at 3B; +3.1 at 7B), and Flickr (+2.5 at 3B; +2.6 at 7B). Notably, our 7B model leads across every benchmark against all 7B competitors, matching the performance of substantially larger 38B to 78B MLLMs.

Reasoning Performance. To evaluate generalizability, we apply **VISTA** to multimodal reasoning (Table 2). Since multi-step reasoning diffuses visual reliance before consolidating it near the answer, we selectively supervise the last-K tokens to target this decision stage without perturbing earlier steps. Under unified Pass@1, **VISTA** improves the VL-Rethinker baseline from 52.3% to 54.7%, with notable gains in visual math (+6.8 on WeMath, +2.9 on MathVision) and consistent gains elsewhere, validating its broader utility.

Generalization to InternVL. Table 3 evaluates transferability to InternVL2.5-4B beyond the primary Qwen2.5-VL series. **VISTA** consistently improves upon VLM-R1 across all seven benchmarks, lifting the macro-average by 1.3 points to 70.2%, with notable gains on LISA-G (+3.3) and RIO (+2.1). These results verify that our regularization objective transfers robustly across model families.

5.3 Design Verification

Visual Credit and Token Selection Ablation. Figure 4 evaluates both our credit formulation (left) and output-token selection strategies (right) on Qwen2.5-VL-7B-Instruct. On credit design,

Table 1: Main results on multimodal grounding benchmarks. Metrics report $\mathrm{Acc@IoU} > 0.5$ (%). REC, Seg., and OOD denote referring expression comprehension, segmentation-aware, and out-of-domain grounding. Avg. is the unweighted mean across all seven benchmarks. Best results among 7B models are in **bold**.

Method	Scale	REC			Seg.	OOD Grounding			Avg.
		RefCOCO	RefCOCO+	RefCOCOg	LISA-G	Flickr	RefCLEF	RIO	
General open-source MLLMs									
Qwen2.5-VL	7B	84.0	81.8	81.4	64.4	35.6	70.7	50.1	66.9
InternVL2.5	8B	88.4	84.0	84.5	59.6	36.5	63.9	46.4	66.2
InternVL2.5-MPO	8B	89.0	84.9	85.5	59.6	36.3	64.1	47.3	66.7
Qwen2.5-VL	32B	86.9	81.0	84.5	68.4	35.6	58.0	54.5	67.0
InternVL2.5	38B	91.8	87.8	89.4	67.5	37.6	73.3	56.8	72.0
Qwen2.5-VL	72B	91.4	84.4	87.4	78.3	36.7	73.8	55.8	72.5
InternVL2.5	78B	93.4	89.9	91.8	72.6	35.8	70.4	55.3	72.7
Grounding-tuned open MLLMs									
LLaVA-Grounding	–	33.0	27.2	36.1	23.8	21.1	7.9	26.9	25.1
GroundingGPT	–	64.9	60.4	55.5	13.2	8.7	44.2	12.8	37.1
Groma	7B	89.0	83.2	84.7	52.1	34.9	56.9	36.2	62.4
VisionReasoner	7B	86.3	80.9	80.2	67.5	35.6	69.8	50.8	67.3
Dr. Seg	7B	74.5	70.5	73.3	62.0	33.4	65.0	47.9	60.9
RegionReasoner	7B	86.7	81.3	81.6	66.6	28.5	71.0	49.1	66.4
Vision-R1	7B	87.5	83.8	84.1	70.5	33.8	67.8	52.5	68.6
VLM-R1	3B	88.0	83.5	84.3	62.9	36.3	73.1	51.3	68.5
	7B	88.2	83.3	84.7	67.8	37.9	74.2	52.7	69.8
VISTA	3B	88.5	84.5	84.5	67.2	38.8	75.0	53.5	70.3
(Ours)	7B	**89.3**	**86.4**	**84.9**	**72.6**	**40.5**	**75.9**	**54.9**	**72.1**

Table 2: Main results on multimodal understanding and reasoning benchmarks. Metrics report Pass@1 (%) under our unified evaluation pipeline, with Avg. representing the unweighted mean across all six benchmarks. The best results among 7B reasoning models are in **bold**.

Model	Size	Visual Math			General Reasoning		Visual Text	Avg.
		MathVista	MathVision	WeMath	MMMU-Pro	MMStar	OCRBench	
Closed-source frontier MLLMs								
GPT-5	–	81.9	72.0	71.1	78.4	75.7	80.7	76.7
GPT-4o	–	63.8	30.4	68.8	51.9	64.7	80.6	60.0
Gemini 2.5-Pro	–	80.9	69.1	78.0	68.0	73.6	86.2	76.0
Open-source general MLLMs								
InternVL2.5	8B	51.0	13.7	27.8	23.4	52.4	76.2	40.8
InternVL2.5-MPO	8B	68.3	22.5	53.2	30.2	57.2	79.7	51.9
Qwen2.5-VL	7B	66.0	23.2	45.9	28.9	58.2	53.8	46.0
InternVL2.5	38B	68.0	25.7	45.1	40.6	63.3	84.2	54.5
Qwen2.5-VL	32B	75.5	34.5	59.2	49.7	66.4	54.3	56.6
InternVL2.5	78B	70.8	24.8	48.8	50.1	66.2	85.7	57.7
Qwen2.5-VL	72B	81.5	43.3	75.5	69.0	82.0	89.3	73.4
Reasoning-tuned open MLLMs								
ThinkLite-VL	7B	70.8	23.5	50.5	23.0	49.3	55.3	45.4
NoisyRollout	7B	72.0	25.7	53.8	33.8	64.0	54.8	50.7
R1-Onevision	7B	65.0	20.9	43.9	31.2	55.1	50.2	44.4
Vision-R1	7B	66.8	15.5	61.7	34.8	61.8	32.5	45.5
MM-Eureka	7B	71.5	28.1	47.6	27.4	50.2	51.7	46.1
X-Reasoner	7B	72.0	25.7	42.7	33.8	64.0	54.8	48.8
R1-ShareVL	7B	72.3	27.8	55.2	42.5	65.3	51.7	52.5
VL-Rethinker	7B	71.3	25.4	55.3	40.8	66.2	54.5	52.3
VISTA (Ours)	**7B**	**73.0**	**28.3**	**62.1**	**42.1**	**67.3**	**55.3**	**54.7**

removing advantage weighting (Reward-Only) drops the average score to 80.8%, while disrupting spatial structure (Permuted) or token granularity (Response-Level) further degrades accuracy to 79.7% and 80.1%, with a uniform target trailing at 78.9%. Regarding token selection along the sequence, supervising the first-K (80.8%), last-K (81.7%), or uniformly spaced tokens (79.5%) underperforms dense full-token regularization (83.2%). Crucially, our targeted supervision on decision tokens achieves the highest score of 83.3% without requiring exhaustive sequence-wide intervention. Corroborating the localized sensitivity observed in Figure 2, these results confirm that

Table 3: Comparison with 4B-scale baselines on multimodal grounding benchmarks.

Method	REC			Seg.	OOD Grounding			Avg.
	RefCOCO	RefCOCO+	RefCOCOg	LISA-G	Flickr	RefCLEF	RIO	
InternVL2.5-4B	89.6	82.5	86.1	50.6	34.5	64.8	44.9	64.7
InternVL2.5-4B-MPO	90.2	83.9	88.2	51.8	35.6	61.9	46.4	65.4
VLM-R1	90.6	83.7	88.4	68.1	36.3	67.5	47.8	68.9
VISTA (Ours)	**91.0**	**84.3**	**89.0**	**71.4**	**36.7**	**69.2**	**49.9**	**70.2**

selectively regularizing decisive tokens avoids imposing unwarranted constraints on context tokens that minimally rely on visual input, achieving an effective and targeted supervision scheme.

