V-ERR: VISUAL EVIDENCE RE-CHECKING REINFORCEMENT IN MULTIMODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite their strong perception and generation abilities, vision-language models (VLMs) still struggle with complex reasoning. While GRPO-style post-training reinforcement learning has improved multimodal reasoning performance, it primarily optimizes external outcomes and underutilizes the model's internal modality signals, especially the visual signals that should be tracked and verified across reasoning steps. To address these challenges, we draw inspiration from cognitive science on multimodal signal processing, and uncover two key phenomena in VLMs: signal partitioning, where modality signals are processed through partitioned pathways, and signal attenuation, where modality-specific signals progressively decays as reasoning unfolds. Building on these insights, we propose V-ERR (**V**isual **E**vidence **R**e-checking **R**einforcement), a post-training framework that reinforces attenuated signals during reasoning. Concretely, we first introduce a partitioning score to identify modality-specialized attention pathways and then incorporate GRPO with modality-specialized reward shaping to reinforce evidence-checking to strengthen the attenuated visual-modality signals. Extensive experiments demonstrate that V-ERR significantly enhances the reasoning capabilities of VLMs, consistently outperforming leading RL optimization strategies by a substantial margin. The code is available at https://anonymous.4open.science/r/V-ERR-BB96.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs), especially vision-language models (VLMs), have achieved remarkable success in bridging visual perception with linguistic intelligence (Peng et al., 2024; Xu et al., 2025; Chen et al., 2024). By coupling a vision encoder with a large language model, these architectures effectively unify visual and textual representations, enabling performance on complex reasoning tasks (Zhang et al., 2024b; Yao et al., 2025). However, despite these impressive advances, MLLMs remain prone to severe reliability issues, particularly in multi-step reasoning scenarios (Zheng et al., 2024; He et al., 2025), where errors compound across steps and small early deviations can derail the final conclusion (Bai et al., 2025; Plaat et al., 2025; Liu et al., 2025a; Wang et al., 2025d).

To improve reasoning capabilities, post-training has increasingly shifted toward reinforcement learning (RL) as a dominant paradigm (Zhang et al., 2024a; Chen et al.; Deng et al., 2025; Wang et al., 2025c; Peng et al., 2025). In particular, the emergence of policy-optimization methods such as Group Relative Policy Optimization (GRPO) (Shao et al., 2024) has further solidified RL's leading role in improving reasoning performance (Tan et al., 2025b; DeepSeek-AI, 2025; Zhou et al., 2025b). These approaches, especially in multimodal settings, have largely converged on GRPO-style, with key design differences primarily reflecting how rewards are assigned across the generation trajectory. One line of work operates at the sequence level, assigning a single reward per complete output (Shao et al., 2024; Yu et al., 2025; Liu et al., 2025b; Xiao et al., 2025). Another line of work operates at the token level, shaping rewards at individual steps or tokens to guide intermediate decisions rather than only the final outcome (Tan et al., 2025a; Lin et al., 2025a; Lee & Tong, 2025; Qiu et al., 2025). However, despite their flexibility in objective design and consistent gains through direct behavior shaping, most existing approaches remain largely black-box, primarily optimizing external outcomes without insight into internal reasoning failures. This limitation becomes even more pronounced in VLMs, where modality signals are transient and unevenly utilized, so outcome-only rewards provide noisy

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.

credit assignment and may encourage shortcut solutions, ultimately pushing further improvements in reasoning toward a bottleneck (Wen et al., 2025; Zhou et al., 2025a).

To address the above challenges, we draw inspiration from cognitive science, where signals from different modalities are often processed through partitioned pathways (Driver & Spence, 1998; Watanabe & Shimojo, 2005; Cloutman, 2013; Freud et al., 2023), and are recruited repeatedly over time to support sustained inference (Miller & Cohen, 2001; Edelman & Gally, 2013; Eriksson et al., 2015; Panagiotaropoulos, 2024). Guided by this perspective, we conduct a fine-grained internal attention analysis to trace how visual and textual signals evolve inside VLMs during reasoning, aiming to compare their processing patterns with those in biological systems. Through this analysis, we identify two key internal phenomena: signals in multimodal models exhibit partitioning along the width dimension and attenuation along the depth dimension. **Signal partitioning** refers to the tendency for different modality signals to become fixed into particular internal structures for processing. Specifically, visual information is consistently processed through only a small subset of specialized attention heads rather than being broadly accessible across the network. **Signal attenuation** denotes the weakening of certain modality signals as the reasoning trajectory progresses. In particular, visual evidence is engaged primarily in the early stage of generation and then diminishes or disappears in later steps. These findings suggest that while VLMs exhibit a partitioned processing pattern similar to pathway-separated architecture in biological cognition, modality-specific signals, particularly visual ones, progressively attenuate over the course of reasoning, potentially causing critical evidence to attenuate or even vanish from the trajectory.

Building on this insight, we propose V-ERR (*Visual Evidence Re-checking Reinforcement*), a post-training framework that enhances multimodal reasoning by drawing inspiration from biological cognition, in which modality signals are processed through partitioned pathways and task-relevant evidence is recurrently re-engaged over time to counteract signal attenuation. Concretely, we first introduce a partitioning score to quantify signal partitioning in VLMs, namely which attention heads consistently specialize in processing language versus vision, allowing us to identify a compact set of modality-specialized heads that mediate text-to-vision information flow. We then apply GRPO as the policy optimizer, using reward shaping that explicitly encourages repeated, decisive evidence-checking through these partitioned pathways, thereby counteracting modality attenuation, especially visual attenuation, along the generation trajectory. V-ERR requires no additional annotations and introduces a limited amount of extra computation, while better aligning VLMs with the multimodal reasoning process observed in biological cognition. Experiments on benchmarks such as MathVista and MathVision, conducted with MLLMs including LLaVA and Qwen-VL, show that compared with strong GRPO-based RL baselines, V-ERR achieves consistent improvements in multimodal reasoning performance.

Our contributions include: (1) Inspired by cognitive science, we reveal two key phenomena in VLMs, namely width-wise signal partitioning and depth-wise signal attenuation, which characterize how modality signals are internally processed and how they attenuate over a reasoning trajectory. (2) Building on these observations, we propose V-ERR, which leverages the signal-fixation structure to identify modality-specialized pathways and augments GRPO with reinforcement that specifically strengthens attenuated modality signals, leading to substantial gains in multimodal reasoning. (3) Through extensive experiments on multimodal benchmarks such as MathVista and MathVision with two mainstream multimodal models (LLaVA and Qwen-VL), we show that V-ERR better aligns MLLMs with these cognitive phenomena and consistently outperforms strong GRPO-based baselines.

2 Related Work

Token-level mechanisms in generation. Mechanistic interpretability has increasingly studied token-by-token generation dynamics, seeking to explain how internal states and context positions steer decoding. A first family is output-oriented and decomposes the final prediction into contributions from intermediate representations and components, providing readouts of how internal computations aggregate into the output distribution (Katz & Belinkov, 2023; Ferrando et al., 2023; Geva et al., 2021). A second family is next-token-oriented and uses causal interventions to pinpoint which activations or context positions drive a particular next-token choice (Meng et al., 2022; Heimersheim & Nanda, 2024; Zhang & Nanda, 2023). A third family is trajectory-oriented and identifies token-level control phenomena that persist across steps, such as attention concentrating on early tokens and stable anchor

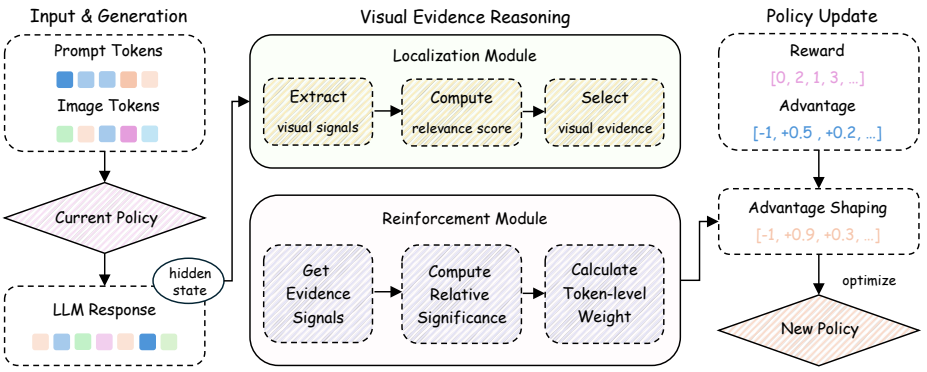


Figure 1: Overview of the proposed V-ERR (*Visual Evidence Re-checking Reinforcement*) framework for localizing critical visual evidence and reinforcing recurrent evidence re-checking in VLMs.

tokens that scaffold later inference (Xiao et al., 2024; Gu et al., 2025; Barbero et al., 2025; Men et al., 2024; Bogdan et al., 2025). A fourth family is exploration-oriented and treats certain tokens as decision points for exploration (Lin et al., 2025b; Vassoyan et al., 2025; Li et al., 2025; Cui et al., 2025; Wang et al., 2025b). Despite these advances, prior work primarily contributes analysis for token-level attribution and control, whereas our method uncovers modality-specific partitioning and attenuation in VLMs and translates them into a post-training procedure that promotes sustained multimodal evidence use.

GRPO-style RL for VLM Reasoning. Recent reinforcement learning for multimodal reasoning has increasingly adopted GRPO-style, critic-free group-based optimization as a practical alternative to PPO, with most differences stemming from how rewards are assigned across the generation trajectory (Wang et al., 2025a; Shen et al., 2025; Zhou et al., 2025c; Meng et al., 2025). At the sequence level, GRPO and its engineering-oriented variants such as DAPO optimize a single completion-level objective and estimate baselines from grouped samples to reduce training cost and improve stability, making them a common choice in reasoning-focused pipelines (Shao et al., 2024; Yu et al., 2025; Liu et al., 2025b; Xiao et al., 2025). At the token level, GRPO-style extensions introduce denser supervision by shaping rewards over steps or tokens, aiming to guide intermediate decisions rather than only the final outcome (Tan et al., 2025a; Lin et al., 2025a; Lee & Tong, 2025; Qiu et al., 2025). More recently, several multimodal RL methods further incorporate visual perception grounding signals into policy optimization (Jian et al., 2025b; Wang et al., 2025e; Kancheti et al., 2026). Despite their effectiveness, existing GRPO-based methods still primarily optimize externally defined reward or grounding signals and leave the internal failure modes of multimodal reasoning under-specified, whereas our approach explicitly targets modality-specific signal partitioning and attenuation and uses GRPO with pathway-aware reward shaping to encourage sustained reasoning.

3 The Anatomy of Modality Signal Dynamics

To overcome the shared limitation of existing RL-based post-training, namely optimizing only external outcomes while remaining largely blind to internal modality failures, we turn to insights from cognitive science, where signals from different modalities are processed through partitioned pathways and task-relevant evidence is recurrently re-engaged over time to support sustained inference. Guided by this perspective, we conduct a mechanistic dissection of VLMs' internal attention dynamics along multi-step reasoning trajectories. Our analysis uncovers two key phenomena: **width-wise signal partitioning** and **depth-wise signal attenuation**.

