Vision-Step: Enhancing Visual Focus in Vision-Language Models through Step-wise Grounding Supervision

Anonymous ACL submission

Abstract

Despite their broad success, Vision-Language Models (VLMs) frequently experience "vision collapse" during complex reasoning, abandoning visual evidence for language priors. Existing reinforcement learning methods rely on either sparse outcome-level rewards or computationally expensive Process Reward Models (PRMs). To address this, we propose Vision-Step, a highly efficient RL framework that enforces visual-logical consistency without relying on external PRMs. At its core, the Exclusive Grounding Match (EGM) mechanism directly evaluates intermediate visual grounding through local spatial alignment and a global redundancy check to penalize hallucinations. To stabilize training and prevent premature policy collapse, we introduce a hierarchical advantage fusion strategy that cleanly integrates step-wise process advantages with global outcome rewards. Supported by a newly curated dataset of 15,213 fine-grained annotated instances, Vision-Step effectively mitigates vision collapse while maintaining high training efficiency. Empirical results show that our approach significantly outperforms traditional methods, offering a more effective solution for aligning VLMs.

1 Introduction

Driven by large-scale pretraining, Vision-Language Models (VLMs)(Zhu et al., 2025; Bai et al., 2025a; Hurst et al., 2024) now excel across diverse multimodal tasks, such as visual question answering(Liu et al., 2023), chart understanding(Lee et al., 2023), and document parsing(Blecher et al., 2024). To further unlock their potential in complex reasoning, reinforcement learning (RL)(Ouyang et al., 2022; Rafailov et al., 2023) has emerged as a powerful paradigm. In particular, recent milestones like DeepSeek-R1(Guo et al., 2025) reveal that leveraging verifiable reward signals in RL effectively strengthens chain-of-thought (CoT)(Wei et al., 2022) multi-step reasoning tasks.

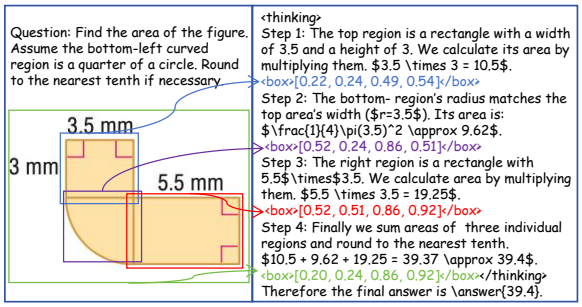


Figure 1: An illustration of the explicit visually-grounded reasoning process in Vision-Step. To mitigate "vision collapse" during complex multimodal tasks, our model explicitly anchors each intermediate logical deduction (e.g., calculating the area of a specific sub-region) to corresponding visual evidence via bounding box coordinates.

Despite these advancements, adapting this RL-driven reasoning paradigm to multimodal contexts reveals a critical bottleneck: existing VLMs struggle to maintain focus on visual inputs during long-horizon reasoning(Li et al., 2023a; Liu et al., 2024a; Lee et al., 2025). We term this phenomenon "vision collapse." As the CoT progresses, the model increasingly prioritizes self-consistent textual continuation over the extraction of fresh visual evidence(Zhao et al., 2025). This tendency primarily stems from the model's heavy reliance on its ingrained language habits and autoregressive generation (Li et al., 2023a; He et al., 2025), leading it to generate convincing descriptions even when visual evidence is weak.

To address this, existing RL alignment methods rely on either sparse or dense supervision, yet both face critical bottlenecks. Sparse outcome-level rewards(Chen et al., 2023; Li et al., 2026b) evaluate only the final prediction, which exacerbates credit assignment issues and inadvertently encourages visual hallucinations—the model guesses the correct textual answer(Yu et al., 2024) while bypassing actual visual evidence. Conversely, dense step-wise

supervision—typically instantiated via Process Reward Models (PRMs)(Lightman et al., 2024; Wang et al., 2024b)—offers fine-grained guidance, but demands prohibitive computational overhead via massive external evaluators or Monte Carlo rollouts(Uesato et al., 2022; Trinh et al., 2024). More crucially, directly injecting these dense step-wise rewards frequently destabilizes autoregressive policies, triggering premature policy collapse(Wu et al., 2023) when models exploit high-variance signals through redundant visual attention.

In response to these challenges, we propose Vision-Step, a novel RL framework designed to explicitly enforce visual-logical consistency in long-horizon reasoning. Vision-Step provides highly efficient, step-wise process supervision directly derived from the model's autoregressive trajectories. To make intermediate visual focus evaluable, Vision-Step decomposes reasoning into steps, each ending with visual bounding box coordinates, as illustrated in Figure 1. Then we introduce the Exclusive Grounding Match (EGM) mechanism to compute visual process rewards. For each reasoning step, EGM evaluates the spatial alignment by calculating the IoU between the predicted bbox and the golden reference bbox, while a global redundancy pool dynamically penalizes hallucinatory or repetitive visual focus. Furthermore, to overcome the premature policy collapse frequently caused by the high variance of dense process rewards, we design a hierarchical advantage fusion strategy. By standardizing process rewards(Shao et al., 2024) within dynamically constructed mini-groups and broadcasting sequence-level outcome advantages to each step, our method harmonizes fine-grained visual exploration with global logical correctness. In summary, our main contributions are:

(i) We release a high-quality dataset of 15,213 multimodal reasoning instances with fine-grained visual grounding annotations, providing a rigorous testbed for step-level visual-logical alignment.

(ii) We propose Vision-Step, a PRM-free RL framework. Its core EGM mechanism directly evaluates intermediate visual grounding, delivering precise step-wise supervision without the massive computational overhead of external evaluators.

(iii) We introduce a hierarchical advantage fusion strategy that cleanly integrates depth-normalized process advantages with global outcome rewards, effectively stabilizing RL dynamics and preventing premature policy collapse.

2 Related Work

2.1 Vision Reasoning Model

To enhance the complex reasoning capabilities of VLMs, early research primarily relied on explicit prompt engineering and supervised fine-tuning (SFT) over curated rationale trajectories (Wei et al., 2022; Kojima et al., 2022; Liu et al., 2023; Dai et al., 2023; Yu et al., 2024). As static SFT proved insufficient for multi-step logic, subsequent paradigms introduced search algorithms and reinforcement learning (Trinh et al., 2024; Sun et al., 2024). Most notably, catalyzed by the recent breakthroughs of DeepSeek-R1 (Guo et al., 2025), a new wave of methodologies has attempted to elicit deeper cognitive exploration in multimodal contexts (Lu et al., 2024; Peng et al., 2025). But existing models predominantly default to a text-centric reasoning paradigm(Li et al., 2026a; Kavuri et al., 2025). By heavily favoring autoregressive textual continuity over active visual evidence extraction (Lee et al., 2025; He et al., 2025; Xu et al., 2025), they fail to achieve the interleaved, visually-grounded reasoning characteristic of human cognition. This fundamental limitation strongly motivates the development of frameworks that explicitly anchor logical deductions to visual evidence.