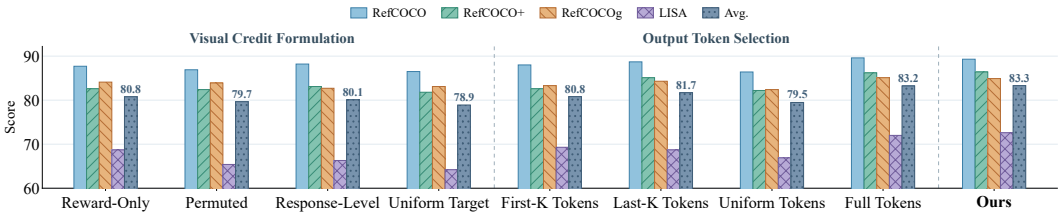


Figure 4: **Ablation on visual credit and token selection.** Comparisons of credit formulations (left) and token subsets (right) on grounding benchmarks. Selective supervision at decision tokens reflects localized visual sensitivity, surpasses position heuristics, and matches full alignment.

Target Attention Boosts Grounding. Figure 5 shows a pronounced positive correlation between target region attention and grounding performance. As attention concentrates on the ground truth region, grounding IoU rises sharply and remains high. Quantitatively, a Pearson correlation coefficient of $r = 0.3994$ demonstrates a positive linear dependency, suggesting that reinforcing token-level visual focus serves as an inductive bias for improving spatial localization during policy optimization.

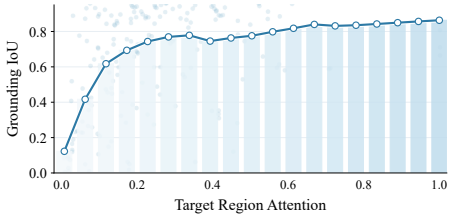


Figure 5: **Focused target visual attention boosts grounding.** Concentrating attention on target regions consistently yields higher grounding IoU.

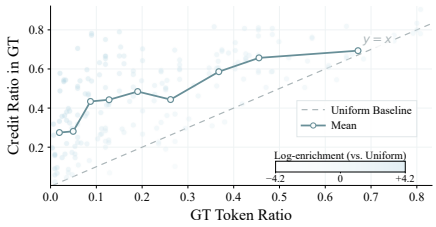
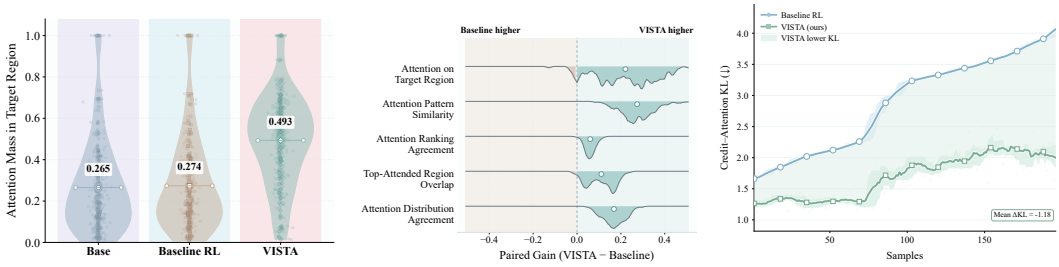


Figure 6: **Credit enrichment on GT tokens.** GT credit fraction consistently exceeds the uniform baseline ($y = x$), demonstrating targeted credit allocation.

5.4 Mechanism Analysis

Credit Enrichment on Ground-Truth Tokens. Figure 6 validates our credit formulation. Uninformative credits would trace the uniform $y = x$ baseline, whereas empirical binned means consistently surpass it across nearly all GT ratios. This enrichment peaks when GT tokens are sparse, supporting that supervision concentrates on visually critical tokens rather than token frequency. As the GT ratio increases, the margin naturally narrows. This suggests that our credit design provides an effective selective signal for visual regularization.

Visual Regularization Analysis. Figure 7 shows how **VISTA** reshapes visual attention allocation. Notably, attention mass dedicated to target regions nearly doubles, rising from 0.274 under baseline RL to 0.493 (Figure 7a), producing widespread positive paired gains across target attention metrics (Figure 7b). Crucially, **VISTA** systematically reduces the KL divergence between credit and attention (Figure 7c), confirming that learned attention profiles align well with target visual credit. Together,



(a) Attention on target regions. (b) Paired visual alignment gains. (c) Credit and attention divergence.

Figure 7: **Mechanistic analysis of VISTA. VISTA** increases attention to target regions, improves visual alignment across metrics, and minimizes credit-attention distribution divergence.

these findings demonstrate that **VISTA** effectively encourages localized visual focus consistent with credit during policy optimization.

Layer-wise Analysis. In figure 8, the target attention gain is negligible in early layers but increases sharply in deeper layers, with several heads exhibiting consistently strong gains. This pattern is consistent with **VISTA** primarily reshaping high-level visual routing near the output stage.

Global Attention Regularization. Figure 9 visualizes spatial attention centered on ground-truth targets. Compared to baseline RL, **VISTA** concentrates visual focus tightly within target regions while dampening peripheral background noise. The difference map confirms this redistribution: positive attention shifts cluster near the target, with widespread reductions across distant regions. This demonstrates that **VISTA** systematically improves target-selective visual allocation across samples, rather than relying on sparse outliers.

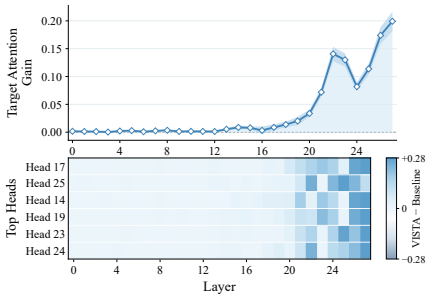


Figure 8: **Layer-wise target attention gain.** Average head gains (top) and top-head gains across layers (bottom).

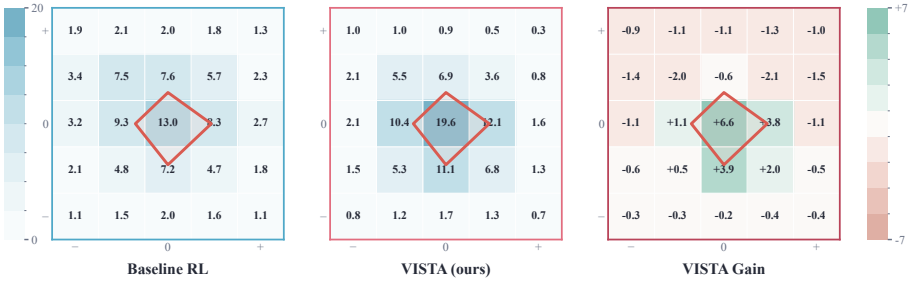


Figure 9: **Global attention alignment maps.** Centered on GT targets and averaged across samples, attention maps show that **VISTA** redirects focus from the background onto the target.

6 CONCLUSION

In this work, we address key limitations in multimodal RL: maximizing outcome rewards fails to guarantee reliance on correct visual evidence, especially as visual dependence varies across tokens. We present **VISTA**, which complements response-level optimization with token-specific visual supervision to steer policy updates toward target visual regions. Across 13 grounding and multimodal reasoning benchmarks, **VISTA** consistently outperforms strong RL baselines and transfers across model scales and task domains. Mechanistic analyses confirm these gains are accompanied by more concentrated target attention, reduced peripheral distraction, and tighter alignment between visual cues and decisions. Coupling outcome rewards with fine-grained visual guidance paves the way for developing more visually faithful, accountable multimodal reasoning models.