3.1 Signal Partitioning

Motivated by the cognitive-science view that multimodal signals may be processed through partitioned pathways, we ask whether textual and visual signals in VLMs are broadly distributed across the

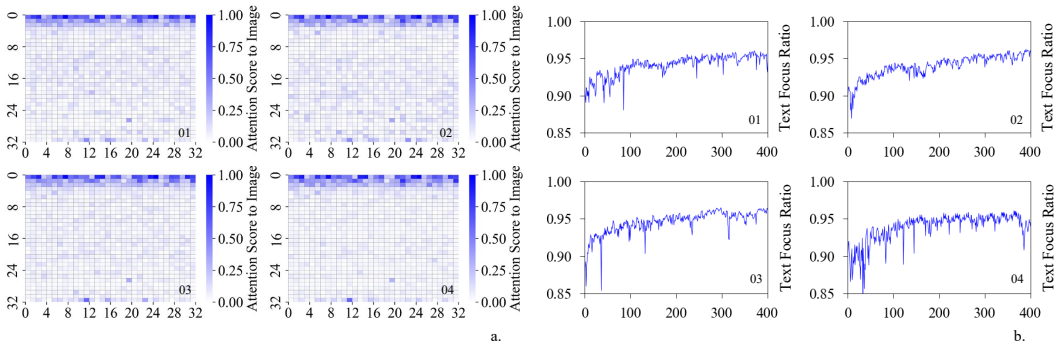


Figure 2: **a.** Visual Propensity Score. The x-axis represents Head IDs and the y-axis represents Layer IDs. Brighter colors indicate higher Visual Propensity. The heatmaps are shown for four randomly sampled instances. Additional examples and statistical analysis are provided in the Appendix A. **b.** Four samples of the Visual Signal Attenuation Effect. The Text Focus Ratio (y-axis) climbs steadily as the response token index (x-axis) increases. More examples and statistical analysis are provided in the Appendix A.

network or concentrated in a small set of internal components. To this end, we characterize each attention head by how strongly it directs information toward textual and visual tokens during reasoning. Specifically, we define the **Visual Propensity (VP)** and **Textual Propensity (TP)** of head h in layer l as the fractions of its attention mass assigned to image tokens and input text tokens, respectively:

$$VP^{(l,h)} = \frac{\sum_{t\in R}\sum_{v\in V_{\text{img}}} A_{t,v}^{(l,h)}}{\sum_{t\in R}\sum_{k\in V_{\text{all}}} A_{t,k}^{(l,h)}}, \qquad TP^{(l,h)} = \frac{\sum_{t\in R}\sum_{u\in V_{\text{text}}} A_{t,u}^{(l,h)}}{\sum_{t\in R}\sum_{k\in V_{\text{all}}} A_{t,k}^{(l,h)}}, \tag{1}$$

where $A_{t,k}^{(l,h)}$ denotes the attention weight from response position t to key position k, R is the set of response tokens, V_{img} and V_{text} denote the image-token set and input text-token set, respectively, and V_{all} denotes the full context-token set. Intuitively, a higher VP or TP indicates that the attention head relies more strongly on visual or textual information when processing the response.

In practice, we run the VLM on substantial instances and collect attention maps during autoregressive decoding. We then compute $VP^{(l,h)}$ and $TP^{(l,h)}$ for every head by aggregating attention mass over the full generated response tokens. The resulting statistics reveal an asymmetric partitioning pattern between the two modalities. For textual signals, most heads exhibit consistently high TP values, indicating that textual context is broadly accessible across the network and does not form a clearly sparse partition. In contrast, visual signals show a much more concentrated pattern. As shown in Figure 2.a, the VP heatmap reveals that visual information is not uniformly accessible, but is processed through a sparse and structured subset of heads. In particular, only a small fraction of heads account for the majority of attention mass directed to image tokens and we refer to these high-VP heads as visual heads, or more formally, **visual pathways**. Due to space limitations, we provide supporting evidence in Appendix A, including large-scale VP heatmap analyses across different model architectures and quantitative metrics comparison on the separation between visual and non-visual heads.

This asymmetry suggests that, unlike textual information, which is broadly distributed across attention heads, text-to-vision information flow is mediated by a compact, modality-specialized substrate. Therefore, the signal partitioning phenomenon in VLMs mainly manifests in the visual modality: visual signals depend on a small set of specialized pathways, making them potentially more vulnerable during long-horizon reasoning.

3.2 Signal Attenuation

Having established that only visual signals are processed through partitioned pathways along the width dimension, we next examine how these signals evolve along the depth dimension of multi-step reasoning. Hence, we introduce a diagnostic that compares the model's reliance on textual versus visual signals at each generation step.

For a generated response sequence R with tokens $o_1, \ldots, o_T$, we compute the **Text Focus Ratio (TFR)** at each step t, which quantifies the fraction of attention mass allocated to textual input as opposed to visual. Let $A^{(h)}_{t,\text{text}}$ denote the aggregate attention weight assigned to input prompt text tokens, and $A^{(h)}_{t,\text{img}}$ denote the attention weight assigned to image tokens, for head h at step t. Aggregating over all heads H, we define:

$$TFR_t = \frac{\sum_{h \in H} A^{(h)}_{t,\text{text}}}{\sum_{h \in H} \left(A^{(h)}_{t,\text{text}} + A^{(h)}_{t,\text{img}} \right)}. \tag{2}$$

Correspondingly, the visual focus ratio can be written as $VFR_t = 1 - TFR_t$. A higher TFR indicates stronger reliance on textual input, while a higher VFR indicates stronger reliance on visual input.

We trace the decoding process and compute step-wise modality focus profiles on many instances and the results show an asymmetric pattern. As shown in Figure 2.b, textual signals remain stable or even become increasingly dominant as generation proceeds, meaning that visual signals will attenuate steadily over later reasoning steps. Namely, the model gradually drifts toward a text-dominant regime, and image tokens become progressively underutilized or practically ignored. Similarly, more experimental results and related analyses can be found in the Appendix A.

3.3 Implications for Post-Training Design.

Collectively, width-wise signal partitioning and depth-wise signal attenuation provide a concrete design principle for post-training. Partitioning shows that only visual signals are processed through a compact subset of modality-specialized pathways. Attenuation further implies that the main temporal failure mode is not a symmetric loss of all modality signals, but a disproportionate weakening of visual signals as reasoning progresses. Therefore, our method focuses on visual pathways because visual signals constitute the more vulnerable bottleneck in multimodal reasoning. An effective intervention should strengthen the pathways that carry visual information and explicitly incentivize recurrent, late-stage evidence rechecking through these pathways, preventing the model from drifting into a text-dominant regime during long-horizon reasoning.

4 Method

Motivated by our goal of turning internal modality-signal diagnostics into actionable post-training improvements, we propose V-ERR (*Visual Evidence Re-checking Reinforcement*). First, to extract effective visual signals through visual pathways to localize the critical visual evidence that genuinely supports multi-step reasoning from dense image tokens, we introduce a **Localization** module (Section 4.1) that filters a task-relevant visual evidence set. Next, to mitigate the failure mode of visual signal attenuation, we introduce a **Reinforcement** module (Section 4.2) that explicitly encourages visual evidence re-checking.

4.1 Localization

Goal. Vision encoders typically represent an image as a dense grid of tokens, but for any given reasoning problem only a small subset constitutes task-relevant evidence. Our first module aims to distill such visual evidence.

Key idea. Building on the *signal partitioning* finding in Section 3.1, we locate evidence by reading out attention specifically from the modality-specialized visual pathways that mediate text-to-vision information flow and using the prompt semantics to determine which visual tokens are relevant of the task.

Step 1: Extract Visual Modality Signals. Let $\mathcal{H}_{\text{vis}}$ denote the set of visual pathways identified by high visual propensity (VP, as defined in Section 3.1) among the attention heads. Rather than aggregating attention over all heads, where signals from multiple modalities are mixed, we form a consolidated attention matrix only on this partitioned substrate:

$$\bar{A} = \frac{1}{|\mathcal{H}_{vis}|} \sum_{(l,h) \in \mathcal{H}_{vis}} A^{(l,h)}, \tag{3}$$

where $A^{(l,h)} \in \mathbb{R}^{N \times N}$ is the attention map of attention head h in layer l and N is the full sequence length.

Step 2: Compute Task-Relevant Scores. Next, we define the relevance of each image token v based on how strongly it is interrogated by the textual prompt. Let $\mathcal{P}$ denote the set of indices corresponding to the prompt tokens, and $\mathcal{V}_{img}$ denote the indices of the image tokens. We calculate the **Task-Relevant Score** (S_v) for each image token $v \in \mathcal{V}_{img}$ by summing the attention it receives from all prompt tokens:

$$S_v = \sum_{p \in \mathcal{P}} \bar{A}_{p,v}. \tag{4}$$

This formulation ensures that the score is dynamic and query-dependent; an image token representing a red shirt will have high S_v if the question asks about clothing, but low S_v if the question concerns the background scenery.

Step 3: Select Visual Evidence Set. Finally, we discretize S_v into a binary mask by selecting the top-K image tokens to form the visual-evidence set:

$$\mathcal{V}_{\text{evi}} = \{\, v \in \mathcal{V}_{\text{img}} \mid S_v \geq \tau_K \,\}, \tag{5}$$

where τ_K is the K-th largest score. This filtering step effectively removes background noise, ensuring that subsequent reinforcement targets re-checking of semantically meaningful visual evidence, rather than indiscriminately increasing attention to the entire image.

4.2 REINFORCEMENT

Goal. Having localized task-relevant visual evidence, our second module reinforces re-checking of it during multi-step generation to counteract the signal attenuation observed in 3.2.

Key idea. Instead of a brittle global threshold on attention, we detect locally salient evidence re-engagement relative to the surrounding temporal context, and shape token-level credit only at those moments to amplify evidence-checking signals in order to encourage this kind of behaviour.

Step 1: Compute Visual Evidence Score. For each generated token o_t in the response, we compute a focused visual-evidence score by aggregating attention mass directed to the visual evidence:

$$a_t = \sum_{v \in \mathcal{V}_{\text{evi}}} \bar{A}_{t,v}, \tag{6}$$

where $\bar{A}$ is the modality-specialized attention aggregation from visual pathways. This definition filters out attention to irrelevant image regions and tracks only task-relevant evidence.

Step 2: Calculate Relative Significance. Because attention magnitudes vary across prompts and reasoning stages, we assess whether a_t is locally salient by normalizing it with a sliding-window statistic. Let μ_t and σ_t be the mean and standard deviation of $\{a_{t'}\}$ within a window of size W centered at t. The local normalized score is defined as:

$$z_t = \frac{a_t - \mu_t}{\sigma_t + \epsilon}. \tag{7}$$

We treat a token as exhibiting *evidence re-checking* if it satisfies relative significance ($z_t > \delta$) and absolute validity ($a_t > \eta$), where δ prevents rewarding routine fluctuations and η filters numerical leakage when visual signals are near-zero.

Step 3: Shape Token-level Advantage. We build on GRPO and introduce a scaling coefficient γ_t to shape the advantage $\hat{A}_i$ at the token level:

$$\tilde{A}_{i,t} = \gamma_t \cdot \hat{A}_i, \tag{8}$$

where

$$\gamma_t = 1 + (\gamma_{\text{amp}} - 1) \cdot \mathbb{I}[z_t > \delta \ \wedge \ a_t > \eta]\,. \tag{9}$$

Intuitively, γ_t increases credit specifically at tokens where the model exhibits a locally salient reliance on visual evidence, encouraging evidence-checking and counteracting the signal attenuation observed in Section 3.2. For training stability, we normalize γ_t over the sequence to keep $\mathbb{E}_t[\gamma_t] = 1$, preventing global reward inflation.

Table 1: Results on Benchmarks.

Method	MathVista	CSVQA	MathVision	ScienceQA	MMMU	Avg.
LLaVA-1.5-7B						
GRPO	31.5	23.7	15.1	66.8	31.9	33.8
DAPO	35.6	22.4	13.9	66.9	32.1	34.2
Dr. GRPO	31.8	21.7	15.6	66.5	31.5	33.4
BNPO	32.6	24.5	12.7	66.6	31.9	33.7
Rhythm	36.3	20.7	14.8	68.1	32.2	34.4
V-ERR	**38.2** +6.7	**26.0** +2.3	**16.0** +0.9	**70.2** +3.4	**33.5** +1.6	**36.8** +3.0
Qwen2.5-VL-3B-Instruct						
GRPO	64.8	31.9	22.3	80.5	41.9	48.3
DAPO	62.9	30.9	16.6	79.5	42.7	46.5
Dr. GRPO	61.8	33.4	26.2	80.2	43.1	48.9
BNPO	65.5	30.9	21.7	80.2	43.8	48.4
Rhythm	67.4	29.9	23.8	78.9	42.5	48.5
V-ERR	**67.8** +3.0	**35.7** +3.8	**26.5** +4.2	**80.6** +0.1	**44.0** +2.1	**50.9** +2.6

5 EXPERIMENTS

In this section, we evaluate the performance of the proposed V-ERR by addressing the following three research questions:

- RQ1: Does V-ERR consistently improve multimodal reasoning performance across diverse benchmarks and model families, and does it remain stable to optimize?
- RQ2: How do key design factors influence the performance?
- RQ3: Does V-ERR actually induce visual evidence rechecking behaviors, rather than only improving final accuracy?