2.2 Process Rewarding for Reinforcement Learning

Reinforcement Learning (RL) has evolved from human-preference alignment (Ouyang et al., 2022; Rafailov et al., 2023) to a powerful engine for eliciting multi-step reasoning, as demonstrated by models like DeepSeek-R1 (Guo et al., 2025). However, relying on sparse outcome-level supervision creates major scalability bottlenecks and exacerbates credit assignment issues. To provide fine-grained guidance, PRM have emerged, utilizing automated Monte Carlo rollouts (Uesato et al., 2022; Wang et al., 2024b) to bypass costly human annotations.

However, existing PRM implementations suffer from critical structural drawbacks when adapted to multimodal reasoning. First, they heavily depend on massive external reward models (Zhang et al., 2025; Luo et al., 2024) or computationally prohibitive Monte Carlo Tree Search (Cui et al., 2025; Cao et al., 2025). Second, and more crucially, optimizing autoregressive policies with dense process rewards frequently destabilizes RL training dynamics, triggering premature policy collapse (Wu et al., 2023). Vision-Step overcomes this instability by

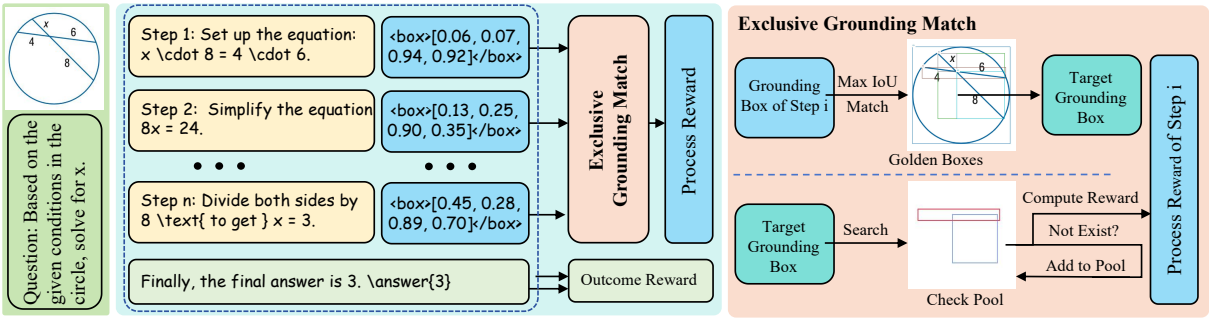


Figure 2: Overview of the reasoning decomposition and the Exclusive Grounding Match (EGM) mechanism. The generated reasoning trajectory is parsed into discrete steps to compute step-wise process rewards. EGM evaluates visual grounding through local spatial alignment (Max IoU) and a global redundancy check (Check Pool) to penalize redundant visual exploration. The final answer strictly determines the sequence-level outcome reward.

designing grounded process rewards that anchor step-wise supervision to explicit visual evidence.

3 Method

3.1 Preliminary

While a SFT cold start is utilized to initialize the basic reasoning structure, we primarily formulate the training of VLMs for complex visual reasoning as a RL problem. Formally, a multimodal prompt, which accommodates multi-image reasoning contexts, is defined as a tuple $x = (I, q)$, where $I = \{i_1, i_2, \ldots, i_{N_v}\}$ represents a collection of N_v input images ($N_v \geq 1$), and q denotes a strictly singular textual query. Given the prompt x, the reasoning policy, parameterized by a large language model with weights θ, is denoted as $\pi_\theta(\tau|x)$. This policy autoregressively decodes a complete reasoning trajectory $\tau = [y_1, y_2, \ldots, y_L] \in \mathcal{Y}^*$, where $\mathcal{Y}$ represents the multimodal vocabulary, y_t is the generated token at time step t, and $\mathcal{Y}^*$ denotes the space of all possible token sequences.

To evaluate the visual grounding capabilities during the thinking process, we formally decompose a complete trajectory τ into a final predicted answer $\hat{a}$ preceded by a sequence of K discrete reasoning steps $\mathcal{S} = \{s_1, s_2, \ldots, s_K\}$. Each intermediate step is formulated as a tuple:

$$s_k = (t_k, \hat{b}_k), \tag{1}$$

where $t_k \in \mathcal{Y}^*$ represents the textual logical thought, $\hat{b}_k \in \mathbb{R}^4 \cup \{\emptyset\}$ denotes the predicted visual grounding box.

To optimize the reasoning policy π_θ, we adopt the Group Relative Policy Optimization (GRPO) framework (Shao et al., 2024). Given a multimodal prompt x, the policy generates a group of G rollout trajectories, denoted as $\mathcal{G} = \{\tau_1, \tau_2, \ldots, \tau_G\}$. To evaluate the global quality of each trajectory $\tau_i \in \mathcal{G}$, we define a sequence-level outcome reward R^i_{out} that encompasses both logical correctness and structural compliance:

$$R^i_{out} = r_{acc}(\hat{a}_i, a^*) + r_{fmt}(\tau_i), \tag{2}$$

where $r_{acc} \in \{0, 1\}$ denotes the accuracy reward, $\hat{a}_i$ is the predicted final answer, a^* represents the ground truth, and r_{fmt} denotes the format reward.

Beyond sequence-level outcomes, we introduce a step-wise process reward to evaluate intermediate grounding quality:

$$R^i_{step,k} = f_{EGM}(s_{i,k}, B^*, \mathcal{S}_{i,<k}), \tag{3}$$

where $s_{i,k}$ is the k-th intermediate reasoning step of the i-th trajectory, B^* is the global set of golden bounding boxes, and $\mathcal{S}_{i,<k}$ denotes the preceding trajectory history up to step $k-1$. f_{EGM} represents our proposed EGM mechanism.

Following the standard GRPO policy optimization, the reward signals R^i_{out} and $R^i_{step,k}$ are respectively transformed into the outcome advantage A^i_{out} and step-wise process advantage $A^i_{step,k}$. These advantages are subsequently fused into a composite token-level advantage $\hat{A}_{i,t}$ to update the policy parameters θ.