AI USE STATEMENT

During the preparation of this manuscript, large language models (LLMs) were employed to polish draft sections for improved readability and phrasing, while all technical arguments, findings, and claims remain entirely the authors' own. Additionally, LLMs were consulted during the early stages solely as an exploratory retrieval aid to help locate relevant research in the field; the paper's core ideas were conceived by the authors after independently reading and synthesizing this literature. All works cited in the Related Work section were subsequently retrieved, manually verified, and curated by the authors without relying on model-generated citations. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We have taken comprehensive steps to facilitate the reproducibility of our training pipelines. The core methodology and overall training workflow are described in Section 5.1 and Appendix B.2, with implementation details, core hyperparameter configurations, and experimental environments elaborated in Appendix B and Appendix C. Furthermore, comprehensive specifications regarding benchmark datasets, evaluation protocols, prompt templates, and scoring metrics are detailed in Appendix B and Appendix E, while supplementary ablation studies are reported in Appendix D. Finally, we provide clean, anonymized source code packaged as a zip archive in the **Supplementary Material**, complete with execution scripts and a comprehensive README to guide environment configuration and pipeline execution.

REFERENCES

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the 42nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29915–29926, 2025.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 27056–27087, 2024a.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 24185–24198, 2024c.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Lingxiao Du, Fanqing Meng, Zongkai Liu, Zhixiang Zhou, Ping Luo, Qiaosheng Zhang, and Wenqi Shao. Mm-prm: Enhancing multimodal mathematical reasoning with scalable step-level supervision. *arXiv preprint arXiv:2505.13427*, 2025.

Haowen Gao, zhenyu zhang, Liang Pang, Fangda Guo, Hongjian Dou, Guannan Lv, ShaoGuo Liu, Tingting Gao, Huawei Shen, and Xueqi Cheng. DIVA-GRPO: Enhancing multimodal reasoning through difficulty-adaptive variant advantage. In *Proceedings of the 14th International Conference on Learning Representations (ICLR)*, 2026.

Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. In *Proceedings of the 14th International Conference on Learning Representations (ICLR)*, 2026.

Sarthak Jain and Byron C Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, volume 2025, pp. 87676–87703, 2025a.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the 42nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9339–9350, 2025b.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 787–798, 2014.

Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the 41st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9579–9589, 2024.

Zhaowei Li, Qi Xu, Dong Zhang, Hang Song, Yiqing Cai, Qi Qi, Ran Zhou, Junting Pan, Zefeng Li, Vu Tu, et al. GroundingGPT: Language enhanced multi-modal grounding model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 6657–6678, 2024.

Qianchu Liu, Sheng Zhang, Guanghui Qin, Timothy Ossowski, Yu Gu, Ying Jin, Sid Kiblawi, Sam Preston, Mu Wei, Paul Vozila, Tristan Naumann, and Hoifung Poon. X-reasoner: Towards generalizable reasoning across modalities and domains. *arXiv preprint arXiv:2505.03981*, 2025a.

Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. Noisyrollout: Reinforcing visual reasoning with data augmentation. In *Proceedings of the 39th Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 2923–2957, 2025b.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024.

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 *Proceedings of the 14th International Conference on Learning Representations (ICLR)*, volume 2026, pp. 94069–94086, 2026.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, volume 2024, pp. 23439–23554, 2024.

Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. Groma: Localized visual tokenization for grounding multimodal large language models. In *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, pp. 417–435, 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 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11–20, 2016.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

Yingqian Min, Kun Zhou, Yifan Li, Yuhuan Wu, Han Peng, Yifan Du, Wayne Xin Zhao, Min Yang, and Ji-Rong Wen. Improving vision-language models with perception-centric process reward models. In *Proceedings of the 43rd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33099–33109, 2026.

OpenAI. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the 15th IEEE International Conference on Computer Vision (ICCV)*, pp. 2641–2649, 2015.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma GongQue, Shanglin Lei, YiFan Zhang, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Xiao Zong, Yida Xu, Peiqing Yang, Zhimin Bao, Muxi Diao, Chen Li, and Honggang Zhang. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 20023–20070, 2025.

Mengxue Qu, Yu Wu, Wu Liu, Xiaodan Liang, Jingkuan Song, Yao Zhao, and Yunchao Wei. Rio: A benchmark for reasoning intention-oriented objects in open environments. In *Proceedings of the 37th Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 43041–43056, 2023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

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.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, et al. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*, 2025.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Haoxiang Sun, Tao Wang, Chenwei Tang, Li Yuan, and Jiancheng Lv. Dr. seg: Revisiting grpo training for visual large language models through perception-oriented design. In *Proceedings of the 43rd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24320–24329, 2026a.

Wenfang Sun, Hao Chen, Yingjun Du, Yefeng Zheng, and Cees GM Snoek. Regionreasoner: Region-grounded multi-round visual reasoning. *arXiv preprint arXiv:2602.03733*, 2026b.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL): Findings*, pp. 13088–13110, 2024.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. In *Proceedings of the 39th Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 38, 2025a.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 95095–95169, 2024a.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, et al. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*, 2024b.

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, et al. Visualprm: An effective process reward model for multimodal reasoning. *arXiv preprint arXiv:2503.10291*, 2025b.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. In *Proceedings of the 39th Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 38, 2025c.

Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. Grounded chain-of-thought for multimodal large language models. In *Proceedings of the 43rd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 33577–33587, 2026.

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Perception-r1: Advancing multimodal reasoning capabilities of MLLMs via visual perception reward. In *Proceedings of the 14th International Conference on Learning Representations (ICLR)*, 2026.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In *Proceedings of the 20th IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2376–2385, 2025.

Huanjin Yao, Qixiang Yin, Jingyi Zhang, Min Yang, Yibo Wang, Wenhao Wu, Fei Su, Li Shen, Minghui Qiu, Dacheng Tao, et al. R1-sharevl: Incentivizing reasoning capabilities of multimodal large language models via share-grpo. In *Proceedings of the 39th Annual Conference on Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Hao Yin, Guangzong Si, and Zilei Wang. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the 42nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14625–14634, 2025.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Proceedings of the 14th European Conference on Computer Vision (ECCV)*, pp. 69–85, 2016.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the 41st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13807–13816, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 15134–15186, 2025.

Yufei Zhan, Yousong Zhu, Shurong Zheng, Hongyin Zhao, Fan Yang, Ming Tang, and Jinqiao Wang. Vision-r1: Evolving human-free alignment in large vision-language models via vision-guided reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5807–5817, 2026.

Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Lei Zhang, Chunyuan Li, et al. Llava-grounding: Grounded visual chat with large multimodal models. In *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, pp. 19–35, 2024.

Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. In *Proceedings of the 20th IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1859–1869, 2025.

Jianfei Zhao, Feng Zhang, Xin Sun, Chong Feng, and Zhixing Tan. Tell model where to look: Mitigating hallucinations in mllms by vision-guided attention. In *Proceedings of the 43rd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 32582–32591, 2026.

APPENDIX

A Algorithm of VISTA **16**

A.1 Group Relative Policy Optimization (GRPO) 16

A.2 Training Algorithm Overview 16

B Experiment Details **16**

B.1 Models and Environments 16

B.2 Training Details 18

B.3 Training Datasets 18

B.4 Evaluation Datasets 19

B.5 Computing Resources 20

B.6 Baselines 20

B.7 Evaluation Metrics 22

C Hyperparameters **23**

C.1 Grounding 23

C.2 Multimodal Reasoning 23

D Additional Ablation Studies **23**

E Prompts **25**

F Case Study **26**

A ALGORITHM OF VISTA

A.1 GROUP RELATIVE POLICY OPTIMIZATION (GRPO)

To optimize π_θ efficiently, we build upon GRPO (Shao et al., 2024). For each query x, the reference policy $\pi_{\theta_{old}}$ samples a group of G trajectories $\{\tau_i\}_{i=1}^G$ and receives corresponding rewards $\{r_i\}_{i=1}^G$. The advantage $\hat{A}_i$ for the i-th trajectory is computed via group-level normalization:

$$\hat{A}_i = \frac{r_i - \mu}{\sigma} \tag{15}$$

where μ and σ denote the empirical mean and standard deviation of the group's rewards. The policy parameters θ are updated by maximizing the clipped surrogate objective at the token level:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\sum_{j}\min\left(\gamma_{i,j}(\theta)\hat{A}_i, \text{clip}(\gamma_{i,j}(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_i\right) - \beta\mathbb{D}_{KL}\right] \tag{16}$$

where $\gamma_{i,j}(\theta) = \frac{\pi_\theta(y_{i,j}|\cdot)}{\pi_{\theta_{old}}(y_{i,j}|\cdot)}$ represents the token-level likelihood ratio, ϵ is the clipping hyperparameter, and $\mathbb{D}_{KL}$ is the Kullback-Leibler divergence (Kullback & Leibler, 1951) penalty applied to anchor the optimization to the base model.