5.1 EXPERIMENTAL SETUP

Benchmarks. We test on question-answering (QA) benchmarks and challenging mathematical reasoning datasets, including CSVQA (Jian et al., 2025a), ScienceQA (Lu et al., 2022), MathVision (Wang et al., 2024), MathVista (Lu et al., 2024) and MMMU (Yue et al., 2024). These benchmarks collectively offer a diverse and rigorous evaluation suite, including multimodal scientific understanding with diagram-aware questions and advanced mathematical problem-solving that requires multi-step symbolic reasoning, visual interpretation, and integration of domain knowledge, which enables us to evaluate model performance from factual retrieval and logical inference to complex quantitative and perceptual reasoning.

Baselines. To validate the effect of our partition-conditioned evidence reinforcement, we implement V-ERR on top of GRPO and use GRPO as the primary baseline. Beyond GRPO, we include three widely used strong GRPO-style variants: DAPO, which focuses on stabilizing large-scale RL training by addressing issues such as reward noise and training instability (Yu et al., 2025); Dr. GRPO, which corrects biases in GRPO to improve token efficiency while maintaining reasoning performance (Liu et al., 2025b); and BNPO, which introduces beta-based normalization to reduce variance and generalizes GRPO under binary reward settings (Xiao et al., 2025). We also compare to the recent attention-structure-guided fine-grained policy optimization method *Rhythm*, which identifies preplan and anchor tokens via attention-derived metrics and reallocates token-level credit to these critical nodes (Li et al., 2025).

Backbone Models and Implementation Details. Experiments are conducted using the LLaVA-1.5-7B (Liu et al., 2023) and Qwen2.5-VL-3B-Instruct (Team, 2025a) models. We retain the full input prompt without explicit truncation and implement our training pipeline on the TRL framework (von Werra et al., 2020) by extending its *GRPOTrainer*. More details can be found in the Appendix.

Table 2: Results for different top-k ratios of Visual Pathways.

Method	MathVista	CSVQA	MathVision	ScienceQA	MMMU
			LLaVA-1.5-7B		
GRPO	31.5	23.7	15.1	66.8	31.9
10% Heads	37.8 +6.3	25.2 +1.5	15.1 +0.0	69.1 +2.3	33.3 +1.4
20% Heads	**38.2 +6.7**	**26.0 +2.3**	**16.0 +0.9**	**70.2 +3.4**	**33.5 +1.6**
30% Heads	35.6 +4.1	24.5 +0.8	15.7 +0.6	68.7 +1.9	32.2 +0.3
			Qwen2.5-VL-3B-Instruct		
GRPO	64.8	31.9	22.3	80.5	41.9
10% Heads	67.8 +3.0	**36.7 +4.8**	22.9 +0.6	80.0 -0.5	42.6 +0.7
20% Heads	**67.8 +3.0**	35.7 +3.8	**26.5 +4.2**	**80.6 +0.1**	**44.0 +2.1**
30% Heads	64.4 -0.4	32.1 +0.2	25.0 +2.7	80.5 +0.0	42.9 +1.0

5.2 Evaluation of Effectiveness & Efficiency (RQ1)

Results on Benchmarks. Table 1 and Figure 8 (shown in the appendix) reports the final accuracy and reward curves on the benchmarks respectively. Across all datasets, our proposed V-ERR achieves the best performance, consistently outperforming the GRPO baseline and its extensions, and the reward curves indicate that V-ERR does not make optimization harder to converge. In particular, V-ERR yields substantial gains over GRPO on MathVista (+6.7%) and ScienceQA (+3.4%) in LLaVA, and on MathVision (+4.2%) and CSVQA (+3.8%) in Qwen, where effective alignment between visual evidence and multi-step reasoning is critical. Building upon Rhythm (Li et al., 2025), which introduces rhythm-aware credit assignment for structured reasoning, V-ERR extends this paradigm to the multimodal setting by explicitly leveraging partitioned visual pathways and shaping vision-conditioned credit to reward recurrent evidence-checking, leading to consistent performance improvements across all benchmarks and indicating that modality-aware credit provides more precise optimization signals for multimodal reasoning. Other GRPO variants, including DAPO, Dr. GRPO, and BNPO, exhibit limited or dataset-dependent improvements, whereas V-ERR maintains stable gains across diverse tasks, highlighting its robustness and general effectiveness.

Answer to RQ1. V-ERR yields consistent gains across models and datasets while maintaining smooth reward convergence, suggesting that our method is both effective and optimization-friendly.

5.3 Evaluation of Design Factors (RQ2)

In this section, we analyze how V-ERR's performance varies with the amount of selected visual pathways and critical evidence, and whether task-relevance conditioning is necessary.

Top-K ratios of Visual Pathways. Table 2 reports V-ERR under different top-K ratios for selecting Visual Pathways, evaluated on multiple benchmarks with LLaVA-1.5 and Qwen2.5-VL. We observe a clear sweet spot that selecting a moderate fraction generally yields the best results. Using too few pathways underestimates the partitioned substrate and misses useful vision-routing capacity, while including too many introduces noise that dilute the evidence signal used for shaping.

Top-K ratios of Visual Evidence. Table 3 varies the top-K ratio for selecting Visual Evidence. A moderate ratio of about 15% performs best, suggesting that effective reinforcement requires a compact yet sufficient evidence set. When K is too small, the anchor set becomes overly selective and may drop necessary supporting regions, leading to brittle re-checking. When K is too large, anchors start to include background and distractors, causing the re-checking signal to become less targeted and weakening the advantage shaping effect.

The Necessity of Task-relevance. To validate the design of V-ERR, we ask whether the gains come from merely amplifying any strong visual engagement, or from reinforcing re-checking of task-relevant evidence. We introduce a variant termed **Focus**, which rewards visual signals purely based on its spatial concentration, without conditioning on the task-relevant visual evidence. As

Table 3: Results for different top-k ratios of Visual Evidence.

Method	MathVista	CSVQA	MathVision	ScienceQA	MMMU
			LLaVA-1.5-7B		
GRPO	31.5	23.7	15.1	66.8	31.9
5% Anchors	35.6 +4.1	24.0 +0.3	**18.1 +3.0**	68.4 +1.6	33.0 +1.1
15% Anchors	**38.2 +6.7**	**26.0 +2.3**	16.0 +0.9	**70.2 +3.4**	**33.5 +1.6**
30% Anchors	36.0 +4.5	22.7 -1.0	15.4 +0.3	67.8 +1.0	32.0 +0.1

Table 4: Pure Visual Focus vs V-ERR.

Method	MathVista	CSVQA	MathVision	ScienceQA	MMMU
			LLaVA-1.5-7B		
GRPO	31.5	23.7	15.1	66.8	31.9
Focus	34.4 +2.9	23.5 -0.2	15.3 +0.2	68.0 +1.2	32.7 +0.8
V-ERR	**38.2 +6.7**	**26.0 +2.3**	**16.0 +0.9**	**70.2 +3.4**	**33.5 +1.6**

shown in Table 4, while encouraging the model to focus intently on something in the image, rather than distributing weak visual signals uniformly, is mostly beneficial for reasoning, V-ERR consistently achieves superior performance, indicating that localizing task-relevant visual evidence is essential.

Answer to RQ2. V-ERR benefits from a moderate amount of pathway capacity and a compact evidence set, and crucially depends on task-relevance conditioning.

5.4 Evaluation of Evidence Rechecking Behavior (RQ3)

To examine whether V-ERR promotes repeated use of visual evidence during generation, we visualize the model's attention to image tokens. For each selected example, we extract attention maps at three representative stages of the response—Early, Middle, and Late, and compare the base model with the V-ERR-tuned model, as shown in Figure 7. The base model often shows strong attention to the image at the beginning, but this attention becomes increasingly diffuse or vanishes as generation proceeds, reflecting the tendency of visual evidence usage to diminish in later steps. In contrast, the V-ERR-tuned model exhibits more persistent attention in the middle and late stages. The results align with our motivation: enforcing visual evidence rechecking helps counteract the late-stage attenuation of visual signal utilization during multi-step reasoning.

Answer to RQ3. Attention visualizations show that V-ERR induces revisits to task-relevant regions, consistent with the intended evidence-rechecking mechanism, which is additionally supported by Appendix C.4. However, improved attention alone is insufficient; reasoning gains emerge only when recurrent evidence engagement is coupled with outcome-driven supervision. Accordingly, V-ERR optimizes attention not as an end in itself, but as a mechanism to improve evidence utilization. Case studies provided in the Appendix D further demonstrate this point.

6 Conclusion

Motivated by cognitive science, we investigate why VLMs still struggle with multi-step reasoning, and uncover two internal phenomena that help explain this gap: width-wise signal partitioning, where visual signals are processed through a small set of modality-specialized attention heads, and depth-wise signal attenuation, where visual evidence is actively used early but progressively weakens later in the reasoning trajectory. Hence we propose V-ERR, a post-training framework that identifies visual pathways and integrates GRPO-based policy optimization with reward shaping that encourages decisive evidence-checking. V-ERR requires no extra supervision and introduces a modest amount of computational overhead. Experiments on multimodal benchmarks demonstrate consistent improvements over strong GRPO-style RL baselines, suggesting that explicitly reinforcing recurrent engagement of visual evidence is a practical approach to improving multimodal reasoning.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the writing. We have not used generative AI tools for other tasks. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The code is available at https://anonymous.4open.science/r/V-ERR-BB96.

REFERENCES

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey, 2025. URL https://arxiv.org/abs/2404.18930.

Federico Barbero, Álvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. Why do llms attend to the first token?, 2025. URL https://arxiv.org/abs/2504.02732.

Paul C. Bogdan, Uzay Macar, Neel Nanda, and Arthur Conmy. Thought anchors: Which llm reasoning steps matter?, 2025. URL https://arxiv.org/abs/2506.19143.

Liang Chen et al. R1-v: Reinforcing super generalization ability in visionlanguage models with less than 3.

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*, pp. 24185–24198, 2024.

Lauren L. Cloutman. Interaction between dorsal and ventral processing streams: Where, when and how? *Brain and Language*, 127(2):251–263, 2013. ISSN 0093-934X. doi: https://doi.org/10.1016/j.bandl.2012.08.003. URL https://www.sciencedirect.com/science/article/pii/S0093934X12001459.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. *arXiv preprint arXiv:2503.17352*, 2025.

Jon Driver and Charles Spence. Crossmodal attention. *Current Opinion in Neurobiology*, 8(2):245–253, 1998. ISSN 0959-4388. doi: https://doi.org/10.1016/S0959-4388(98)80147-5. URL https://www.sciencedirect.com/science/article/pii/S0959438898801475.

Gerald M. Edelman and Joseph A. Gally. Reentry: a key mechanism for integration of brain function. *Frontiers in Integrative Neuroscience*, Volume 7 - 2013, 2013. ISSN 1662-5145. doi: 10.3389/fnint.2013.00063. URL https://www.frontiersin.org/journals/integrative-neuroscience/articles/10.3389/fnint.2013.00063.

Johan Eriksson, Edward K. Vogel, Anders Lansner, Fredrik Bergström, and Lars Nyberg. Neurocognitive architecture of working memory. *Neuron*, 88(1):33–46, 2015. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2015.09.020. URL https://www.sciencedirect.com/science/article/pii/S0896627315007771.

Javier Ferrando, Gerard I. Gállego, Ioannis Tsiamas, and Marta R. Costa-jussà. Explaining how transformers use context to build predictions. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5486–5513, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.301. URL https://aclanthology.org/2023.acl-long.301/.

Erez Freud, Tzvi Ganel, and Galia Avidan. *Functional Specialization Across the Visual Cortex*. Oxford University Press, December 2023. doi: 10.1093/acrefore/9780190264086.013.390.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.446. URL https://aclanthology.org/2021.emnlp-main.446/.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view, 2025. URL https://arxiv.org/abs/2410.10781.

Jinghan He, Kuan Zhu, Haiyun Guo, Junfeng Fang, Zhenglin Hua, Yuheng Jia, Ming Tang, Tat-Seng Chua, and Jinqiao Wang. Cracking the code of hallucination in lvlms with vision-aware head divergence, 2025. URL https://arxiv.org/abs/2412.13949.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching, 2024. URL https://arxiv.org/abs/2404.15255.

Ai Jian, Weijie Qiu, Xiaokun Wang, Peiyu Wang, Yunzhuo Hao, Jiangbo Pei, Yichen Wei, Yi Peng, and Xuchen Song. Csvqa: A chinese multimodal benchmark for evaluating stem reasoning capabilities of vlms, 2025a. URL https://arxiv.org/abs/2505.24120.