3.2 Reasoning Decomposition and Reward Calculation

To enable fine-grained credit assignment across long-horizon visual reasoning tasks, we must first parse the continuous token sequence generated by the policy into discrete, evaluable steps. We achieve this structural predictability by initializing the policy model using the cold start stage. This phase explicitly trains the model to externalize

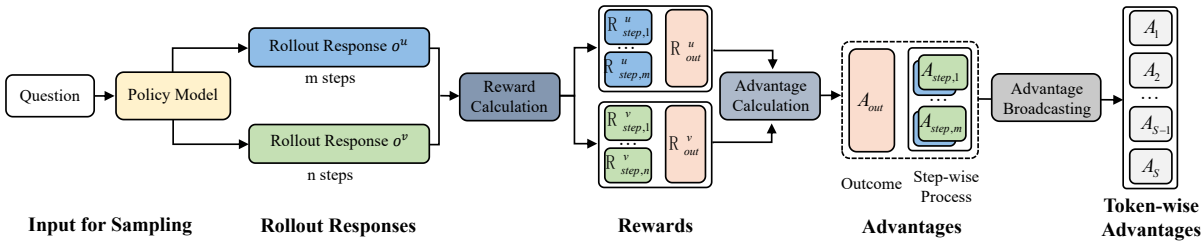


Figure 3: The hierarchical advantage alignment pipeline for policy optimization. To mitigate premature policy collapse caused by variable-length rollouts, we compute global Outcome Advantages alongside step-wise Process Advantages (standardized within dynamic mini-groups of comparable depths). These signals are fused via Advantage Broadcasting to yield dense token-wise advantages for stable GRPO training.

its reasoning process using a predefined template. Specifically, the model is guided to encapsulate its thinking steps and corresponding visual evidence within `<thinking>` tags, followed by the final answer enclosed in `\answer{}`. A standardized trajectory thus follows the format:

```
<thinking>
   Step 1: . . . <box> b̂_1 </box>
   . . .
   Step n: . . . <box> b̂_n </box>
</thinking>
. . . \answer{â}
```

To accurately compute step-wise process rewards, the continuous sequence generated within the `<thinking>` block must be parsed into discrete, evaluable steps. By synchronizing textual step transitions with token-level boundaries, we construct a valid boundary index list $\mathcal{I}_{start}$, constrained by a maximum reasoning depth K_{max}. To ensure strict mathematical alignment across variable-length rollouts, the final valid step count is governed entirely by $\mathcal{I}_{start}$, with trajectories systematically truncated or zero-padded accordingly. Detailed parsing implementations are provided in the Appendix.

Building upon the parsed step sequence $\mathcal{S} = \{s_1, \ldots, s_K\}$, we instantiate the step-wise process reward $R_{step,k}$ for each intermediate step $s_k = (t_k, \hat{b}_k)$ via our proposed Exclusive Grounding Match (EGM) mechanism. To circumvent the labor-intensive requirement of step-aligned bounding box annotations, EGM evaluates spatial grounding accuracy against a global, unaligned pool of golden boxes B^*. For steps entailing purely textual or mathematical deduction without explicit visual grounding ($\hat{b}_k = \emptyset$), we directly assign $R_{step,k} = 0$. Conversely, if $\hat{b}_k \in \mathbb{R}^4$, EGM first performs a **Local Grounding Match** to measure the spatial alignment between the predicted bbox $\hat{b}_k$ and candidates in B^*. The optimal match score M_k and its corresponding target golden box $T_k^* \in B^*$ are identified using the Intersection over Union (IoU) metric:

$$\begin{aligned} M_k &= \max_{b_j^* \in B^*} \mathrm{IoU}(\hat{b}_k, b_j^*), \\ T_k^* &= \arg\max_{b_j^* \in B^*} \mathrm{IoU}(\hat{b}_k, b_j^*). \end{aligned} \tag{4}$$

Second, to mitigate reward hacking—where a policy repeatedly attends to the most salient region across multiple steps to exploit dense rewards—EGM executes a **Global Redundancy Check**. We maintain a dynamic check pool $\mathcal{P}_{<k} = \{T_j^* \mid s_j \in \mathcal{S}_{<k},\ \hat{b}_j \neq \emptyset\}$ archiving previously anchored target boxes. For the current target T_k^*, novel discoveries receive the full match score M_k, whereas targets already existing in $\mathcal{P}_{<k}$ trigger an exclusivity penalty. Formally, the final process reward is computed as:

$$R_{step,k} = M_k \left(1 - \lambda \cdot \mathbf{1}(T_k^* \in \mathcal{P}_{<k})\right), \tag{5}$$

where $\lambda \in (0, 1)$ denotes the penalty coefficient, strictly parameterized to discourage redundant exploration while still tolerating necessary logical backtracking. The check pool is subsequently updated as $\mathcal{P}_k = \mathcal{P}_{<k} \cup \{T_k^*\}$ for succeeding steps.

3.3 Multi-Advantages Alignment and Policy Optimization

While the EGM provides fine-grained step-wise feedback $R_{step,k}$, empirical observations reveal that training an autoregressive policy *purely* on dense process rewards often leads to premature policy collapse. Specifically, the high variance and conflicting gradient signals from intermediate rewards penalize necessary exploratory reasoning steps, causing the policy to degenerate into generating structurally malformed outputs that evade intermediate penalties. To resolve this, we propose a **Hierarchical Advantage Alignment** pipeline. As illustrated in Figure 3, this pipeline safely anchors

step-level visual exploration to global response correctness through *Advantage Calculation* and *Advantage Broadcasting*.

In the *Reward Calculation* stage, for the generated group of G trajectories $\mathcal{G}$, we compute the sequence-level outcome reward R^i_{out} and the step-wise process reward vector $\{R^i_{step,k}\}_{k=1}^{K_i}$ for each rollout τ_i as formulated previously. Figure 3 illustrates this process using two exemplary rollouts with varying reasoning depths. Subsequently, in the *Advantage Calculation* stage, the outcome advantage A^i_{out} is obtained by standardizing the outcome reward across the full group $\mathcal{G}$:

$$A^i_{out} = \frac{R^i_{out} - \mu_{out}}{\sigma_{out}}, \tag{6}$$

where μ_{out} and σ_{out} denote the mean and standard deviation of $\{R^j_{out}\}_{j=1}^{G}$. This yields a single global scalar per trajectory.

However, deriving a stable step-wise advantage is non-trivial due to the variable reasoning step counts across rollouts. To ensure fair step-aligned comparison and mitigate excessive computational overhead, we propose a **Dynamic Survival Normalization** strategy. For each step index k, we define the active rollout subset as $\mathcal{G}_k = \{\tau_j \in \mathcal{G} \mid K_j \geq k\}$, which dynamically filters out trajectories that have already terminated. We then construct the step-specific mini-group $\mathcal{M}_k = \{R^j_{step,k} \mid \tau_j \in \mathcal{G}_k\}$ by horizontally slicing the process rewards of the surviving rollouts. By introducing a minimum survival threshold G_{min}, the step-wise process advantage $A^i_{step,k}$ is standardized exclusively within this active group:

$$A^i_{step,k} = \begin{cases} \dfrac{R^i_{step,k} - \mu_{\mathcal{M}_k}}{\sigma_{\mathcal{M}_k} + \epsilon}, & |\mathcal{G}_k| \geq G_{min} \\ 0, & |\mathcal{G}_k| < G_{min}, \end{cases} \tag{7}$$

where ϵ is a small constant for numerical stability. Notice that when surviving rollouts at extreme reasoning depths fall below G_{min} ($|\mathcal{G}_k| < G_{min}$), group statistics become highly unreliable; we thus conservatively assign zero advantage to these tail steps to prevent uncalibrated gradient explosion, while seamlessly preserving the dense advantage signals for all preceding valid exploration steps.