A.2 TRAINING ALGORITHM OVERVIEW

Algorithm 1 outlines the overall training procedure of **VISTA**. For each sampled rollout, we begin by computing the standard response-level advantage using the base RL algorithm. Rather than supervising all output positions uniformly, we isolate a sparse subset of decision-critical tokens for visual regularization; for grounding tasks, these correspond to the numeric coordinates specifying the bounding box. For each identified token, **VISTA** estimates its signed visual credit across all valid image patches, distills a sparse target distribution over the top-K_v positive credit values, and guides the final-layer cross-modal attention toward this target. This auxiliary regularization loss integrates seamlessly into the original RL objective, enhancing grounding and reasoning precision without disrupting the underlying policy optimization pipeline.

The temporary gates z_b serve only as auxiliary differentiable variables for visual-credit estimation and are excluded from the optimizer state. During credit extraction, we differentiate the advantage-weighted token log-probability with respect to the gates applied to the projected visual embeddings, while stopping gradients through the response-level advantage A_b. The resulting credit-derived target $Q_{b,t}$ is also detached during attention alignment, such that the auxiliary objective updates the policy through the attention distribution $P_{b,t}$. This visual-only attention is extracted from the final transformer layer, averaged across attention heads, and normalized over valid visual-token positions. For responses containing a valid coordinate span, **VISTA** applies supervision to the selected coordinate tokens; otherwise, it falls back to the final K_t valid output tokens. A selected token is skipped when no positive visual credit is available to construct a valid target distribution. Consequently, $\mathcal{L}_{\text{VISTA}}$ contributes only when a valid credit-derived visual target can be established, and the optimization reduces to the original RL objective when no selected token provides such supervision.

B EXPERIMENT DETAILS

B.1 MODELS AND ENVIRONMENTS

Backbone Models. We instantiate **VISTA** across two representative open-source MLLM families, spanning model scales from 3B to 7B. Specifically, we evaluate Qwen2.5-VL-3B, Qwen2.5-VL-7B (Bai et al., 2025), and InternVL2.5-4B (Chen et al., 2024b;c). This selection enables us to examine whether our visual alignment mechanism generalizes effectively across distinct model architectures and parameter scales.

Qwen2.5-VL incorporates a native dynamic-resolution vision encoder that preserves the spatial layout of visual inputs, employing window attention to lower visual encoding costs. The extracted visual features are compressed and mapped into the language embedding space via an MLP-based

Algorithm 1 VISTA Training for Multimodal Reinforcement Learning

Require: Policy π_θ; batch $\{x_b = (I_b, q_b)\}_{b=1}^{B}$; output-token budget K_t; visual-token budget K_v; λ_{VISTA}; ϵ
Ensure: Updated policy parameters θ

Sample responses $y_b = (y_{b,1}, \ldots, y_{b,T_b}) \sim \pi_\theta(\cdot \mid I_b, q_b)$
Compute response-level advantages A_b and original RL loss $\mathcal{L}_{\text{GRPO}}$
Initialize $\mathcal{N} \leftarrow 0,\ \mathcal{D} \leftarrow 0$
for $b = 1, \ldots, B$ **do**
  Encode I_b into projected visual tokens $V_b = (v_{b,1}, \ldots, v_{b,N_b})$
  Parse y_b for valid coordinate-token indices $\mathcal{B}_b$
  if $\mathcal{B}_b \neq \varnothing$ **then**
    $\mathcal{T}_b \leftarrow \text{SpreadSample}(\mathcal{B}_b, K_t)$ ▷ coordinate tokens
  else
    $\mathcal{T}_b \leftarrow \text{LastK}(y_b, K_t)$ ▷ terminal output tokens
  end if
  Set temporary gates $\widetilde{v}_{b,j} = z_{b,j} v_{b,j}$ with $z_{b,j} = 1$
  Run π_θ with $\widetilde{V}_b$ to obtain token log-probabilities $\ell_{b,t}$ and visual-only attention $P_{b,t}$
  for all $t \in \mathcal{T}_b$ **do**
    Compute signed visual credit

$$C_{b,t} \leftarrow \nabla_{z_b}[\text{sg}(A_b)\ell_{b,t}]|_{z_b=\mathbf{1}}$$

    Retain positive credit $C^{+}_{b,t,j} \leftarrow \max(C_{b,t,j}, 0)$
    Select support $\mathcal{S}^{K_v}_{b,t} \leftarrow \text{TopK}(C^{+}_{b,t}, K_v)$
    if $\sum_{j \in \mathcal{S}^{K_v}_{b,t}} C^{+}_{b,t,j} \leq 0$ **then**
      continue ▷ no positive visual credit
    end if
    Construct the stop-gradient visual target

$$Q_{b,t,j} \leftarrow \text{sg}\left(\frac{\mathbb{I}[j \in \mathcal{S}^{K_v}_{b,t}]\, C^{+}_{b,t,j}}{\sum_{\ell \in \mathcal{S}^{K_v}_{b,t}} C^{+}_{b,t,\ell} + \epsilon} \right)$$

    Compute visual credit mass $m_{b,t} \leftarrow \sum_{j=1}^{N_b} |C_{b,t,j}|$
    Compute credit–attention divergence $D_{b,t} \leftarrow D_{\text{KL}}(Q_{b,t} \| P_{b,t})$
    Accumulate $\mathcal{N} \leftarrow \mathcal{N} + m_{b,t} D_{b,t}$, $\mathcal{D} \leftarrow \mathcal{D} + m_{b,t}$
  end for
end for
Compute selective visual regularization $\mathcal{L}_{\text{VISTA}} \leftarrow \dfrac{\mathcal{N}}{\mathcal{D} + \epsilon}$
Form joint objective $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{GRPO}} + \lambda_{\text{VISTA}} \mathcal{L}_{\text{VISTA}}$
Update $\theta \leftarrow \text{OPTIMIZERSTEP}(\theta, \nabla_\theta \mathcal{L}_{\text{total}})$
return θ

vision-language projector. We evaluate its 3B and 7B variants, which share identical visual frontends but differ in language-backbone capacities (Bai et al., 2025).

InternVL2.5 adopts an alternative modular architecture composed of an InternViT vision encoder, a lightweight MLP projector, and a pretrained language model. We employ InternVL2.5-4B, which couples InternViT-300M-448px-V2.5 with Qwen2.5-3B-Instruct. InternVL2.5 handles high-resolution inputs via dynamic 448×448 patch tiling and compresses visual tokens prior to language ingestion (Chen et al., 2024b). These structural contrasts between Qwen2.5-VL and InternVL2.5 offer a rigorous testbed for assessing the architecture-agnostic efficacy of **VISTA**.

Grounding Training Environment. For multimodal grounding benchmarks, we build on the VLM-R1 framework (Shen et al., 2025). VLM-R1 adapts R1-style reinforcement learning to vision-language domains via rule-based rewards, establishing a dedicated GRPO training pipeline for Referring Expression Comprehension (REC). Its implementation natively accommodates both QwenVL and InternVL backbones, allowing us to maintain a uniform training environment across both model families. We preserve the original rollout generation, reward calculation, group-relative advantage estimation, and policy optimization routines intact, seamlessly incorporating **VISTA** as an auxiliary

visual alignment regularizer. Consequently, the core RL optimization remains strictly controlled, ensuring a faithful comparison against standard outcome-RL baselines.