Pu Jian, Junhong Wu, Wei Sun, Chen Wang, Shuo Ren, and Jiajun Zhang. Look again, think slowly: Enhancing visual reflection in vision-language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9251–9270, Suzhou, China, November 2025b. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.470. URL https://aclanthology.org/2025.emnlp-main.470/.

Sai Srinivas Kancheti, Aditya Kanade, Rohit Sinha, Vineeth N Balasubramanian, and Tanuja Ganu. Faithful grpo: Improving visual spatial reasoning in multimodal language models via constrained policy optimization, 2026. URL https://arxiv.org/abs/2604.08476.

Shahar Katz and Yonatan Belinkov. VISIT: Visualizing and interpreting the semantic information flow of transformers. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14094–14113, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.939. URL https://aclanthology.org/2023.findings-emnlp.939/.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images, 2016.

Alan Lee and Harry Tong. Token-efficient rl for llm reasoning, 2025. URL https://arxiv.org/abs/2504.20834.

Yang Li, Zhichen Dong, Yuhan Sun, Weixun Wang, Shaopan Xiong, Yijia Luo, Jiashun Liu, Han Lu, Jiamang Wang, Wenbo Su, et al. Attention illuminates llm reasoning: The preplan-and-anchor rhythm enables fine-grained policy optimization. *arXiv preprint arXiv:2510.13554*, 2025.

Xingyu Lin, Yilin Wen, En Wang, Du Su, Wenbin Liu, Chenfu Bao, and Zhonghou Lv. Token-level policy optimization: Linking group-level rewards to token-level aggregation via markov likelihood, 2025a. URL https://arxiv.org/abs/2510.09369.

Zicheng Lin, Tian Liang, Jiahao Xu, Qiuzhi Lin, Xing Wang, Ruilin Luo, Chufan Shi, Siheng Li, Yujiu Yang, and Zhaopeng Tu. Critical tokens matter: Token-level contrastive estimation enhances llm's reasoning capability, 2025b. URL https://arxiv.org/abs/2411.19943.

Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models, 2025a. URL https://arxiv.org/abs/2505.21523.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. In *Conference on Language Modeling (COLM)*, 2025b.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

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 *International Conference on Learning Representations (ICLR)*, 2024.

Tianyi Men, Pengfei Cao, Zhuoran Jin, Yubo Chen, Kang Liu, and Jun Zhao. Unlocking the future: Exploring look-ahead planning mechanistic interpretability in large language models, 2024. URL https://arxiv.org/abs/2406.16033.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. 2025. URL https://api.semanticscholar.org/CorpusID:276929382.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022.

Earl Miller and Jonathan Cohen. An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24:167–202, 02 2001. doi: 10.1146/annurev.neuro.24.1.167.

Theofanis I. Panagiotaropoulos. An integrative view of the role of prefrontal cortex in consciousness. *Neuron*, 112(10):1626–1641, 2024. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2024.04.028. URL https://www.sciencedirect.com/science/article/pii/S0896627324002897.

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

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Representation Learning*, volume 2024, pp. 51575–51598, 2024.

Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. Multi-step reasoning with large language models, a survey, 2025. URL https://arxiv.org/abs/2407.11511.

Weikang Qiu, Tinglin Huang, Ryan Rullo, Yucheng Kuang, Ali Maatouk, S. Raquel Ramos, and Rex Ying. Rephqa: Evaluating readability of large language models in public health question answering, 2025. URL https://arxiv.org/abs/2509.16360.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300, 2024. URL https://api.semanticscholar.org/CorpusID:267412607.

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

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models, 2024. URL https://arxiv.org/abs/2406.17294.

Hongze Tan, Jianfei Pan, Jinghao Lin, Tao Chen, Zhihang Zheng, Zhihao Tang, and Haihua Yang. Gtpo and grpo-s: Token and sequence-level reward shaping with policy entropy, 2025a. URL https://arxiv.org/abs/2508.04349.

Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Minglan Lin, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning. *arXiv preprint arXiv:2503.20752*, 2025b.

Qwen Team. Qwen2.5-vl, January 2025a. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

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

Jean Vassoyan, Nathanaël Beau, and Roman Plaud. Ignore the kl penalty! boosting exploration on critical tokens to enhance rl fine-tuning, 2025. URL https://arxiv.org/abs/2502.06533.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. TRL: Transformers Reinforcement Learning, 2020. URL https://github.com/huggingface/trl.

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 *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=QWTCcxMpPA.

Peiyu Wang, Yichen Wei, Yi Peng, Xiaokun Wang, Weijie Qiu, Wei Shen, Tianyidan Xie, Jiangbo Pei, Jianhao Zhang, Yunzhuo Hao, Xuchen Song, Yang Liu, and Yahui Zhou. Skywork r1v2: Multimodal hybrid reinforcement learning for reasoning, 2025a. URL https://arxiv.org/abs/2504.16656.

Shenzhi Wang, Le Yu, Chang Gao, Chujie Zheng, Shixuan Liu, Rui Lu, Kai Dang, Xionghui Chen, Jianxin Yang, Zhenru Zhang, Yuqiong Liu, An Yang, Andrew Zhao, Yang Yue, Shiji Song, Bowen Yu, Gao Huang, and Junyang Lin. Beyond the 80/20 rule: High-entropy minority tokens drive effective reinforcement learning for llm reasoning, 2025b. URL https://arxiv.org/abs/2506.01939.

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. *arXiv preprint arXiv:2504.07934*, 2025c.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey, 2025d. URL https://arxiv.org/abs/2503.12605.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Mingshi Yan, Fei Huang, and Heng Ji. Perception-aware policy optimization for multimodal reasoning. *ArXiv*, abs/2507.06448, 2025e. URL https://api.semanticscholar.org/CorpusID:280150762.

Katsumi Watanabe and Shinsuke Shimojo. Chapter 89 - crossmodal attention in event perception. In Laurent Itti, Geraint Rees, and John K. Tsotsos (eds.), *Neurobiology of Attention*, pp. 538–543. Academic Press, Burlington, 2005. ISBN 978-0-12-375731-9. doi: https://doi.org/10.1016/B978-012375731-9/50093-8. URL https://www.sciencedirect.com/science/article/pii/B9780123757319500938.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, et al. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base llms. *arXiv preprint arXiv:2506.14245*, 2025.

Changyi Xiao, Mengdi Zhang, and Yixin Cao. Bnpo: Beta normalization policy optimization. *arXiv preprint arXiv:2506.02864*, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks, 2024. URL https://arxiv.org/abs/2309.17453.

Yunqiu Xu, Linchao Zhu, and Yi Yang. Mc-bench: A benchmark for multi-context visual grounding in the era of mllms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17675–17687, October 2025.

Huanjin Yao, Jiaxing Huang, Yawen Qiu, Michael K. Chen, Wenzheng Liu, Wei Zhang, Wenjie Zeng, Xikun Zhang, Jingyi Zhang, Yuxin Song, Wenhao Wu, and Dacheng Tao. Mmreason: An open-ended multi-modal multi-step reasoning benchmark for mllms toward agi, 2025. URL https://arxiv.org/abs/2506.23563.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024.

Di Zhang, Jianbo Wu, Jingdi Lei, Tong Che, Jiatong Li, Tong Xie, Xiaoshui Huang, Shufei Zhang, Marco Pavone, Yuqiang Li, et al. Llama-berry: Pairwise optimization for o1-like olympiad-level mathematical reasoning. *arXiv preprint arXiv:2410.02884*, 2024a.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. *ArXiv*, abs/2309.16042, 2023. URL https://api.semanticscholar.org/CorpusID:263131114.

Zhuosheng Zhang, Aston Zhang, Mu Li, hai zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024b. ISSN 2835-8856. URL https://openreview.net/forum?id=y1pPWFVfvR.

Haojie Zheng, Tianyang Xu, Hanchi Sun, Shu Pu, Ruoxi Chen, and Lichao Sun. Thinking before looking: Improving multimodal llm reasoning via mitigating visual hallucination, 2024. URL https://arxiv.org/abs/2411.12591.

Guanghao Zhou, Panjia Qiu, Cen Chen, Jie Wang, Zheming Yang, Jian Xu, and Minghui Qiu. Reinforced mllm: A survey on rl-based reasoning in multimodal large language models. *arXiv preprint arXiv:2504.21277*, 2025a.

Hengguang Zhou, Xirui Li, Ruochen Wang, Minhao Cheng, Tianyi Zhou, and Cho-Jui Hsieh. R1-zero's" aha moment" in visual reasoning on a 2b non-sft model, 2025. *URL https://arxiv.org/abs/2503.05132*, 2025b.

Jiayi Zhou, Jiaming Ji, Boyuan Chen, Jiapeng Sun, Wenqi Chen, Donghai Hong, Sirui Han, Yike Guo, and Yaodong Yang. Generative rlhf-v: Learning principles from multi-modal human preference, 2025c. URL https://arxiv.org/abs/2505.18531.

APPENDIX

TABLE OF CONTENTS

A General Discussions **15**

A.1 Sampling Analysis in Section 3 15

A.2 Statistical Analysis in Section 3 16

A.3 Limitations and Future Work 18

A.4 Broader Impacts 19

B Additional Implementation Details **19**

B.1 The Details of Conversation 19

B.2 The Details of Reward 20

B.3 The Algorithm Process 20

B.4 The Details of Training 20

C Supplementary Results **22**

C.1 Impact of Response Length on Reasoning 22

C.2 Additional Comparison between *V-ERR* with *Rhythm* 22

C.3 Visual Signal Attenuation Quantification 23

C.4 Validation of Task-Relevant Visual Focus 24

D Case Study **34**

A GENERAL DISCUSSIONS

A.1 SAMPLING ANALYSIS IN SECTION 3

To corroborate the universality of the "Signal Partitioning" phenomenon and the "Signal Attenuation" observed in our analysis, we analyze attention dynamics using two distinct vision-language model architectures: LLaVA-1.5-7B (Liu et al., 2023) and Qwen2.5-VL-3B-Instruct (Team, 2025a), on the MathVista dataset (Lu et al., 2024), with sampling temperature T = 1.0. For each model, we randomly selected 8 samples. Then we followed a rigorous "generate-then-extract" protocol: the model first autoregressively generated a full response, after which we performed a single forward pass on the concatenated sequence through the model to extract the causal attention matrices for analysis via the Hugging Face transformers library.

Text Focus Ratio. We first examined the temporal evolution of attention by calculating the Text Focus Ratio (TFR, as defined in Section 3.2) across the generated sequences. The comparative results are visualized in 3.

The analysis reveals distinct behavioral signatures for each architecture: (1) LLaVA-1.5-7B: This model exhibits a consistently high reliance on textual tokens throughout the generation process. The TFR fluctuates within a relatively narrow and high band, typically ranging between 0.85 and 0.95. This suggests that LLaVA may rely heavily on the language model's priors from the very onset of reasoning, with visual tokens exerting a steady but limited influence. (2) Qwen2.5-VL-3B-Instruct: In contrast, Qwen displays a dynamic "attenuation" trajectory. In the initial phase of generation, the TFR is marginally lower, which indicates a deeper initial engagement with visual signals. However, as the reasoning chain elongates, the TFR rises sharply, eventually reaching near-saturation levels.