Finally, the *Advantage Broadcasting* stage fuses these signals into dense, token-wise advantages. Let $k(t)$ denote the reasoning step to which the t-th token belongs. The composite token advantage $\hat{A}_{i,t}$ is computed via a location-parameter shift scaled by a hyperparameter α:

$$\hat{A}_{i,t} = \alpha A^i_{step,k(t)} + A^i_{out}, \tag{8}$$

where $\alpha \in [0, 1]$ explicitly controls the trade-off between step-wise process guidance and global outcome verification. This weighted additive fusion seamlessly compensates tokens with negative process rewards when the final response is correct, preventing premature policy collapse.

To optimize the policy π_θ, we formalize the GRPO objective using a compact theoretical formulation. The final surrogate loss is defined as the expectation over the sampled trajectories:

$$\mathcal{L} = \mathbb{E}_{\tau_i}\left[\sum_{t=1}^{|\tau_i|} \hat{A}_{i,t} \log \frac{\pi_\theta(a_t^{(i)}|x, a_{<t}^{(i)})}{\pi_{\text{ref}}(a_t^{(i)}|x, a_{<t}^{(i)})}\right], \tag{9}$$

where $a_t^{(i)}$ denotes the t-th token in trajectory τ_i, and π_{ref} is the frozen reference policy. This formulation maximizes the advantage-weighted likelihood while implicitly penalizing severe deviations from the reference distribution.

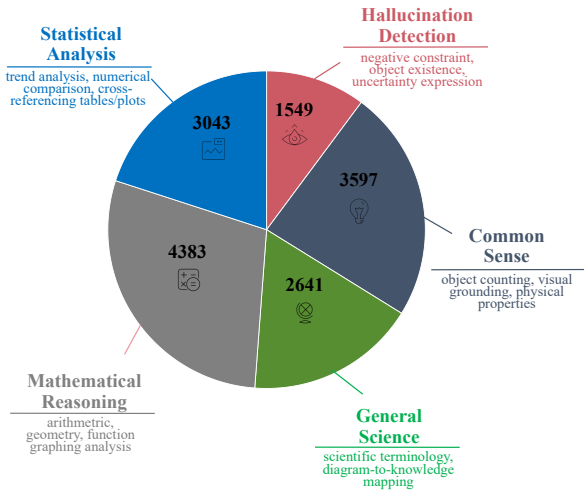


Figure 4: Composition of the multi-domain training corpus. We categorize the data into five core domains: Mathematical Reasoning, Statistical Analysis, General Science, Common Sense and Hallucination Detection.

4 Experiments

4.1 Data Preparation

We curated a multi-domain dataset of 15,213 samples across five specialized dimensions to enhance both fine-grained perception and high-level

Table 1: Quantitative comparison on six general multimodal reasoning and hallucination benchmarks. The best results among open-source models at the same scale are highlighted in **bold**.

Models	MathVista	MathVerse	MathVision	MMMU	MMStar	POPE	Avg.
Close-source Models							
GPT-4o	64.2	43.7	30.6	60.7	65.1	86.9	58.5
Claude-3.5	60.2	41.8	34.1	59.3	58.8	86.2	56.7
Gemini-2.5 Pro	70.3	55.2	36.7	71.2	67.5	91.3	65.4
3B-Scaled Open Vision-Language Models							
Qwen2.5-VL-3B	61.2	30.9	20.2	24.2	53.4	86.1	46.0
Visionary-R1	52.6	35.9	17.2	27.8	44.3	72.4	41.7
ViGoRL-3B	54.5	25.3	19.0	28.1	50.6	87.1	44.1
VLAA-Thinker-3B	61.0	36.4	24.4	26.5	57.3	87.9	48.9
Vision-Step-3B (Ours)	**67.4**	**39.5**	**28.7**	**32.1**	**58.9**	**89.3**	**52.7**
7B-Scaled Open Vision-Language Models							
Qwen2.5-VL-7B	67.1	42.1	25.1	31.1	55.3	87.4	51.4
MM-Eureka-7B	70.4	45.4	28.1	34.9	56.6	86.7	53.7
Vision-R1-7B	68.7	48.2	32.3	32.0	59.2	80.8	53.5
ViGoRL-7B	62.1	33.2	23.3	35.4	52.8	**88.9**	49.3
VLAA-Thinker-7B	68.0	**48.8**	24.0	30.9	59.9	85.8	52.9
Vision-Step-7B (Ours)	**72.4**	46.4	**35.6**	**43.6**	**62.5**	88.5	**58.2**

reasoning in RL the training step. This corpus aggregates high-quality data from the following datasets: mathematical reasoning (CLEVR-Math(Lindström and Abraham, 2022), Geometry3K(Lu et al., 2021)), statistical analysis (FigureQA(Kahou et al., 2017), ChartQA(Masry et al., 2022)), general science (ScienceQA(Saikh et al., 2022)), common sense (Super-CLEVR(Li et al., 2023b), V*(Wu and Xie, 2024)), and hallucination detection (HaloQuest(Wang et al., 2024c)). The domain distribution of these five domains within the overall corpus is illustrated in Figure 4. To enhance the spatial anchoring of the model's reasoning, we employed GPT-5(Singh et al., 2025) to generate precise bbox coordinates for the key elements most likely to be involved in visual reasoning within the data. This automated grounding process ensures that logical deductions—ranging from geometric theorems to chart trends—are directly mapped to specific visual regions, thereby reducing "blind" reasoning.

From this curated corpus, we sampled 9,734 high-quality instances to serve as the Cold-Start SFT training set. This subset was carefully balanced to maintain a diverse representation of task complexities while prioritizing data with verified reasoning chains. By leveraging this curated mixture during the initial supervised fine-tuning phase, the model acquires specialized domain expertise and robust factuality, providing a stable foundation for subsequent alignment and optimization.

4.2 Experimental Setting

Implementation Details. We conducted experiments using Qwen-2.5-VL-3B-Instruct and Qwen-2.5-VL-7B-Instruct. For the training configuration, the SFT phase is conducted for 1 epoch with a batch size of 16, while the RL phase runs for 360 steps with a batch size of 8 and a learning rate of $1 \times e^{-6}$. The rollout numbers for the group (G) and mini-group (G_{mini}) are set to 8 and 4. The α was set to 0.6. We used a default sampling temperature of 1 and a maximum response length of 512 tokens.