Reasoning Training Environment. For multimodal reasoning tasks, we build upon the training framework introduced by VL-Rethinker (Wang et al., 2025a), which is developed within the OpenRLHF ecosystem. VL-Rethinker enhances GRPO-based multimodal RL with techniques such as Selective Sample Replay to prevent vanishing advantages and Forced Rethinking to foster self-reflective reasoning behaviors. We adopt its complete OpenRLHF infrastructure for rollout generation, reward and advantage estimation, and policy updates, integrating **VISTA** strictly as an auxiliary attention alignment objective during policy optimization. This modular design allows **VISTA** to augment existing reasoning RL pipelines without disrupting their primary response-level optimization dynamics.

B.2 Training Details

Training Frameworks. We use two task-specific open-source reinforcement learning frameworks. For grounding, we build on VLM-R1 (Shen et al., 2025), a GRPO-based pipeline for referring-expression comprehension. We retain its original rollout, reward computation, advantage estimation, and optimization procedures, applying **VISTA** as an auxiliary regularization objective at every actor update. For multimodal reasoning, we adopt the OpenRLHF-based VL-Rethinker framework (Wang et al., 2025a). To balance supervisory efficacy with training efficiency, we introduce an intermittent alignment interval of 4: **VISTA** is activated every four actor mini-batch updates (i.e., updates $1, 5, 9, \ldots$), while the remaining updates optimize the standard RL objective without computing visual credits or attention regularization loss.

Checkpoint Saving and Selection. We periodically save intermediate checkpoints throughout training to track optimization dynamics and enable model selection. For grounding experiments, checkpoints are saved every 100 steps. For reasoning experiments, checkpoints are saved every 5 steps, allowing finer-grained tracking of the rapidly changing reward dynamics during reinforcement learning. We select checkpoints according to the reward dynamics observed during training. Specifically, we monitor the training reward trajectory and choose checkpoints from stages where the policy reaches strong and stable reward performance. This selection is performed solely using training-time signals and does not use results from the reported evaluation benchmarks. The same selection principle is applied consistently across different model scales and experimental settings.

B.3 Training Datasets

We use task-specific training data for multimodal reasoning and grounding. For reasoning, we follow the data setup of VL-Rethinker and train on ViRL39K (Wang et al., 2025a). For grounding, we follow VLM-R1 (Shen et al., 2025) using a unified training set by combining the training splits of RefCOCO, RefCOCO+, and RefCOCOg (Yu et al., 2016; Mao et al., 2016).

Reasoning Training Data. For multimodal reasoning, we use ViRL39K (Wang et al., 2025a), a curated dataset designed for vision-language reinforcement learning. ViRL39K contains 38,870 verifiable visual question–answer pairs spanning eight broad categories, including mathematical, scientific, diagrammatic, tabular, document, and spatial reasoning problems. The dataset is constructed from both newly collected examples and existing multimodal reasoning resources through cleaning, reformatting, rephrasing, and answer verification. Importantly, the answers are designed to support rule-based verification, making the dataset particularly suitable for outcome-based RL. ViRL39K also provides model-capability annotations based on empirical pass rates, which characterize the difficulty of individual examples for models at different scales. We use ViRL39K as the reasoning training corpus within the VL-Rethinker training environment.

Grounding Training Data. For referring-expression grounding, we use the training splits of RefCOCO (Yu et al., 2016), RefCOCO+ (Yu et al., 2016), and RefCOCOg (Mao et al., 2016). All three datasets are built on MS-COCO images and associate natural-language referring expressions with localized objects. RefCOCO and RefCOCO+ were collected through an interactive referring game; RefCOCO+ further prohibits absolute spatial descriptions, placing greater emphasis on

Table 4: Composition of the grounding training data. The three training splits are directly concatenated for reinforcement learning.

Dataset	# Training Instances
RefCOCO (Yu et al., 2016)	120,624
RefCOCO+ (Yu et al., 2016)	120,191
RefCOCOg (Mao et al., 2016)	80,512
Total	**321,327**

appearance-based disambiguation (Yu et al., 2016). RefCOCOg is collected in a non-interactive setting and generally contains longer and more descriptive referring expressions (Mao et al., 2016).

The processed training files contain 120,624 examples from RefCOCO, 120,191 from RefCOCO+, and 80,512 from RefCOCOg. We directly concatenate the three training sets without dataset-specific reweighting, resulting in 321,327 grounding training instances in total. Each instance consists of an image, a referring expression, and its corresponding target bounding box, which enables rule-based reward computation during grounding RL.

B.4 Evaluation Datasets

We evaluate **VISTA** across a comprehensive suite of multimodal reasoning and visual grounding benchmarks. To maintain evaluation efficiency, we construct standardized evaluation subsets via uniform random sampling (fixed seed 42) and evaluate all compared models on these exact subsets. Before sampling, we strictly decontaminate the candidate evaluation pools by filtering out any instances that overlap with our training data, ensuring zero data leakage in reported metrics.

Multimodal Reasoning.

- **MathVista** (Lu et al., 2024) assesses mathematical reasoning across diverse visual formats, including diagrams, charts, geometric figures, and natural images, challenging models to combine visual perception with deep quantitative reasoning. We evaluate on 400 instances uniformly sampled from the 1,000-example `testmini` split.
- **MathVision** (Wang et al., 2024a) presents a challenging multimodal mathematics benchmark sourced from real-world mathematics competitions, covering diverse disciplines that demand fine-grained perception and advanced multi-step deduction. We evaluate on 608 instances uniformly sampled from its 3,040-example test split.
- **We-Math** (Qiao et al., 2025) evaluates visual mathematical problem-solving from a knowledge-centric view, categorizing problems by fine-grained mathematical concepts and their composite relations. We evaluate on 522 instances uniformly sampled from our 1,740-example `testmini` split.
- **MMMU-Pro** (Yue et al., 2025) extends MMMU with rigorous visual reasoning challenges explicitly designed to eliminate language-only shortcuts, demanding genuine cross-modal perception. We evaluate on 461 instances uniformly sampled from the 1,153 examples available in our evaluation split.
- **MMStar** (Chen et al., 2024a) serves as a vision-indispensable benchmark curated across six core capabilities and 18 fine-grained dimensions, enforcing visual reliance to prevent language-prior exploitation. We evaluate on 450 instances uniformly sampled from its 1,500-example validation split.
- **OCRBench** (Liu et al., 2024) benchmarks text-centric visual comprehension across text recognition, scene-text perception, document VQA, and key information extraction. We evaluate on 600 instances uniformly sampled from its 1,000-example test split.

Multimodal Grounding.

- **RefCOCO** (Yu et al., 2016) is a foundational referring-expression comprehension benchmark based on MS-COCO, requiring models to predict spatial coordinates for targeted entities described

in natural language. We evaluate on 1,083 instances uniformly sampled from the 10,834-example validation split.

- **RefCOCO+** (Yu et al., 2016) mirrors the RefCOCO setup while strictly excluding absolute directional language, forcing models to infer localization from visual appearance and relational context. We evaluate on 1,076 instances uniformly sampled from the 10,758-example validation split.
- **RefCOCOg** (Mao et al., 2016) contains more descriptive and syntactically complex referring expressions on MS-COCO, testing nuanced linguistic understanding for spatial grounding. We evaluate on 490 instances uniformly sampled from the 4,896-example validation split.
- **LISA-Grounding (LISA-G)** (Lai et al., 2024; Shen et al., 2025) adapts the complex reasoning segmentation benchmark LISA to bounding-box localization. Featuring implicit instructions that require commonsense deduction, it serves as an out-of-domain grounding testbed following VLM-R1 (Shen et al., 2025). We evaluate on 332 instances uniformly sampled from its 1,658-example test split.
- **Flickr30K Entities** (Plummer et al., 2015) augments natural image captions with fine-grained phrase-to-region alignments across diverse open-domain scenes. We evaluate on 1,448 instances uniformly sampled from our 14,481-example single-box test split.
- **RefCLEF (ReferItGame)** (Kazemzadeh et al., 2014) is a real-world referring benchmark collected via an interactive two-player game, associating informal natural descriptions with image regions. We evaluate on 1,003 instances uniformly sampled from the 10,032-example `refclef_unc_test` split.
- **RIO** (Qu et al., 2023) probes intention-driven grounding in complex scenes. Rather than specifying explicit visual categories, queries outline task intentions or functional requirements that candidate objects must fulfill. We evaluate on 2,934 instances uniformly sampled from the 29,344-example `common_all_thing` evaluation split.