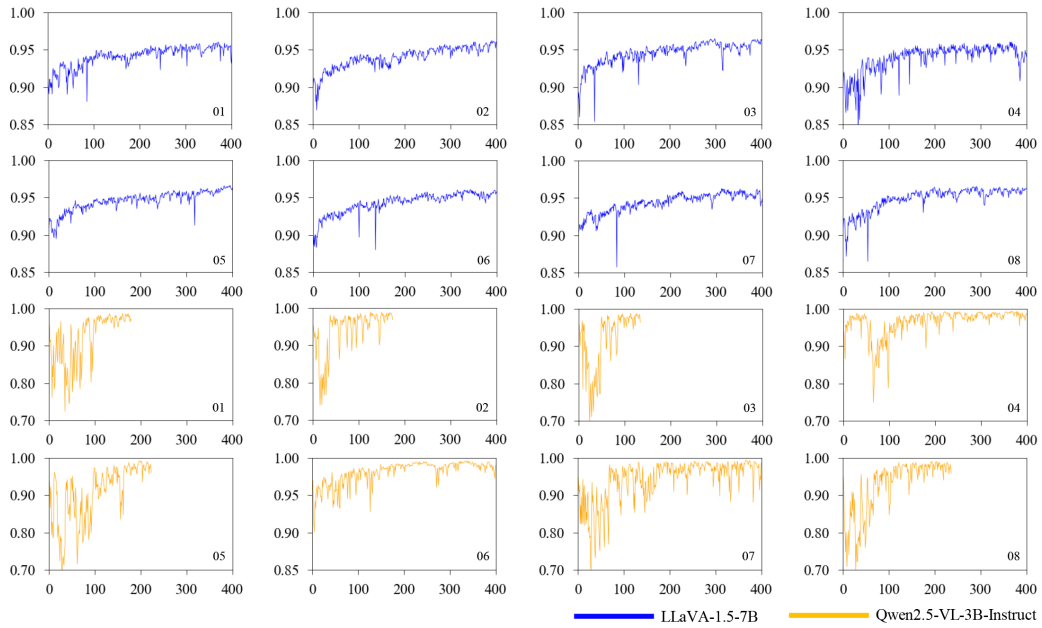


Figure 3: Comparative analysis of attention dynamics across different VLM architectures. The y-axis represents the text focus ratio, and the x-axis represents the response token index. The figure illustrates the temporal evolution of the Text Focus Ratio over the generation sequence for LLaVA-1.5-7B and Qwen2.5-VL-3B-Instruct, characterizing the progressive attenuation of visual information.

While Qwen exhibits stronger early-stage utilization of visual-modality signals, it also shows a more pronounced "signal attenuation" trend as reasoning progresses. In later stages of generation, the visual signal contribution drops sharply, and the model's inference becomes dominated by text-context signals, indicating that long-chain reasoning can drive visual signals to decay over time regardless of their initial strength.

Visual Propensity. We further investigated the structural location of the "Visual Pathways"—the specific attention heads responsible for processing image tokens. We computed the Visual Propensity (VP, as defined in Section 3.1) heatmaps for both models to identify where visual processing is physically localized. The comparative results are visualized in Figure 4.

The heatmaps reveal fundamentally different structural strategies: (1) LLaVA-1.5-7B: LLaVA's visual pathways are overwhelmingly concentrated in the early layers. The deep reasoning layers are largely "blind" to the image, relying entirely on the information propagated from these early layers. (2) Qwen2.5-VL-3B-Instruct: Qwen exhibits a relatively more uniform distribution. Its visual pathways appear with significant frequency in the middle and posterior layers.

This structural divergence highlights a critical insight: structural presence does not guarantee temporal persistence. Although Qwen possesses more visual pathways in deeper layers, the TFR analysis proves that it may not choose to utilize them as generation progresses. This validates our hypothesis that visual signal attenuation is a behavioral optimization issue rather than purely a structural capability bottleneck, further motivating the need for the recurrent reinforcement provided by V-ERR.

A.2 Statistical Analysis in Section 3

To further validate the generalizability of the "Signal Partitioning" and "Signal Attenuation" phenomena beyond case-level observations, we conducted a large-scale statistical analysis across broader data coverage. Specifically, we sampled 1,000 examples from CSVQA (Jian et al., 2025a), ScienceQA (Lu et al., 2022), MathVision (Wang et al., 2024), and MathVista (Lu et al., 2024), and evaluated three distinct vision-language models: LLaVA-1.5-7B (Liu et al., 2023), Qwen2.5-VL-3B-Instruct (Team, 2025a), and Qwen3-VL-2B-Instruct (Team, 2025b).

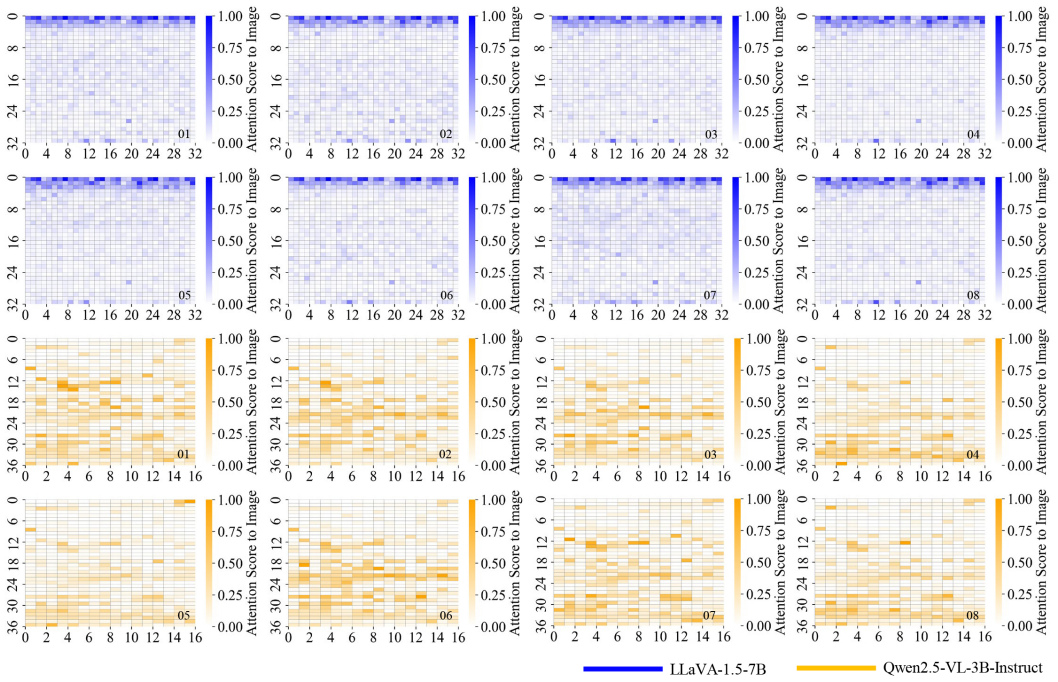


Figure 4: Comparison of Visual Propensity Score heatmaps across different VLM architectures. The heatmaps illustrate the layer-wise and head-wise distribution of visual heads in LLaVA-1.5-7B and Qwen2.5-VL-3B-Instruct, revealing distinct structural localization patterns of visual processing.

Table 5: Quantitative separation metrics between visual and non-visual heads for LLaVA-1.5 and Qwen2.5-VL.

	V-Heads Mean VP	Non-V-Heads Mean VP	VP Gap	VP Ratio	AUC	Cohen's d
LLaVA	0.219	0.050	0.169	4.378	0.953	2.201
Qwen	0.121	0.019	0.102	6.432	0.969	2.887

Visual Propensity. For each sample, we computed the VP matrix, and then stacked them to obtain the mean VP score, the standard deviation, and the variance for each head. In addition, to evaluate the stability of the visual heads, we also computed the selection frequency of each head being identified as a visual head across samples. For each architecture, we provide four heatmaps, as shown in Fig. 5. Furthmore, we computed explicit quantitative separation metrics comparing selected visual heads and non-visual heads. As shown in Table 5, for both llava and qwen, the selected visual heads have substantially higher VP than non-visual heads, with strong separation reflected by AUC and Cohen's d, showing a strong and statistically meaningful separation between visual and non-visual heads quantitatively.

Text Focus Ratio. To quantify how training objectives modulate the model's focus on textual versus visual tokens over the course of generation, we computed the TFR for each of the 1,000 sampled examples across the full autoregressive sequence. To enable aggregate comparison, we temporally aligned all TFR trajectories by resampling each curve to a fixed length of 100 uniform bins, then stacked the normalized curves to compute the mean trajectory and its standard deviation. We conducted this analysis under three conditions: (1) the base model, (2) fine-tuned with GRPO, and (3) fine-tuned with V-ERR.

Figure 6 presents the aggregated TFR curves for LLaVA and Qwen, with shaded regions indicating the 95% confidence interval (CI). Although TFR is informative about macro-level attention allocation, it does not capture where within the image the model focuses. V-ERR is not simply increasing

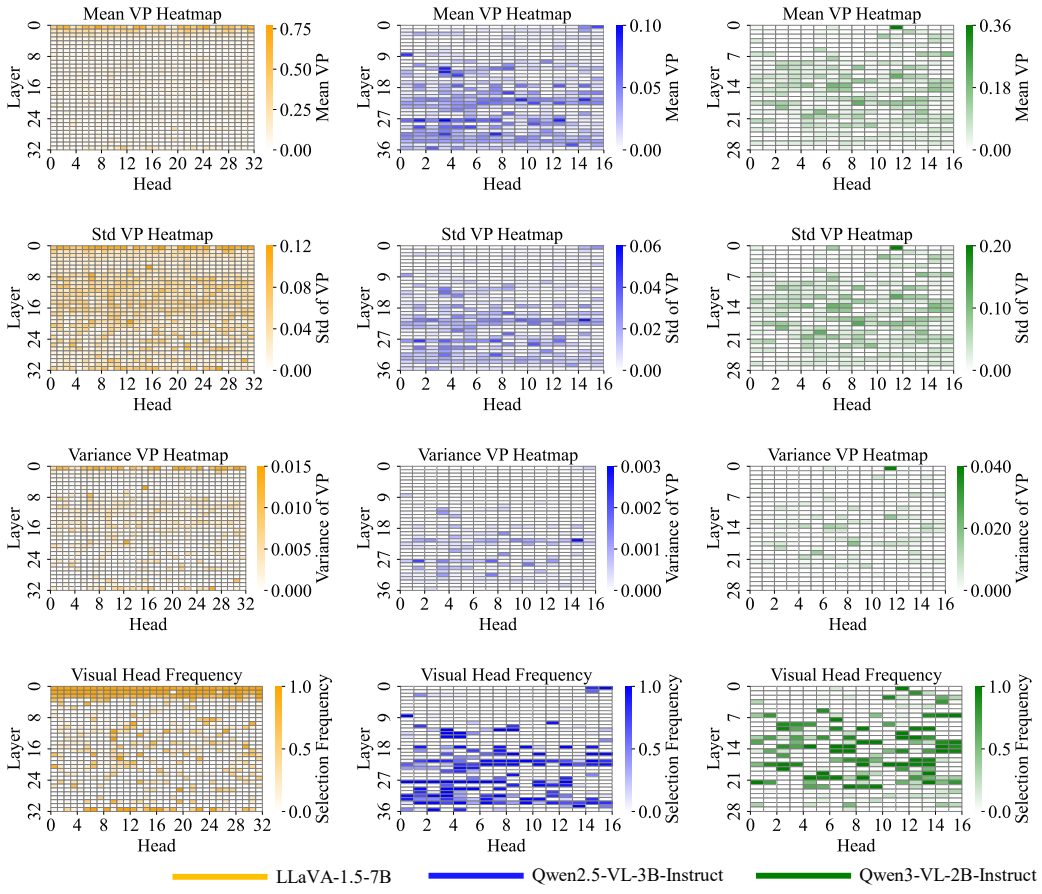


Figure 5: Comparison of statistical Visual Propensity heatmaps across different VLM architectures. The heatmaps illustrate the layer-wise and head-wise distribution of visual heads in LLaVA and Qwen, revealing distinct structural localization patterns of visual processing. High selection frequency reflects stable visual-head specialization across diverse inputs, reinforcing the reliability of the identified attention patterns.

the overall proportion of attention allocated to visual tokens, but rather encouraging the model to repeatedly attend to task-relevant regions within the image. To complement this global metric, the supplementary results in Section C.4 reveals the fine-grained attention of V-ERR-fine-tuned models, specifically pinpointing which image regions receive heightened focus.

A.3 Limitations and Future Work

While V-ERR consistently improves multimodal reasoning, it has several limitations. First, the effectiveness of our partitioning score and the resulting modality-specialized head set can be model- and architecture-dependent; different VLM backbones, training recipes, or attention implementations may yield different partitioning patterns, requiring re-identification and potentially reducing plug-and-play transfer. Second, our reward shaping encourages evidence-checking through specific pathways, but this proxy may not always correspond to correct verification: the model can over-attend to irrelevant visual regions or learn superficial "checking" behaviors that satisfy the reward without improving faithfulness, especially on tasks where visual grounding is subtle. Third, as a RL post-training method, V-ERR demands careful tuning and may be costly for large-scale adoption.