Baselines. To comprehensively evaluate our method, we compare Vision-Step against three categories of strong baselines: (1) Closed-source models (GPT-4o(Achiam et al., 2023), Claude-3.5(The), Gemini-2.5 Pro(Comanici et al., 2025)) to benchmark against the commercial upper bound; (2) Base models (Qwen2.5-VL-3B/7B(Bai et al., 2025b)) to quantify the direct reasoning improvements yielded by our training pipeline; and (3) Recent open-source reasoning models at comparable 3B and 7B scales (e.g., Vision-R1(Huang et al., 2025), Visionary-R1(Xia et al., 2025), ViGoRL(Sarch

Table 2: Quantitative comparison on fine-grained visual perception, document understanding, and logical reasoning benchmarks. The best results among open-source models are highlighted in **bold**.

Models	ChartQA Pro	OCRBench	LogicVista	HallusionBench	DocVQA
GPT-4o	41.7	736	65.1	56.2	92.8
Claude-3.5	53.7	798	55.2	54.8	95.2
Gemini-2.5 Pro	60.2	832	73.8	63.7	96.7
Qwen2.5-VL-7B	46.7	804	44.1	51.7	93.0
MM-Eureka-7B	39.9	797	43.7	56.3	89.6
Vision-R1-7B	39.7	**834**	45.4	58.2	90.1
ViGoRL-7B	48.7	752	46.2	61.3	88.2
VLAA-Thinker-7B	45.5	768	48.5	64.0	94.0
Vision-Step-7B (Ours)	**51.1**	829	**50.2**	**69.2**	**94.3**

et al., 2026), VLAA-Thinker(Chen et al., 2025), MM-Eureka-7B(Meng et al., 2025)) to demonstrate our superiority over the latest RL-based and R1-style multimodal reasoning models.

4.3 Main Results

Results on General Benchmarks. We evaluate Vision-Step across six general multimodal benchmarks: MathVista(Lu et al., 2024), MathVerse(Zhang et al., 2024), MathVision(Wang et al., 2024a), MMMU(Yue et al., 2024), MMStar(Chen et al., 2024), POPE(Li et al., 2023a). For all these benchmarks, we adopt accuracy as the unified evaluation metric and employ the same system prompt across all benchmarks. As shown in Table 1, Vision-Step establishes new state-of-the-art results across both scales. Vision-Step-3B achieves the highest 3B average score (52.7), outperforming its base model by +6.7. Vision-Step-7B dominates the 7B category with an average of 58.2.

Crucially, Vision-Step excels on complex cognitive tasks: Vision-Step-7B scores 72.4 on MathVista and 43.6 on MMMU, with the former surpassing even GPT-4o (64.2) and Claude-3.5 (60.2). This leap validates that our hierarchical advantage alignment effectively optimizes long-horizon reasoning trajectories. Concurrently, top-tier POPE scores (89.3 for 3B, 88.5 for 7B) confirm that our EGM mechanism effectively prevents visual hallucinations. By forcing intermediate thoughts to anchor on concrete visual evidence, EGM prevents over-reliance on spurious linguistic priors. Ultimately, these results prove our method unlocks deep reasoning while preserving foundational perception reliability.

Results on Specific Domain Benchmarks. To further evaluate the model's proficiency in fine-grained visual perception, document understanding, and rigorous logical deduction, we evaluate the 7B model on five specialized benchmarks: ChartQA Pro(Masry et al., 2025), OCRBench(Liu et al., 2024b), LogicVista(Xiao et al., 2024), HallusionBench(Guan et al., 2024), and DocVQA(Mathew et al., 2021). Regarding the evaluation metrics, we report the absolute score (out of 1000) for OCRBench, while standard accuracy is adopted for the remaining four benchmarks.

Beyond overall metrics, Table 2 highlights Vision-Step-7B's proficiency in scenarios demanding deep structural and logical comprehension. Its exceptional capability in parsing dense text and structural data directly validates the efficacy of the proposed EGM mechanism; by explicitly grounding intermediate thoughts to visual coordinates, the model avoids losing context in visually cluttered environments. Furthermore, its strong resistance to multimodal hallucinations, as evidenced by its performance on LogicVista and HallusionBench, underscores a critical advantage of our hierarchical RL framework: it fosters genuine logical deduction rather than superficial pattern matching. Even on pure text recognition tasks (OCRBench), Vision-Step-7B maintains highly competitive robustness against both proprietary and specialized open-source models, proving that its reasoning gains do not come at the cost of foundational perception. Collectively, these capability-driven improvements confirm that Vision-Step successfully bridges the gap between low-level visual perception and high-level reasoning.

Table 3: Ablation study of Vision-Step-7B on core multimodal reasoning benchmarks. We systematically evaluate the performance drops when removing key components.

	MathVista	MathVision	MMMU
Qwen2.5-VL-7B	67.1	25.1	31.1
Vision-Step-7B	72.4	35.6	43.6
Δ w/o Adv. Fusion	69.7	32.0	41.4
Δ w/o Step reward	70.2	28.2	39.6
Δ w/o RL	70.1	26.9	40.1
Δ w/o SFT	68.3	25.7	37.5

4.4 Ablation Study

We perform a systematic ablation study on three multimodal reasoning benchmarks, aiming to isolate the individual contributions of each proposed module in Vision-Step-7B. As detailed in Table 3, the severe degradation upon removing either the SFT cold-start or the subsequent RL optimization phase underscores that high-quality initialization and RL-driven exploration are strictly complementary for eliciting complex reasoning. Crucially, ablating the proposed EGM step reward exposes the vulnerability of long-horizon multimodal tasks to reward sparsity, proving our dense visual-logical supervision is essential for sustained reasoning. Finally, the consistent performance decline observed without Advantage Fusion validates its design intent: effectively mitigating gradient variance and stabilizing the composite optimization process.

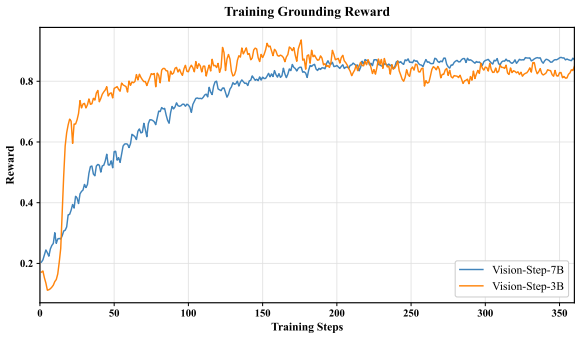


Figure 5: Training Grounding Reward curves during RL optimization. Vision-Step-7B demonstrates smoother learning and superior stable convergence compared to the more volatile 3B model.

We track the training optimization trajectories across different model scales (Figure 5) to evaluate the stability and effectiveness of our proposed step-level grounding reward during reinforcement learning. Ultimately, both the Vision-Step-3B and Vision-Step-7B models successfully achieve robust convergence, demonstrating the universal reliability of our reward design. Specifically, while the 3B model exhibits a rapid initial surge followed by slight mid-stage variance, the 7B model maintains a smoother and highly stable ascent toward its final reward plateau. These converging trends validate that our step-wise alignment provides consistent optimization signals, with the larger parameter capacity of the 7B model naturally affording superior training stability.