B.5 Computing Resources

All training experiments, covering both baseline reproductions and **VISTA** optimization, were conducted using a single node equipped with $8\times$ NVIDIA H100 80GB HBM3 GPUs, dual Intel(R) Xeon(R) Platinum 8468V CPUs (192 logical cores), and 2.0 TB of system RAM. The software stack is configured with Ubuntu 22.04.4 LTS, NVIDIA Driver 580.95.05, and CUDA 12.4.

Model evaluations were carried out on a dedicated benchmarking node equipped with $8\times$ NVIDIA GeForce RTX 5090 GPUs (32GB VRAM), an Intel(R) Xeon(R) Gold 6530 CPU (128 logical cores), and 512 GB of system RAM, running NVIDIA Driver version 580.95.05 and CUDA 13.0. This environment was utilized exclusively for inference and empirical evaluation across benchmarks, strictly separate from the training infrastructure and bearing no policy optimization workload.

B.6 Baselines

We provide additional details of the baselines used in our grounding and multimodal reasoning evaluations. We include both general-purpose MLLMs and task-specific models to cover representative training paradigms, including supervised instruction tuning, preference optimization, and reinforcement learning.

Multimodal Grounding Baselines.

- **Qwen2.5-VL** (Bai et al., 2025) is a general-purpose MLLM with native dynamic-resolution visual encoding and strong spatial perception. It supports direct object localization through points and bounding boxes and serves as a strong generalist baseline for referring-expression grounding.
- **InternVL2.5** (Chen et al., 2024b) is an open-source MLLM family developed through model, data, and test-time scaling. We include its different model scales to compare against strong general-purpose multimodal representations without grounding-specific reinforcement learning.
- **InternVL2.5-MPO** (Wang et al., 2024b) further post-trains InternVL2.5 with Mixed Preference Optimization (MPO), using large-scale multimodal preference data to jointly model relative response preference, response quality, and preferred generation behavior.

- **Groma** (Ma et al., 2024) introduces localized visual tokenization for region-level understanding. It decomposes images into regions of interest and represents them with region tokens, enabling a unified interface for referring, grounding, and region description.
- **GroundingGPT** (Li et al., 2024) is an end-to-end multimodal grounding model designed for fine-grained localization across images, videos, and audio. It adopts a coarse-to-fine, multi-stage training strategy together with multimodal, multi-granularity grounding data.
- **LLaVA-Grounding** (Zhang et al., 2024) extends large multimodal models with dedicated visual-prompt and grounding modules. By connecting the language model with pretrained grounding and segmentation components, it supports both object-level and pixel-level grounded visual conversations.
- **VisionReasoner** (Liu et al., 2026) formulates detection, segmentation, counting, and related perception tasks within a unified reasoning framework. It uses reinforcement learning with task-aware rewards and multi-object cognitive learning to produce structured reasoning before perception outputs.
- **Dr. Seg** (Sun et al., 2026a) revisits GRPO for perception-oriented tasks and introduces a Look-to-Confirm mechanism together with a Distribution-Ranked Reward. The framework targets fine-grained and stable reinforcement signals for detection and reasoning segmentation without modifying the underlying VLM architecture.
- **RegionReasoner** (Sun et al., 2026b) is an RL framework for region-grounded multi-round visual reasoning. It explicitly grounds reasoning traces with reference regions and combines grounding fidelity with global–local semantic consistency during optimization.
- **Vision-R1** (Zhan et al., 2026) introduces vision-guided reinforcement learning for LVLM alignment. It employs criterion-driven rewards derived from visual-task feedback together with progressive rule refinement, avoiding learned reward models and human preference annotations.
- **VLM-R1** (Shen et al., 2025) extends R1-style reinforcement learning to general vision-language tasks with rule-based rewards. It studies RL on tasks with directly verifiable visual annotations, including referring-expression comprehension and open-vocabulary detection, and serves as our primary outcome-RL baseline.

Multimodal Reasoning Baselines.

- **GPT-5** (Singh et al., 2025) is OpenAI's frontier model system that combines efficient general inference with deeper reasoning for challenging problems. We include it as a proprietary reference for strong multimodal reasoning performance.
- **GPT-4o** (OpenAI, 2024) is an end-to-end multimodal model jointly trained across text, vision, and audio. Its strong general visual understanding and reasoning capabilities make it a representative proprietary multimodal baseline.
- **Gemini 2.5 Pro** (Comanici et al., 2025) is a natively multimodal thinking model with strong reasoning, long-context understanding, and multimodal perception. We use it as another frontier proprietary reference.
- **InternVL2.5** (Chen et al., 2024b) is a general-purpose open-source MLLM family that improves multimodal capabilities through model, data, and test-time scaling. We evaluate multiple model sizes to provide strong non-reasoning-specific baselines.
- **InternVL2.5-MPO** (Wang et al., 2024b) enhances InternVL2.5 using Mixed Preference Optimization and the large-scale MMPR multimodal preference dataset, with particular emphasis on improving multimodal reasoning and chain-of-thought generation.
- **Qwen2.5-VL** (Bai et al., 2025) is a general-purpose open-source MLLM with strong image, document, diagram, and spatial understanding. We include several model scales as representative general MLLM baselines before reasoning-specific post-training.
- **ThinkLite-VL** (Wang et al., 2025c) studies data-efficient reinforcement fine-tuning for visual reasoning. It uses Monte Carlo Tree Search to estimate sample difficulty and select a compact set of challenging training examples, enabling self-improvement without knowledge distillation.
- **NoisyRollout** (Liu et al., 2025b) improves RL exploration by generating rollouts from both clean and moderately perturbed visual inputs. A noise-annealing schedule promotes diverse exploration

early in training while retaining stability at later stages, without changing the underlying RL objective.

- **R1-OneVision** (Yang et al., 2025) develops a cross-modal formalization pipeline that translates visual information into structured textual representations for reasoning. It combines multimodal chain-of-thought supervision with reinforcement learning to improve reasoning across mathematics, science, and general visual understanding.
- **Vision-R1** (Huang et al., 2026) adopts a cold-start-and-RL recipe for multimodal reasoning. It first constructs multimodal chain-of-thought data for initialization and then applies Progressive Thinking Suppression Training and GRPO to improve complex visual reasoning while controlling overthinking.
- **X-Reasoner** (Liu et al., 2025a) investigates whether reasoning learned from text can generalize across modalities. It post-trains a VLM using text-only long-chain-of-thought supervision followed by reinforcement learning with verifiable rewards, and transfers the resulting reasoning ability to multimodal tasks.
- **MM-Eureka** (Meng et al., 2025) extends large-scale rule-based reinforcement learning to multimodal reasoning. It demonstrates that R1-style optimization can elicit stronger reasoning and reflection behaviors in MLLMs using verifiable multimodal rewards.
- **R1-ShareVL** (Yao et al., 2025) proposes Share-GRPO to alleviate sparse rewards and advantage vanishing. It expands each problem into semantically related variants and shares reasoning trajectories and reward information across them, enabling more effective exploration and advantage estimation.
- **VL-Rethinker** (Wang et al., 2025a) targets self-reflective multimodal reasoning through reinforcement learning without reasoning distillation. It introduces Selective Sample Replay to mitigate vanishing advantages and Forced Rethinking to explicitly encourage reflection and self-verification during RL.

B.7 Evaluation Metrics

We use task-appropriate metrics for grounding and multimodal reasoning, and report all results on a percentage scale.