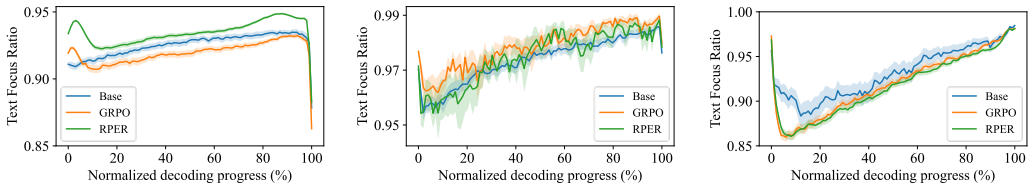


Figure 6: Aggregated TFR trajectories for LLaVA and Qwen under three training conditions: base pre-training, GRPO fine-tuning, and V-ERR fine-tuning. Solid lines denote mean TFR; shaded regions indicate 95% confidence intervals.

For future work, we plan to extend V-ERR beyond question-answering and math reasoning to more diverse, open-world settings where relevant evidence is distributed across many objects and long contexts. A natural direction is to couple pathway identification with region- or instance-level representations so that "re-checking" targets explicit, task-relevant entities rather than diffuse attention over image tokens. We also aim to develop stronger, less gameable verification rewards—e.g., rewards tied to consistent intermediate claims, cross-step visual entailment, or calibrated uncertainty—so that re-engagement reflects genuine validation. Finally, an exciting direction is model self-evolution, where the model iteratively improves its own pathway discovery and reward design, for example, periodically re-estimating modality-specialized heads as capabilities shift and proposing new evidence-checking behaviors from failure cases to continually enhance multimodal reasoning.

A.4 Broader Impacts

The V-ERR framework significantly advances the reliability of vision-language models in complex, multi-step reasoning tasks, offering broad benefits for education, scientific research, and technical analysis. By explicitly reinforcing visual evidence checking without requiring additional human annotations, V-ERR enables more accurate and interpretable multimodal reasoning, which can enhance intelligent tutoring systems, automated diagram and data interpretation, and AI-assisted scientific workflows.

However, enhancing multimodal reasoning also introduces important responsibility considerations. Stronger reasoning capabilities could be misused to generate highly convincing misinformation or automate high-stakes decisions if deployed without proper safeguards. Additionally, the model's reliance on specialized visual pathways may inadvertently amplify biases present in training data or over-rely on spurious visual correlations in novel scenarios.

B Additional Implementation Details

B.1 The Details of Conversation

We follow the native chat-based input interface of the underlying VLMs and construct each training or evaluation instance using the model's chat-template. Concretely, each sample is represented as a single-turn conversation, where the image and the text prompt are packed into the same user message as a structured content list:

```
conversation = [
    {
        "role": "user",
        "content": [
            {"type": "image", "image": image},
            {"type": "text", "text": prompt}
        ]
    }
]
```

This unified representation ensures that both modalities are aligned with the model's expected formatting and avoids discrepancies introduced by manual token concatenation. We then apply `processor.apply_chat_template` to convert the conversation into the final serialized input sequence used by the model.

Further, the text prompt is written in a structured and parseable format to reduce prompt-induced variance and enable automatic answer extraction. For all examples, we use the same prompt format for both RL post-training and evaluation to avoid train–test mismatch:

Input Prompt

System: Please reason step by step, and put your reasoning process inside `<think>` and `</think>` tags.
After reasoning, provide the final answer in the format: 'So the answer is X.'.
Question: ...

B.2 The Details of Reward

Accuracy Reward. The accuracy reward evaluates whether the response is correct. Following (Shao et al., 2024), we assess correctness solely based on the answer field: the reward is 1 if the answer is correct and 0 otherwise.

Format Reward. The format reward evaluates whether the model's response conforms to the required output structure. Specifically, the response must contain a reasoning section enclosed within `<think>...</think>` tags and conclude with a final answer in the exact format "So the answer is X.". The reward is computed as the sum of two binary sub-rewards: 0.5 if the `<think>` block is present, and 0.5 if the answer string matches the required pattern. The total format reward thus ranges from 0 to 1, with 1 indicating full compliance with the specified format.

Finally, we have the response reward R:

$$R = R_{\text{Acc}} + R_{\text{Fmt}}. \tag{10}$$

B.3 The Algorithm Process

Modern training frameworks often discard intermediate attention maps during the forward pass to save memory. So we implement an **Auxiliary Forward Pass** within the training loop based on the TRL framework that outputs attentions, inspired by (Li et al., 2025). The complete procedure of V-ERR is summarized in Algorithm 1.

B.4 The Details of Training

For every benchmark we post-train and evaluate separately: a model is trained with GRPO (or a GRPO variant, or V-ERR) on the training split of a given benchmark and then evaluated on the held-out test split of the same benchmark. No cross-benchmark data mixing is used, and the training and test splits of every benchmark are disjoint. All questions are cast as multiple-choice; the exact sample lists (question, choices, answer, image path) for both splits are released in `datasets/<benchmark>/{train,test}.json` of our code repository.

All baselines are implemented in the same code base and trained under identical conditions: the same backbone checkpoint, the same training split, the same prompt, the same reward functions, the same optimizer settings and number of optimizer steps, the same rollout budget (G=4, temperature 1.0), and the same evaluation protocol. Concretely:

- **GRPO / DAPO / Dr. GRPO / BNPO** share one training script and differ only in `loss_type` and the corresponding advantage/length normalization of TRL's `GRPOTrainer`.
- **Rhythm** is re-implemented for VLMs following the released description, with the hyper-parameters of the original paper: local/global head ratio $0.3/0.3$, quantile q=0.4 for both the preplan (WAAD-difference) and anchor (FAI) triggers, look-back k=3, α=0.5, and the same

Algorithm 1 V-ERR

Generation and Evaluation:
Generate responses $\mathcal{O} = \{o_1, \ldots, o_G\}$ from policy π_θ
Compute rewards and estimate advantages $\hat{A}_i$

for each response $o_i \in \mathcal{O}$ **do**
    Run an auxiliary forward pass to obtain $\{A^{(l,h)}\}$
    Visual Evidence Localization:
    Identify visual-pathway set $\mathcal{H}_{\text{vis}}$
    Aggregate attention over $\mathcal{H}_{\text{vis}}$ by Eq. equation 3
    Compute the task-relevant score for each image token by Eq. equation 4
    Select visual evidence $\mathcal{V}_{\text{evi}}$ by Eq. equation 5

    Evidence Re-checking Reinforcement:
    for $t = 1$ to $|o_i|$ **do**
        Focus visual evidence strength by Eq. equation 6
        Compute the local Z-score over window W by Eq. equation 7
        Set token weight by Eq. equation 9
    end for
    Normalize $\{\gamma_t\}_{t=1}^{|o_i|}$ such that $\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\gamma_t = 1$
end for

Policy Optimization:
Scale per-token advantages by Eq. equation 8
Update parameters θ with $\tilde{A}_{i,t}$

amplification γ_{amp}=2 as V-ERR, using the "coupled" strategy. It uses the same auxiliary forward pass and the same $\mathbb{E}_t[\gamma_t]$=1 normalization as V-ERR, so the two methods differ *only* in which tokens receive amplified credit (textual preplan/anchor tokens vs. visual-evidence re-checking tokens).

- **Focus** uses the V-ERR pipeline with the task-relevance mask removed, i.e., a_t is computed over all image tokens instead of V_{evi}.

All methods are implemented on top of the `GRPOTrainer` of TRL, by overriding `_generate_and_score_completions` to attach a per-token weight γ_t and `_compute_loss` to apply it. Every run uses a single NVIDIA H100 80GB in bfloat16 with TF32 matrix multiplication enabled. Because the auxiliary forward pass must return attention maps, the policy is loaded with `attn_implementation="eager"` for the attention-based methods (Rhythm and V-ERR).

We train with four sampled rollouts per prompt, using a per-device batch size of 4 and gradient accumulation over two steps, which yields an effective batch size of 8 per device. During training, completions are sampled with a decoding temperature of 1.0. We use a model-dependent learning rate of 5×10^{-6} for LLaVA and 7×10^{-6} for Qwen models. Unless noted, we set the top-k ratio for visual heads to 0.2, the top-k ratio for visual anchors to 0.15, the recurrent window size to 10, the max completion length to 512. All runs are conducted on a single GPU with bfloat16 precision and TF32 enabled to accelerate matrix multiplications. All remaining arguments of `GRPOConfig` keep their TRL defaults, the relevant ones of which are reproduced in the lower block of the table for completeness. The same optimization settings are shared by GRPO, DAPO, Dr. GRPO, BNPO, Rhythm and V-ERR on a given backbone; the variants differ only in the loss aggregation (`loss_type`) or in the token-level weighting.

Moreover, Table 6 presents a detailed comparison of the computational overhead between GRPO and V-ERR. Although V-ERR requires explicit attention matrix extraction, this design is not inherently incompatible with modern accelerated training pipelines such as FlashAttention. The attention preservation requirement applies exclusively to the auxiliary forward pass, while the rollout generation stage, which is typically the major runtime bottleneck in online RL training, can still benefit from existing acceleration techniques. In practice, we integrated vLLM to accelerate the rollout phase,

Table 6: Computational Cost Comparison between GRPO and V-ERR.

Setting	Metric	GRPO	V-ERR	Ratio
samples=600	wall-clock GPU hours	3708 seconds	5054 seconds	1.36
max completion length=512	Throughput (samples/sec)	0.193	0.135	0.70
BF16, TF32=True, True	Memory Usage	56354MB	69878MB	1.24

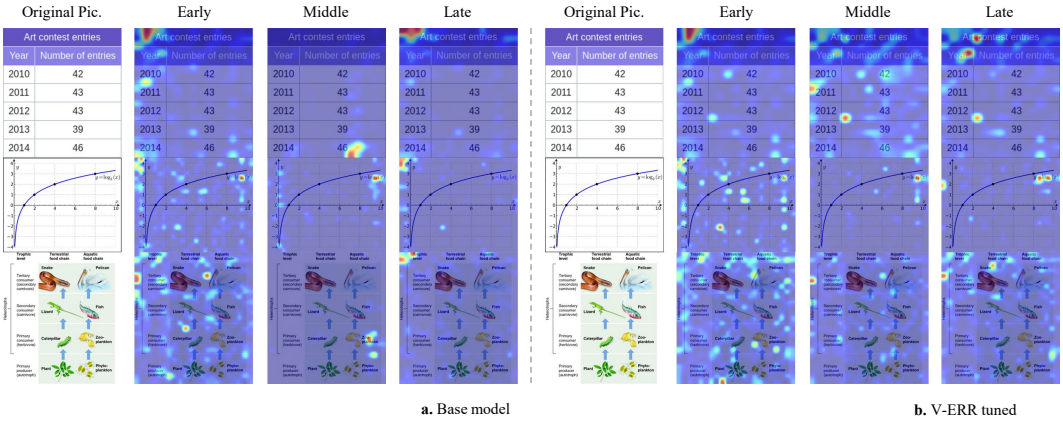


Figure 7: Effect of enforcing visual evidence rechecking. For each example, we overlay attention heatmaps on the input image at three stages of generation (Early / Middle / Late), comparing the base model (a) and the V-ERR-tuned model (b). Warmer colors indicate higher attention to image tokens.

Table 7: Results with 2K Completion Length.

Method	MathVista	CSVQA	MathVision	ScienceQA	MMMU	Avg.
Qwen2.5-VL-3B-Instruct						
GRPO	67.0	34.8	24.4	80.9	43.1	50.0
DAPO	63.3	31.9	21.7	80.0	42.8	47.9
Dr. GRPO	64.0	33.8	25.6	81.1	42.4	49.4
BNPO	66.7	34.1	23.5	80.9	43.8	49.8
V-ERR	**69.3** +2.3	**37.9** +3.1	**27.4** +3.0	**81.3** +0.4	**45.2** +2.1	**52.2** +2.2

which can increase the generation throughput to 0.168 samples/sec. This demonstrates that V-ERR can be incorporated into contemporary training workflows without sacrificing most of the speed benefits of established inference acceleration frameworks.

C Supplementary Results

C.1 Impact of Response Length on Reasoning

Our initial experiments employed a maximum completion length of 512 tokens. However, given that mathematical reasoning often necessitates extended chains of thought, we further investigated the potential impact of longer generations on model performance. We observed that response lengths of Qwen could reach up to approximately 1,900 tokens. Consequently, we adjusted *max_completion_length* to 2048 for Qwen, with the corresponding results shown in Table 7.