Figure 6: Composition of the multi-domain training corpus. We categorize the data into five core domains: Mathematical Reasoning, Statistical Analysis, General Science, Common Sense and Hallucination Detection.

Figure 6 tracks the evolution of visual attention via step-wise Grounding Scores on MathVision. While the Qwen-VL baseline suffers a precipitous early drop and severe variance—indicating rapid vision-language detachment and high susceptibility to hallucinations—Vision-Step-7B maintains stable, high-fidelity grounding throughout the multi-step process. This persistent attention confirms that our method effectively mitigates modality decoupling, keeping the model firmly anchored to visual evidence across extended reasoning horizons.

5 Conclusion

We present Vision-Step, a PRM-free method that aligns step-wise visual grounding with sparse outcome rewards. Vision-Step evaluates intermediate grounding scores, aligns them with outcome advantages, and delivers stable per-token guidance. Experiments on multimodal reasoning benchmarks show consistent gains over strong baselines while retaining efficiency, thereby establishing a scalable and robust foundation for future agents.

Limitations

While our EGM heavily relies on 2D bboxes to anchor reasoning, this spatial representation is inherently optimized for discrete objects or localized regions. For abstract visual concepts—such as overall image style, complex textures, or holistic chart trends—forcing the model to predict rigid box coordinates may not fully capture the continuous nature of the visual evidence. Exploring more flexible grounding representations (e.g., segmentation masks or latent patch attention) for abstract reasoning remains an important direction for future work.

To enable stable step-wise reward assignment, Vision-Step relies on a strictly predefined SFT formatting template (e.g., explicit <thinking> and <box> tags). While this structural rigidity is crucial for preventing premature policy collapse during RL, it inherently constrains the free-form generative diversity of the base language model. This trade-off suggests that achieving a balance between structured verifiable reasoning and open-ended creative problem-solving is a subtle challenge in VLM alignment.

References

The claude 3 model family: Opus, sonnet, haiku.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Ke qin 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, and 8 others. 2025b. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923.

Lukas Blecher, Guillem Cucurull Preixens, Thomas Scialom, and Robert Stojnic. 2024. Nougat: Neural optical understanding for academic documents. In *International Conference on Learning Representations*, volume 2024, pages 37646–37663.

Lang Cao, Renhong Chen, Yingtian Zou, Chao Peng, Huacong Xu, Yuxian Wang, Wu Ning, Qian Chen, Mofan Peng, Zijie Chen, and 1 others. 2025. More bang for the buck: Process reward modeling with entropy-driven uncertainty. *arXiv preprint arXiv:2503.22233*.

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. 2025. Sft or rl? an early investigation into training r1-like reasoning large vision-language models. *arXiv preprint arXiv:2504.11468*.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Ganqu Cui, Lifan Yuan, Zefan Wang, Hanbin Wang, Yuchen Zhang, Jiacheng Chen, Wendi Li, Bingxiang He, Yuchen Fan, Tianyu Yu, and 1 others. 2025. Process reinforcement through implicit rewards. *arXiv preprint arXiv:2502.01456*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14375–14385.

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

Yixiao He, Haifeng Sun, Qi Qi, Zirui Zhuang, Pengfei Ren, Huazheng Wang, Yafeng Nan, and Jingyu Wang. 2025. Mitigating object hallucination in large vision-language models via visual attention direct preference optimization. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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

Vivek Hruday Kavuri, Vysishtya Karanam Karanam, Venkamsetty Venkata Jahnavi, Kriti Madumadukala, Balaji Lakshmipathi Darur, and Ponnurangam Kumaraguru. 2025. Freeze and reveal: Exposing modality bias in vision-language models. In *Proceedings of Interdisciplinary Workshop on Observations of Misunderstood, Misguided and Malicious Use of Language Models*, pages 17–26.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. 2025. Vlind-bench: Measuring language priors in large vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4129–4144.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. 2023. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR.

Jiaqi Li, Guangming Wang, Shuntian Zheng, Minzhe Ni, Xiaoman Lu, Guanghui Ye, and Yu Guan. 2026a. Towards mitigating modality bias in vision-language models for temporal action localization. *arXiv preprint arXiv:2601.21078*.

Tianle Li, Jihai Zhang, Yongming Rao, and Yu Cheng. 2026b. Unveiling the compositional ability gap in vision-language reasoning model. *Advances in Neural Information Processing Systems*, 38:126574–126592.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023a. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305.

Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. 2023b. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14963–14973.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let's verify step by step. In *International Conference on Learning Representations*, volume 2024, pages 39578–39601.

Adam Dahlgren Lindström and Savitha Sam Abraham. 2022. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *International Conference on Learning Representations*, volume 2024, pages 57689–57733.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

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

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

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6774–6786.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Meiqi Guo, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, and 1 others. 2024. Improve mathematical reasoning in language models by automated process supervision. *arXiv preprint arXiv:2406.06592*.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and

logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, and 1 others. 2025. Chartqapro: A more diverse and challenging benchmark for chart question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19123–19151.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, and 1 others. 2025. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. 2025. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301.

Gabriel Sarch, Snigdha Saha, Naitik Khandelwal, Ayush Jain, Michael Tarr, Aviral Kumar, and Katerina Fragkiadaki. 2026. Grounded reinforcement learning for visual reasoning. *Advances in Neural Information Processing Systems*, 38:150977–151013.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2024. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110.

Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. 2024b. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439.

Zhecan Wang, Garrett Bingham, Adams Wei Yu, Quoc V Le, Thang Luong, and Golnaz Ghiasi. 2024c. Haloquest: A visual hallucination dataset for advancing multimodal reasoning. In *European Conference on Computer Vision*, pages 288–304. Springer.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Penghao Wu and Saining Xie. 2024. V?: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36:59008–59033.

Jiaer Xia, Yuhang Zang, Peng Gao, Sharon Li, and Kaiyang Zhou. 2025. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. *arXiv preprint arXiv:2505.14677*.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2087–2098.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and 1 others. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, and 1 others. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.

Zhenru Zhang, Chujie Zheng, Yangzhen Wu, Beichen Zhang, Runji Lin, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2025. The lessons of developing process reward models in mathematical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10495–10516.

Jianfei Zhao, Feng Zhang, Xin Sun, Lingxing Kong, Zhixing Tan, and Chong Feng. 2025. Cross-image contrastive decoding: Precise, lossless suppression of language priors in large vision-language models. *arXiv preprint arXiv:2505.10634*.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Dataset Construction Details

A.1 RL Dataset

First, we need to identify which are the key elements in image-text reasoning. To ensure precise and standardized spatial evaluation during the intermediate reasoning steps, we formally define the concept of **Visual Primitives**. We refer to visual primitives as the fundamental, irreducible visual elements within an image that serve as the spatial anchors for the model's logical deduction. Rather than predicting coarse, image-level bounding boxes, our Vision-Step framework requires the policy to ground its intermediate thoughts to these fine-grained visual primitives.