Grounding Metric. For all grounding benchmarks, we use bounding-box accuracy at an IoU threshold of 0.5. Given a predicted box $\hat{b}_i$ and the set of valid ground-truth boxes $\mathcal{G}_i$ for example i, correctness is defined as

$$c_i = \mathbb{I}\left[\max_{b \in \mathcal{G}_i} \mathrm{IoU}(\hat{b}_i, b) > 0.5\right]. \tag{17}$$

The reported score is

$$\mathrm{Acc@IoU} > 0.5 = \frac{100}{N} \sum_{i=1}^{N} c_i, \tag{18}$$

where N is the number of evaluated examples. The threshold is applied strictly as $\mathrm{IoU} > 0.5$. This same bounding-box metric is used for RefCOCO, RefCOCO+, RefCOCOg, LISA-G, Flickr30K Entities, RefCLEF, and RIO. In particular, although LISA-G originates from a segmentation-aware benchmark, we evaluate its bounding-box annotations rather than mask mIoU. For datasets with multiple valid target boxes, including RIO, a prediction is counted as correct if it exceeds the IoU threshold with any valid ground-truth box.

Reasoning Metric. For multimodal reasoning, we report $\mathrm{Pass@1}$ under the evaluation pipeline used in our reasoning framework. Each generated response is converted into a predicted answer by the corresponding answer-extraction and verification procedure, yielding a binary correctness indicator $c_i \in \{0, 1\}$. We compute

$$\mathrm{Pass@1} = \frac{100}{N} \sum_{i=1}^{N} c_i. \tag{19}$$

For OCRBench, we report PASS@1 using the same deterministic decoding and verification pipeline as our other local benchmarks. For each response, the evaluator sequentially parses candidate answers from the final `\boxed{...}` expression, explicit markers (e.g., `Answer:`), the last non-empty line, and the full text. Predictions and references are standardly normalized and marked correct via exact match or flexible containment. All baselines follow this identical protocol to ensure fair comparison.

Average Score. For both grounding and reasoning, the reported **Avg.** is the unweighted macro-average over benchmark-level percentage scores:

$$\text{Avg.} = \frac{1}{M} \sum_{m=1}^{M} s_m, \tag{20}$$

where M is the number of benchmarks and s_m is the percentage score on benchmark m. Consequently, benchmarks with different numbers of evaluation examples contribute equally to the final average.

C HYPERPARAMETERS

C.1 GROUNDING

Grounding experiments are trained with the VLM-R1 framework. The key hyperparameters are summarized in Table 5. For **VISTA**, we select up to $K_t = 8$ coordinate tokens from each generated bounding-box field and retain the top $K_v = 16$ visual tokens according to positive visual credit.

Table 5: Main hyperparameters for grounding experiments.

Hyperparameter	**Value**
Maximum training steps	1,200
Rollouts per prompt	8
Maximum generation length	768
Per-GPU batch size	8
Gradient accumulation steps	2
KL coefficient β	0.04
Alignment coefficient λ_{VISTA}	0.05
Selected output tokens K_t	8
Output-token selector	Coordinate SpreadSample
Selected visual tokens K_v	16

At inference time, we use greedy decoding (`do_sample=False`) with at most 512 newly generated tokens. The per-GPU batch size is 4 for standard benchmarks and 1 for LISA-G.

C.2 MULTIMODAL REASONING

Reasoning experiments are trained with the VL-Rethinker framework using Qwen2.5-VL-7B-Instruct. Table 6 summarizes the key optimization configurations and **VISTA** hyperparameters. Compared with grounding, we employ a larger token budget and execute visual-credit alignment periodically to substantially accelerate training and curb computational overhead. Specifically, with the alignment interval set to four iterations, visual alignment loss and credit attribution are computed only on the first of every four mini-batch actor updates, leaving intermediate steps dedicated solely to the standard RL objective to maximize training throughput.

For reasoning evaluation, we use one deterministic generation per prompt with temperature 0, top-$p = 1.0$, and a maximum generation length of 8,192 tokens.

D ADDITIONAL ABLATION STUDIES

Sensitivity to Loss Weight λ_{VISTA}. Figure 10 sweeps the alignment weight across $\lambda \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$ to characterize its regularization trade-off. The average accuracy

Table 6: Main hyperparameters for multimodal reasoning experiments.

Hyperparameter	**Value**
Training episodes	3
Rollouts per prompt	8
Global training batch size	256
Maximum generation length	3,072
Maximum output tokens	2,048
Actor learning rate	1.0×10^{-6}
KL penalty coefficient	0.02
Alignment coefficient λ_{VISTA}	0.05
Credit target	Positive linear
Selected output tokens K_t	16
Output-token selector	Last-K
Selected visual tokens K_v	64
Alignment interval	Every 4 iterations

follows a clear unimodal trend: it improves from 82.0% at $\lambda = 0.01$ to a peak of 83.3% at $\lambda = 0.05$, before progressively declining to 78.4% as λ reaches 0.5. This performance drop indicates that aggressive regularization overly constrains cross-modal attention, compromising the base model's intrinsic visual comprehension. A moderate range ($\lambda \in [0.02, 0.05]$) maintains an optimal balance, providing targeted visual guidance without perturbing primary policy optimization. For our main experiments, the default setting $\lambda = 0.05$ was selected based on training dynamics, and this offline sweep serves primarily to reveal behavior under varying penalty strengths.

Number of Selected Visual Tokens. Figure 11 examines grounding performance under varying visual token budgets ($K_V \in \{4, 8, 16, 32, 64\}$). Intuitively, constraining supervision to fewer tokens sharpens visual focus onto localized regions, whereas a larger budget disperses attention across broader visual context. Across evaluated benchmarks, average accuracy rises from 81.3% at $K_V = 4$ to 82.7% at $K_V = 8$, peaking at 83.3% with $K_V = 16$. Expanding the budget further leads to a steady decline, reaching 82.6% at $K_V = 32$ and 81.6% at $K_V = 64$. This bell-shaped trajectory reflects an essential trade-off: overly sparse supervision ($K_V \leq 4$) risks omitting informative surrounding context, while excessively broad supervision ($K_V \geq 32$) dilutes the regularization signal away from decisive visual regions.

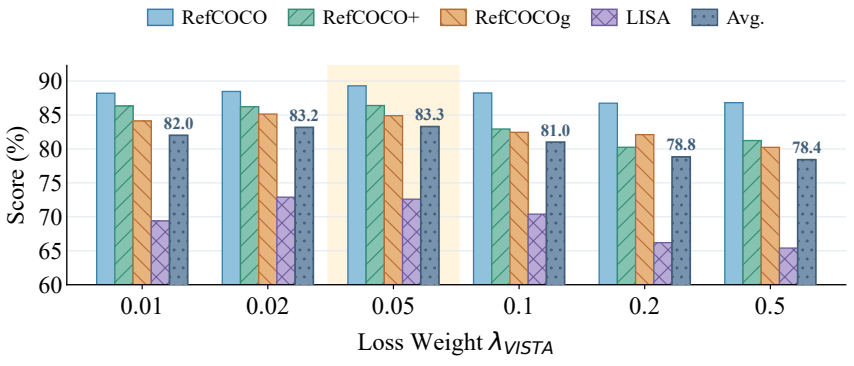


Figure 10: **Sensitivity to loss weight on Qwen2.5-7B-Instruct.** Impact of alignment weight λ_{VISTA} on grounding accuracy across benchmarks.