C.2 Additional Comparison between *V-ERR* with *Rhythm*

We further conduct supplementary experiments on Qwen3-VL-2B-Instruct (Team, 2025b) to compare our *V-ERR* with *Rhythm* (Li et al., 2025), a recent RL paradigm proposed mainly for language models.

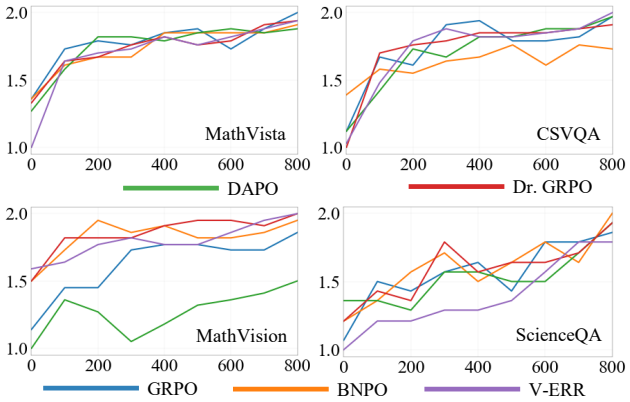


Figure 8: Training reward curves of V-ERR versus baselines. The x-axis denotes training steps, and the y-axis denotes the reward normalized to the range [1,2].

Table 8: Comparison between *V-ERR* and *Rhythm*.

Method	AI2D	CSVQA	ScienceQA	MathVista	MathVision	MathV360K	Avg.
			Qwen3-VL-2B-Instruct				
Rhythm	75.8	33.9	83.8	67.8	24.7	**70.6**	59.4
V-ERR	**76.3** +0.5	**38.5** +4.6	**87.2** +3.4	**68.9** +1.1	**27.7** +3.0	69.5 -1.1	**61.4** +2.0

The key motivation is to examine whether a planning–anchoring rhythm, originally designed to structure text-only reasoning trajectories, can effectively transfer to vision–language reasoning.

Concretely, we evaluate both methods on six multimodal benchmarks, including AI2D (Kembhavi et al., 2016), CSVQA (Jian et al., 2025a), ScienceQA (Lu et al., 2022), MathVista (Lu et al., 2024), MathVision (Wang et al., 2024), and MathV360K (Shi et al., 2024). The results, as shown in Table 8, indicate that directly reinforcing recurrent visual-evidence-checking through visual pathways is more effective than imposing a primarily language-structured rhythm on multimodal reasoning.

These findings suggest that, in VLMs, reliability bottlenecks are not only caused by poor textual planning, but also by "visual signal attenuation" and its weak re-engagement in later steps. While "Preplan-and-Anchor Rhythm" can improve the coherence of intermediate text structure, it does not explicitly prevent the model from drifting away from visual evidence as reasoning progresses. In contrast, V-ERR targets the internal mechanism that mediates text-to-vision information flow and explicitly strengthens repeated visual grounding, leading to more robust multi-step reasoning improvements across tasks.

C.3 Visual Signal Attenuation Quantification

To quantify the extent of visual signal attenuation along the generation trajectory, we introduce a **Visual Signal Attenuation Ratio (VSFR)** based on the **Text Focus Ratio (TFR)**. For each dataset, we randomly sample 10 instances and compute the pre-token TFR_t over the generated response. We then partition each response of length (T) into an early segment and a late segment, taking the first 30% tokens $(E = 1, \ldots, \lfloor 0.3T \rfloor)$ and the last 30% tokens $(L = \lceil 0.7T \rceil, \ldots, T)$. Denoting the average TFR in the early and late segments as

$$e = \frac{1}{|E|} \sum_{t \in E} TFR_t, \qquad l = \frac{1}{|L|} \sum_{t \in L} TFR_t, \tag{11}$$

we define the attenuation ratio as

$$\text{VSFR} = \frac{(1-e) - (1-l)}{1-e} = \frac{l-e}{1-e}. \tag{12}$$

Table 9: Visual Signal Attenuation Ratio (VSFR, %).

Model	01	02	03	04	05	06	07	08
Qwen2.5-VL-3B-Instruct on MathVista								
Base Model	75.8	79.2	78.5	79.2	73.7	79.4	72.2	77.6
V-ERR-Tuned	**73.3** -2.5	**74.2** -5.0	**71.2** -7.3	**72.5** -6.7	**64.9** -8.8	**65.7** -13.7	**63.0** -9.2	**71.1** -6.5
Qwen2.5-VL-3B-Instruct on MathVision								
Base Model	77.5	81.5	75.9	72.8	81.1	79.4	75.6	74.4
V-ERR-Tuned	**72.5** -5.0	**74.2** -7.3	**70.9** -5.0	**68.8** -4.0	**78.8** -2.3	**74.2** -5.2	**70.7** -4.9	**67.6** -6.8
Qwen2.5-VL-3B-Instruct on CSVQA								
Base Model	77.1	**75.4**	78.7	82.4	72.7	73.1	74.9	76.8
V-ERR-Tuned	**71.3** -5.8	78.1 +2.7	**70.7** -7.0	**71.9** -10.5	**65.2** -7.5	**68.3** -4.8	**73.0** -1.9	**73.1** -3.7

Since a higher TFR indicates a stronger shift of attention mass toward text and correspondingly weaker utilization of image tokens, a larger gap between early and late stages reflects more severe attenuation. Therefore, lower VSFR is better, indicating that the model maintains more consistent multimodal attention allocation over time with less late-stage drift toward text. The results are summarized in Table 9. Across datasets, the V-ERR-tuned model exhibits substantially lower VSFR than the base model, demonstrating that our training can mitigate visual signal attenuation to a certain extent on Qwen.

C.4 Validation of Task-Relevant Visual Focus

To further substantiate that the visual tokens recurrently attended to under V-ERR correspond to task-relevant evidence, we conducted a comparative visualization against the GRPO-tuned baseline. Specifically, for each sampled instance, we extracted the top-10 visual tokens receiving the highest cumulative attention during generation and mapped them back to their corresponding spatial patches in the original image. The identified regions were then highlighted with red bounding boxes. Representative cases are presented below.

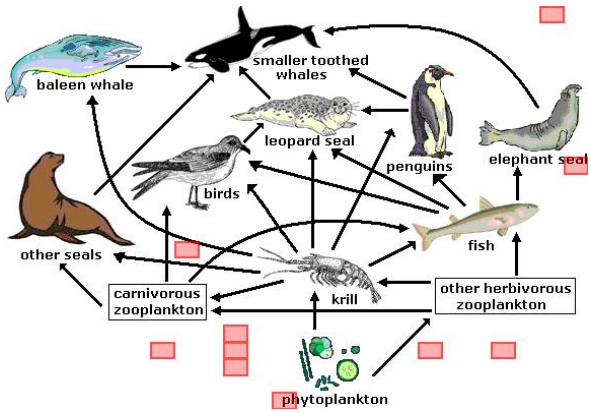
Top 10 Visual Tokens: Case 1

Question:
What could happen that would increase the number of krill?

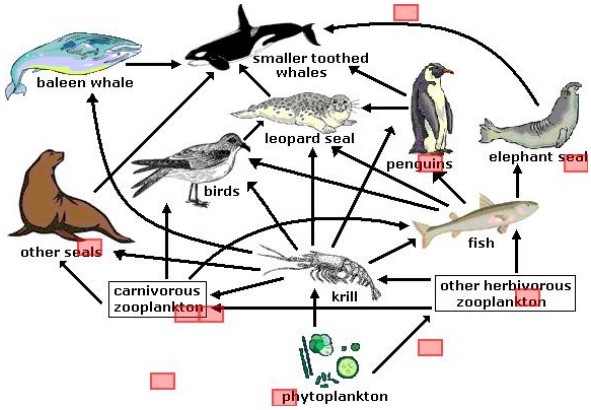
Options:
A. increase in phytoplankton
B. decrease in penguins
C. increase in fish
D. increase in birds

Answer: A

GRPO Focused Tokens:



V-ERR Focused Tokens:



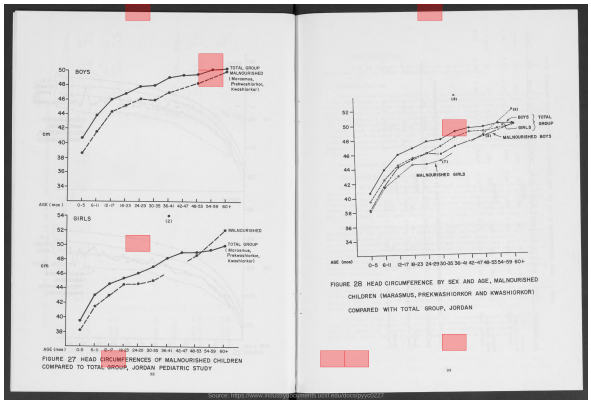
Top 10 Visual Tokens: Case 2

Question:
What is the minimum age group shown in the ‘plots’?

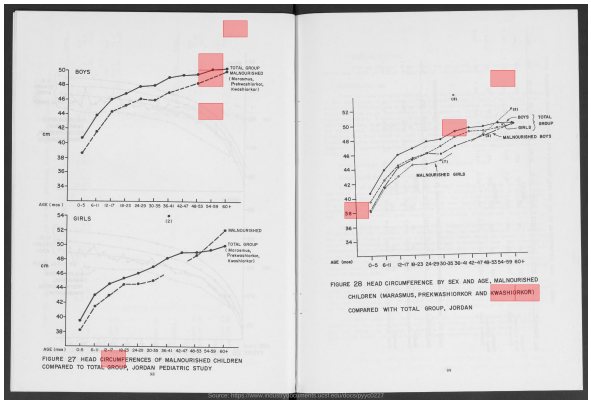
Options:
A. 11-15
B. 21-25
C. 6-10 D. 16-20
E. 0-5

Answer: E

GRPO Focused Tokens:



V-ERR Focused Tokens:



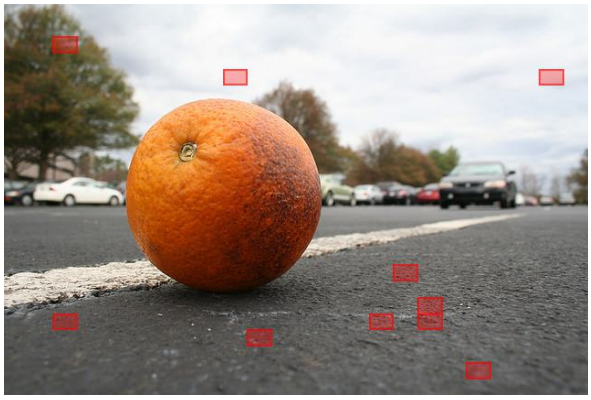
Top 10 Visual Tokens: Case 3

Question:
Is the orange larger than the car?

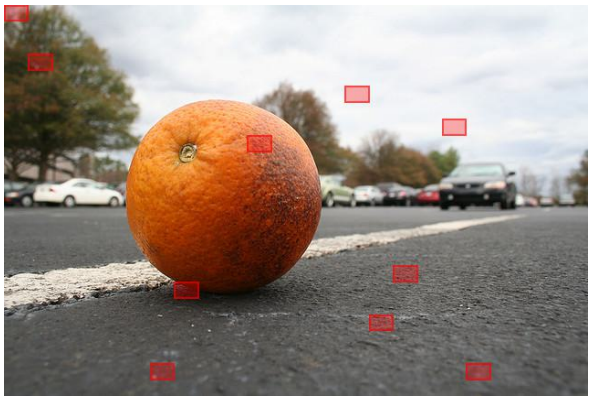
Options:
A. Yes
B. No

Answer: B

GRPO Focused Tokens:



V-ERR Focused Tokens:



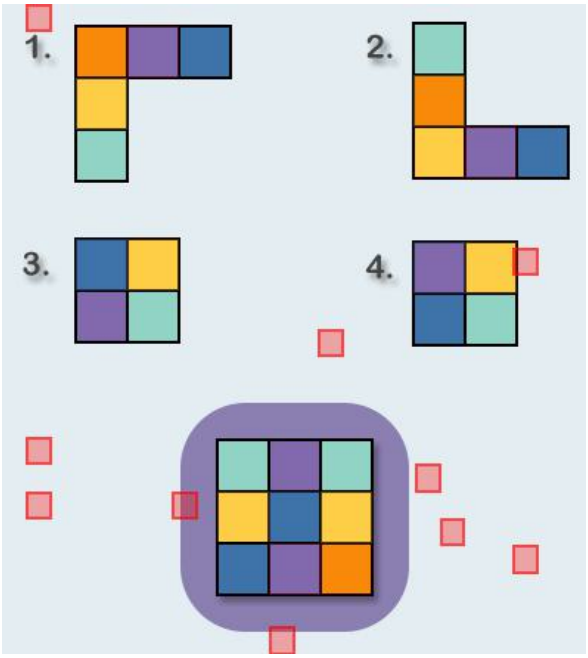
Top 10 Visual Tokens: Case 4

Question:
Which two puzzle pieces form the larger square?