Depending on the distinct characteristics of the multimodal reasoning tasks (e.g., geometry, charts, real-world scenes, and logical puzzles), we systematically categorize the visual primitives space $\mathcal{V}$ into the following specific domains:

- **General Image Analysis:** For broad image comprehension, the primitives include:
  - *Highlighted Region:* Shaded areas, filled contours, or explicitly color-coded zones that draw visual attention.
  - *Sub-figure:* Inset images, zoomed-in local patches, or individual panels within multi-view composite figures.
- **Geometry Deduction:** For mathematical geometry problems, the primitives strictly correspond to geometric entities:
  - *Line & Angle:* Line segments (edges, chords, intersecting lines), and angles (including specific right-angle squares and interior/exterior arcs).
  - *Shape:* Polygons (triangles, quadrilaterals) and standard circles/ellipses.
- **Charts and Function Graphs:** For data visualization interpretation, the primitives are decomposed into structural and data-carrying elements:
  - *Structural Element:* Coordinate axes (X/Y axes, tick marks) and legends.
  - *Data Representation:* Histogram bars, pie chart sectors, and polyline segments in trend charts.
  - *Data Marker:* Key data points, discrete scatter marks, or annotated extreme values.

- **Real-world Scenes:** Primitives consist of distinct physical objects, foreground subjects, or specific structural background regions (e.g., doors, tables, screens) interacting within the scene.

- **Logical Puzzles:** For abstract visual reasoning (e.g., Raven's Progressive Matrices), the primitives denote individual grid cells, abstract symbolic patterns, or state indicators.

By explicitly mapping the model's intermediate reasoning textual tokens to these well-defined visual primitives, our EGM mechanism can rigorously calculate the step-wise process rewards $R^i_{step,k}$ without ambiguity.

Constructing high-quality, step-aligned multimodal reasoning trajectories is fundamentally challenging. In our pilot studies, we observed that attempting to generate both the logical reasoning steps and the corresponding visual grounding coordinates simultaneously (i.e., joint generation) consistently yielded suboptimal results. The competing objectives of complex logical deduction and precise spatial localization often led to severe context dilution and hallucinated bounding boxes.

While state-of-the-art proprietary models (e.g., GPT-5) demonstrated a certain capacity for isolated grounding generation, the pass rate across diverse open-source models remained inadequate without extensive over-generation and filtering. To guarantee the highest caliber of training data, we ultimately designed a rigorous, decoupled **Two-Stage Data Construction Pipeline** incorporating human-in-the-loop verification:

- **Stage 1: Textual Reasoning Generation (Step-Thinking):** In the first phase, we strictly isolate the logical deduction process. The model is prompted to generate intermediate reasoning steps using the standardized `<thinking>` template. To ensure data reliability, we implement a strict outcome-filtering mechanism: only trajectories that arrive at the exact correct final answer are retained. Crucially, to prevent reward hacking (where a model arrives at the correct answer via flawed logic), all retained samples undergo manual verification to identify and eliminate logical mismatches (false-positive trajectories).

- **Stage 2: Visual Grounding Alignment (Step-Grounding):** Building upon the verified logical trajectories from Stage 1, we subsequently generate the corresponding visual grounding boxes for each reasoning step. Because spatial alignment is highly susceptible to subtle visual noise, we enforce a zero-tolerance policy for unverified spatial data. Every generated bounding box is rendered visually onto the source image, and human annotators exhaustively verify the spatial-logical alignment. Any bounding box that fails to accurately encapsulate the referenced visual primitive is manually corrected or discarded.

This decoupled, visually-verified pipeline ensures that our Vision-Step framework is initialized with dense, mathematically rigorous, and fully explainable reasoning trajectories, establishing a solid foundation for the subsequent RL optimization.

A.2 SFT Cold-Start Data Construction

The Supervised Fine-Tuning (SFT) phase is critical for initializing the policy model to output structurally compliant multimodal reasoning trajectories. To construct a high-quality cold-start dataset, we employ a rigorous processing pipeline that standardizes trajectory lengths and explicitly aligns logical steps with visual grounding boxes.

First, to ensure training stability and prevent the model from getting lost in excessively long reasoning loops, we perform **Trajectory Length Normalization**. We identified a marginal subset of the raw data (approximately 1.14%) that exhibited extreme verbosity, containing reasoning chains exceeding 10 steps. We utilized GPT-5.2 to summarize and rewrite these outliers into concise, high-density reasoning trajectories bounded to a maximum of 7 steps, thereby homogenizing the learning distribution.

Next, we enforce strict **Format Alignment**. The data is structured into a standard conversational format. The user instruction is constructed by concatenating the multimodal prompt (the image and the problem text) with a predefined template guiding the step-by-step generation. For the target output, we mandate that every intermediate reasoning step must be strictly followed by a single visual primitive bounding box, formatted as `<box>[x1, y1, x2, y2]</box>`. To handle edge cases where the number of parsed textual steps and available bounding box annotations mismatch in the source data, we apply a heuristic padding strategy: required grounding boxes are sequentially sampled or randomly drawn from the valid annotation pool to

maintain format consistency without breaking the autoregressive sequence. The trajectory is strictly terminated with the final answer encapsulated in `\answer{}`.

B Inference and RL Prompt Template

To elicit structured step-by-step reasoning and precise visual grounding during both the Reinforcement Learning (RL) and inference phases, we format the input prompt using a standardized conversational template. The prompt encapsulates the system role, the multimodal context (images and text), and explicit generation constraints.

Crucially, to ensure strict consistency with our EGM evaluation mechanism (described in Section 3.2), the prompt explicitly instructs the model to format its internal monologue using `<thinking>` tags, output spatial coordinates using `<box>`, and wrap the final solution in `\answer{}`.

Below is the complete prompt template utilized in our framework, instantiated with a geometric reasoning example:

```
<|im_start|>system
You are a helpful assistant.<|im_end|>
<|im_start|>user
<|vision_start|><|image_pad|><|vision_end|>
Find y.
Solve the problem step by step based
on the image.  Write your reasoning
as Step 1, Step 2, etc.  At the end
of EACH step, provide EXACTLY ONE
grounding annotation: <box>[x1, y1, x2,
y2]</box>. Grounding rules: 1. If the
step uses a visual element, ground the
most important one.  2.  If the step
does not use any visual element, ground
a new image element that has NOT been
grounded before. After all steps, give
the final answer in \answer{}.
You FIRST think about the reasoning
process as an internal monologue and
then provide the final answer.  The
reasoning process MUST BE enclosed
within <thinking> </thinking> tags.
The final answer MUST BE put in
\answer{}.<|im_end|>
<|im_start|>assistant
```

C Details of Step Decomposition

To accurately distribute the dense process rewards calculated by our Exclusive Grounding Match (EGM) mechanism during Reinforcement Learning (RL) training, we implement a bidirectional alignment pipeline. This pipeline standardizes step boundaries across both the natural language text view and the tokenized sequence view.