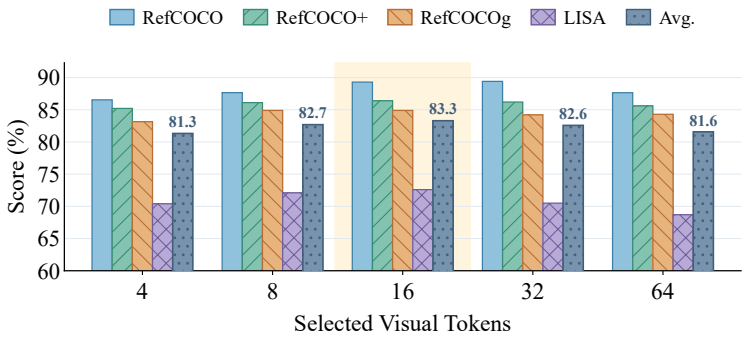


Figure 11: **Effect of visual token budget on Qwen2.5-VL-7B-Instruct.** Impact of the number of selected visual tokens on grounding performance across benchmarks.

E PROMPTS

For referring-expression grounding, we use the same prompt template as VLM-R1 (Shen et al., 2025). The image is provided as a multimodal input block, while the textual instruction is formatted as follows.

Grounding Prompt Template

User Prompt

```
Please provide the bounding box coordinate of the region this sentence
describes: {referring_expression}

First output the thinking process in <think> </think> tags and then
output the final answer in <answer> </answer> tags. Output the final
answer in JSON format.
```

Expected Response Format

```
<think>
{reasoning}
</think>
<answer>
[{"bbox_2d": [x1, y1, x2, y2], "label": "{optional_label}"}]
</answer>
```

For multimodal reasoning, we use the original question text and prompt provided by ViRL39K (Wang et al., 2025a). The image is supplied as a separate multimodal input block, while the textual prompt is formatted as follows.

Reasoning Prompt Template

User Prompt

```
{question}

Please reason step by step, and put your final answer within \boxed{}.
```

Expected Response Format

```
{reasoning}
\boxed{{final_answer}}
```

Here, `{question}` preserves the original problem statement, answer choices, and the associated image placeholder from the dataset. During preprocessing, the image is provided through Qwen's multimodal image content block and aligned with the corresponding visual tokens by the processor. The resulting message is then passed through the native Qwen chat template with the assistant generation prompt appended.

F CASE STUDY

Qualitative Case Study. Figure 12 and Figure 13 illustrate how **VISTA** translates visual credit into precise localization. When identifying the likely driver, both the base model and RL baseline yield less well aligned predictions that include ambiguous margins. In contrast, **VISTA** reasons more distinctly to separate the driver from the passenger, predicting an aligned bounding box. As visualized in Figure 12, while baselines scatter attention across surrounding context, **VISTA** concentrates visual attention directly within the ground truth region (teal contour), confirming that credit-derived guidance effectively grounds decisive reasoning steps in relevant visual evidence.

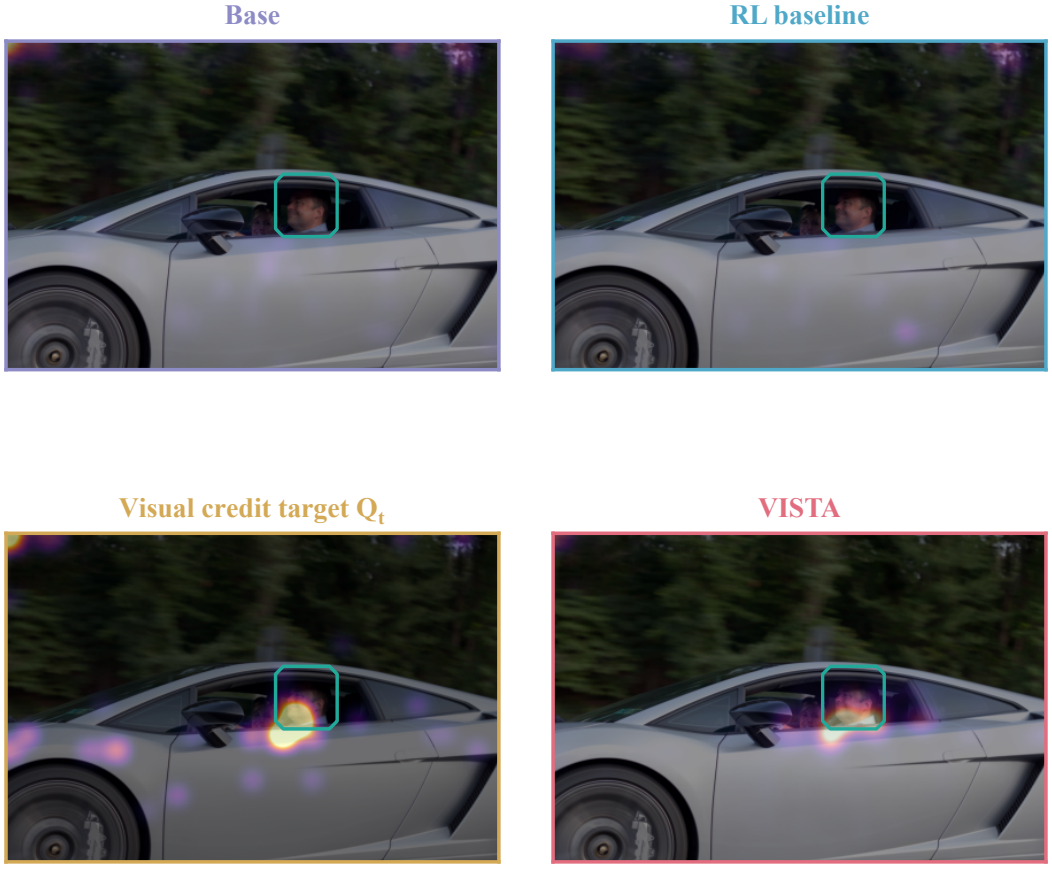
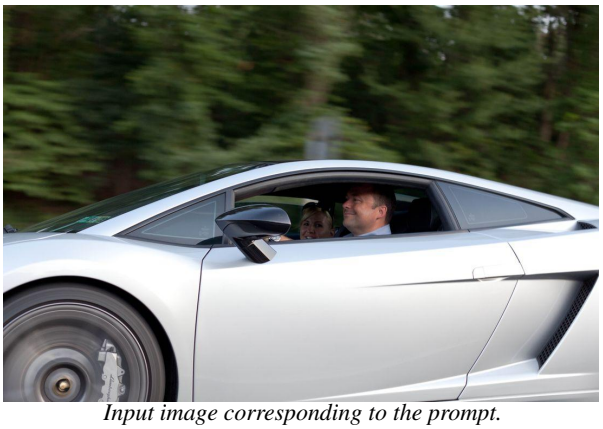


Figure 12: **Case study of visual credit and attention alignment.** The teal contour marks the ground truth region. Compared with both the base model and the RL baseline, **VISTA** concentrates visual attention more consistently on the reward-derived credit target and the referred object.

Prompt

In the picture, there is a stylish and fast car. Who in the picture is most likely the owner or driver of this car? Please provide the bounding box coordinate of this region.



Input image corresponding to the prompt.

Base

Reasoning. The person who appears to be driving the car, as they are located closer to the steering wheel, is likely the owner or driver. The driver is positioned on the right side of the car, which may indicate that this person is more likely to operate or own the vehicle.

Answer.

```
{"bbox_2d": [509, 278, 605, 361], "label": "owner or driver"}
```

RL Baseline

Reasoning. The image shows two individuals inside the car. One appears to be steering while wearing a suit, suggesting that this person is more likely to be the owner or driver. Their position in the car also indicates active participation in operating the vehicle.

Answer.

```
{"bbox_2d": [508, 279, 611, 361], "label": "owner or driver"}
```

VISTA (Ours)

Reasoning. The person closest to the steering wheel, seated on the right, appears to be more actively engaged with the vehicle and is therefore more likely to be the driver. The other person is positioned farther from the controls and is more likely a passenger.

Answer.

```
{"bbox_2d": [480, 279, 605, 361],
"label": "the most likely owner or driver"}
```

Figure 13: **Qualitative case study on intention-aware grounding.** Given the same prompt and input image, **VISTA** produces a more precise localization of the likely owner or driver than the Base and RL baselines.