During the textual reward evaluation pass, the response is scanned sequentially to extract reasoning segments. The extraction protocol operates as follows:

1. **Sequential Scanning:** The parser identifies the initial anchor string `"Step 1"` and retrieves the first visual grounding bounding box immediately following it. If the model fails to output a valid grounding format within this step's context, a dummy box $[0, 0, 0, 0]$ is assigned as a fallback to maintain structural alignment.

2. **Iterative Parsing:** This procedure loops iteratively for subsequent steps (i.e., `"Step 2"`, `"Step 3"`, ...) until either the step index exceeds the maximum threshold ($k > 10$) or the k-th step identifier is missing within the text.

3. **Single-Step Fallback:** In cases where the initial `"Step 1"` anchor cannot be retrieved anywhere in the response, the entire natural language text is conservatively treated as a single monolithic reasoning step for process reward evaluation.

To apply the per-token advantage broadcasting described in Section 3.3, we must locate the exact token indices where each reasoning step initiates, producing a standardized sequence of boundary indices denoted as *start_list*. Due to tokenization variations in the underlying BPE tokenizer, a step prefix can materialize in two distinct tokenized forms depending on the preceding context (i.e., with or without a leading space).

To ensure pixel-perfect localization, we track the tokenized representations of `"Step` k`"` for $k \in [1, 10]$. Specifically, for a given step k, we look up two primary token ID patterns:

$$P_{\text{space}}(k) = [8304,\ 220,\ 15 + k] \tag{10}$$

$$P_{\text{no_space}}(k) = [14822,\ 220,\ 15 + k] \tag{11}$$

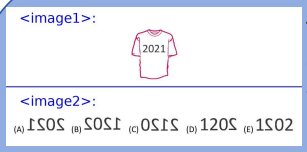
where the terminal token ID codes the step digit (e.g., ID `16` corresponds to digit `1`, ID `17` corresponds to digit `2`, and so forth). The parser searches for the first occurrence of these token patterns sequentially from step 1 to 10. Similar to the text-level fallback, if token pattern $P(1)$ is absent, index `0` is appended as the unique element of *start_list*, designating the entire sequence as a single step.

D Qualitative Results

<image>Which place is located in row C?\n\nChoices:\nA. the diner\nB. the grocery store\nC. the library\nD. the school\n

<thinking>Step 1: Understand the coordinate system of the grid map. The rows are labeled alphabetically from top to bottom (A, B, C) on the left and right margins, while the columns are numbered 1 to 4 from left to right. <box>[0.000, 0.000, 1.000, 1.000]</box>\nStep 2: Locate row C, which is the bottom-most row of the grid. It spans across grid cells C1, C2, C3, and C4. <box>[0.667, 0.000, 1.000, 1.000]</box>\nStep 3: Identify the buildings present within the cells of row C. Cell C1 contains the Police Department, cells C2 and C3 collectively contain the School, and cell C4 contains the Gas Station. <box>[0.667, 0.286, 1.000, 0.714]</box>\nStep 4: Evaluate the given choices against the detected elements in row C. Among the options, \"the school\" (spanning C2 and C3) is the only building listed that resides in row C. <box>[0.667, 0.286, 0.900, 0.580]</box></thinking>\nThe final answer is \\answer{D}"

Figure 7: A qualitative example of geometric reasoning.



<image>Ella puts on this t-shirt and stands in front of a mirror. Which of these images does she see in the mirror?\n<image1>\n<image2>\n\nChoices:\nA. \nB. \nC. \nD. \nE. E\n\n

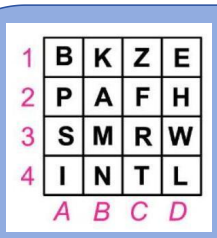
<thinking>Step 1: Identify the text \"2021\" on the shirt. A plane mirror horizontally flips both the character sequence and the individual character shapes. <box>[0.485, 0.150, 0.702, 0.505]</box>\nStep 2: Reverse the sequence from \"2021\" to \"1202\". Next, mirror each character: '1' flips backward, '2' becomes an 'S' shape, and '0' remains symmetric. The visual result is a backward '1', 'S', '0', 'S'. <box>[0.010, 0.605, 0.990, 0.950]</box>\nStep 3: Compare this result with the given options. Option (A) exactly matches this derived pattern. <box>[0.035, 0.750, 0.220, 0.885]</box></thinking>\nThe final answer is \\answer{A}"

Figure 8: A qualitative example of spatial logic decoding.

<image>According to the map provided, is New Mexico state larger than California state?\n\nChoices:\nA. Yes\nB. No\n\n

<thinking>Step 1: Locate the state of California on the western coast of the map and observe its visual area. <box>[0.330, 0.030, 0.640, 0.170]</box>\nStep 2: Locate the state of New Mexico in the southwestern region of the map and observe its visual area. <box>[0.620, 0.260, 0.870, 0.440]</box>\nStep 3: Compare the geographical footprints of both states. California clearly covers a more extensive spatial area on the map than New Mexico, meaning the answer is no. <box>[0.330, 0.030, 0.870, 0.440]</box></thinking>\nThe final answer is \\answer{B}"

Figure 9: A qualitative example of macro-level map comprehension.



<image>Tom encodes words using the board shown. For example, the word PIZZA has the code A2A4C1C1B2. What word did Tom encode as B3B2C4D2?\n\nChoices:\nA. MAZE\nB. MASK\nC. MILK\nD. MATE\nE. MATH\n\n

<thinking>Step 1: Identify the Column-Row coordinate system used on the board. <box>[0.082, 0.084, 0.918, 0.960]</box>\nStep 2: Decode the first coordinate B3. Locating column B and row 3 gives the letter 'M'. <box>[0.336, 0.470, 0.490, 0.630]</box>\nStep 3: Decode the second coordinate B2. Locating column B and row 2 gives the letter 'A'. <box>[0.336, 0.288, 0.490, 0.448]</box>\nStep 4: Decode the remaining coordinates C4 and D2. Column C row 4 is 'T', and column D row 2 is 'H'. <box>[0.490, 0.288, 0.796, 0.812]</box>\nStep 5: Combine all the decoded letters sequentially to form the word \"MATH\", which corresponds to option E. <box>[0.336, 0.288, 0.796, 0.812]</box></thinking>\nThe final answer is \\answer{E}"

Figure 10: A qualitative example of visual relation grounding.