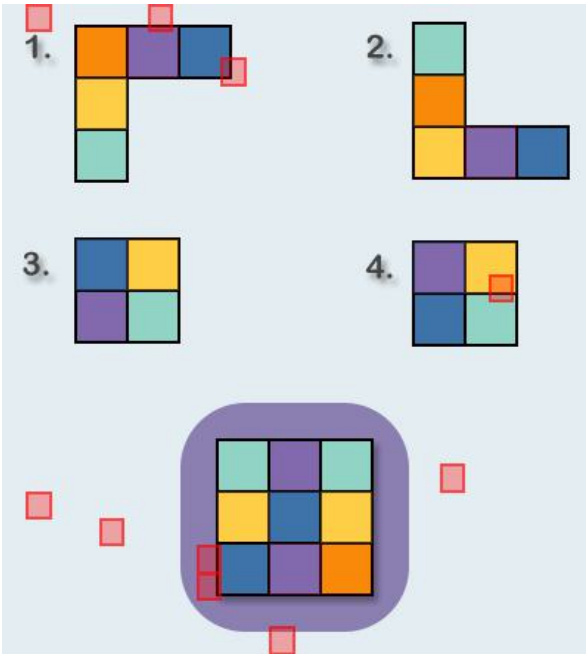
Options:
A. 1 & 2
B. 1 & 3
C. 1 & 4
D. 2 & 3
E. 2 & 4

Answer: B

GRPO Focused Tokens:



V-ERR Focused Tokens:



Top 10 Visual Tokens: Case 5

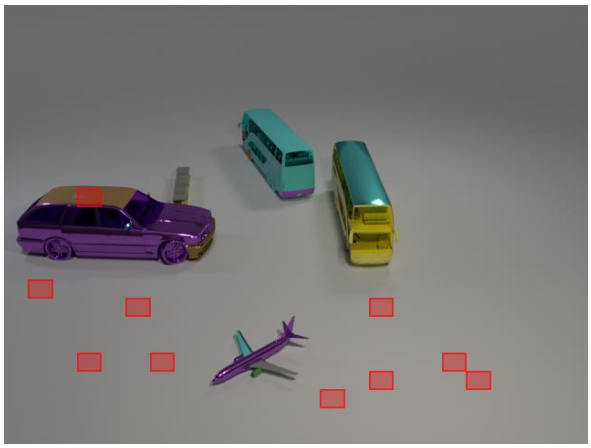
Question:
Are there fewer double-decker buses that are behind the aeroplane than things on the left side of the yellow double bus?

Options:
A. Yes

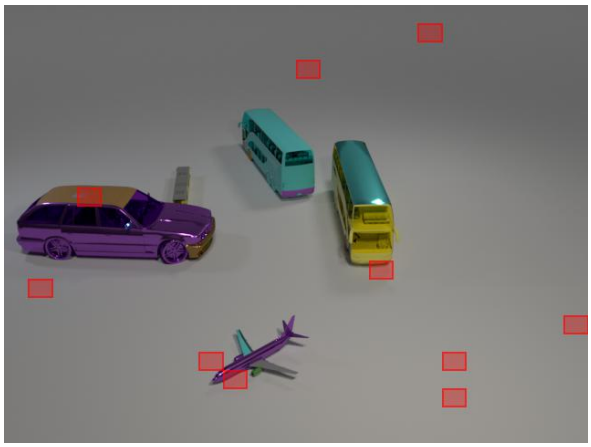
B. No

Answer: A

GRPO Focused Tokens:



V-ERR Focused Tokens:



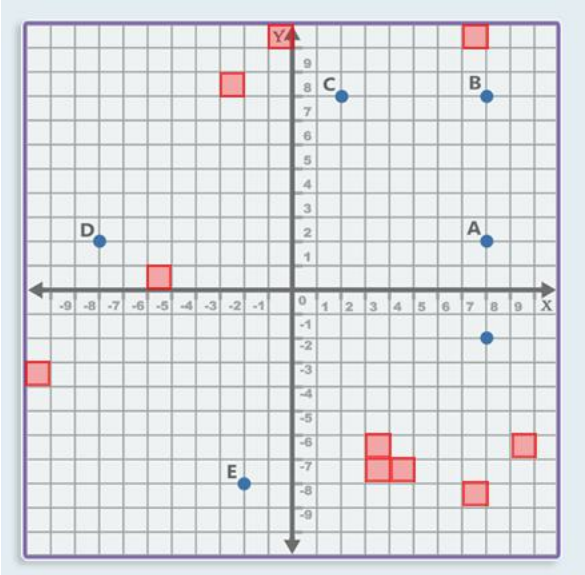
Top 10 Visual Tokens: Case 6

Question:
Find the image of the dot (8,-2) under a clockwise rotation by 270° about the origin.

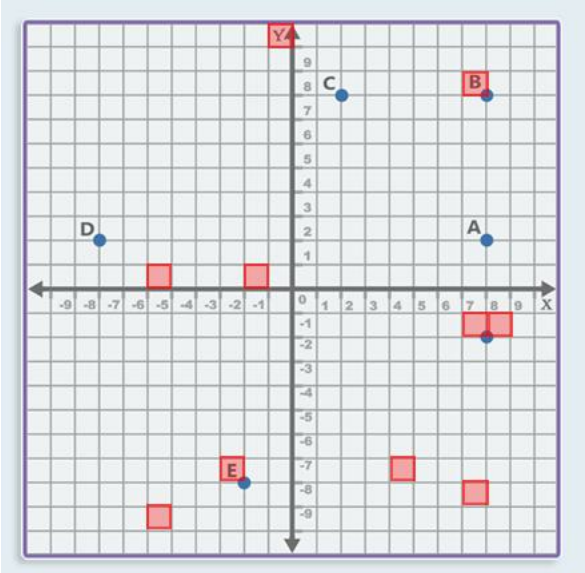
Options:
A. A
B. B
C. C
D. D
E. E

Answer: C

GRPO Focused Tokens:



V-ERR Focused Tokens:



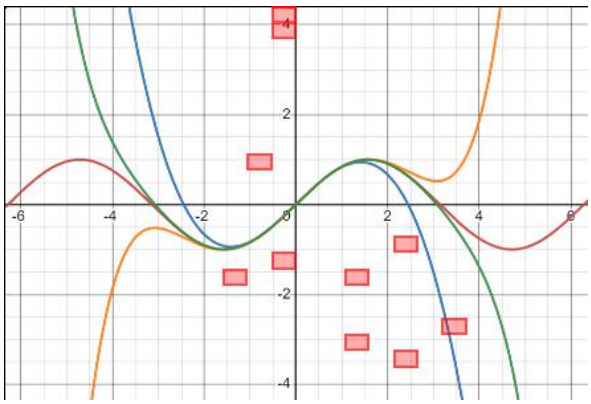
Top 10 Visual Tokens: Case 7

Question:
How many odd functions are in the graph?

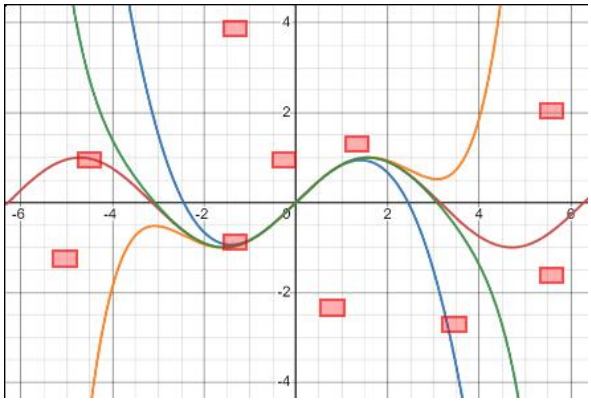
Options:
A. 3
B. 4
C. 5
D. 7

Answer: B

GRPO Focused Tokens:



V-ERR Focused Tokens:



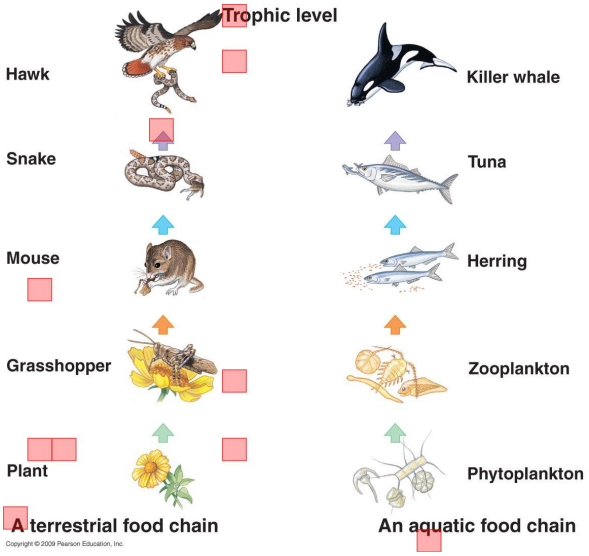
Top 10 Visual Tokens: Case 8

Question:
Which organism would be most affected if there was a shortage of plants?

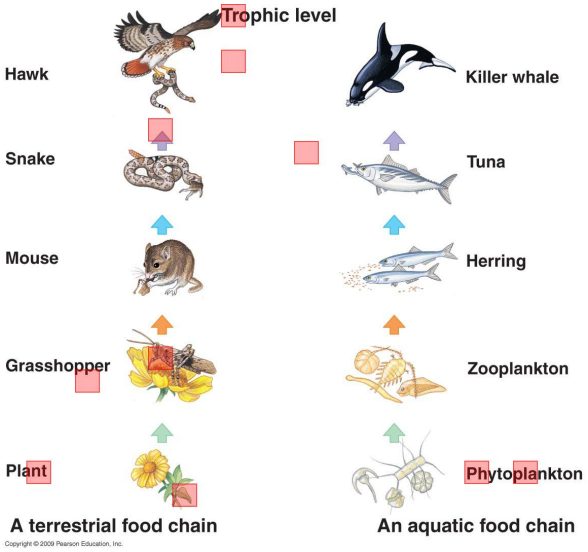
Options:
A. Grasshopper
B. Snake
C. Mouse
D. Hawk

Answer: A

GRPO Focused Tokens:



V-ERR Focused Tokens:



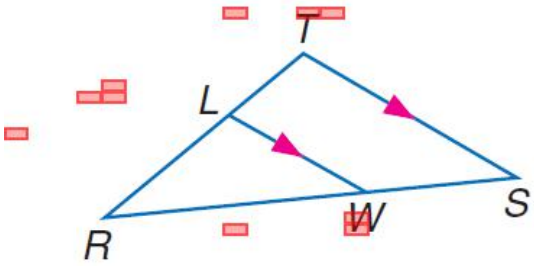
Top 10 Visual Tokens: Case 9

Question:
If RL = 5, RT = 9, and WS = 6, find RW.

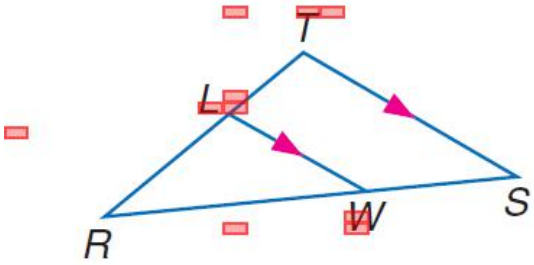
Options:
A. 5.4
B. 6
C. 6.6
D. 7.5

Answer: D

GRPO Focused Tokens:



V-ERR Focused Tokens:



Top 10 Visual Tokens: Case 10

Question:
What is the age gap between these two people in image?

Options:
A. 3
B. 1
C. 2
D. 4

Answer: B

GRPO Focused Tokens:



V-ERR Focused Tokens:



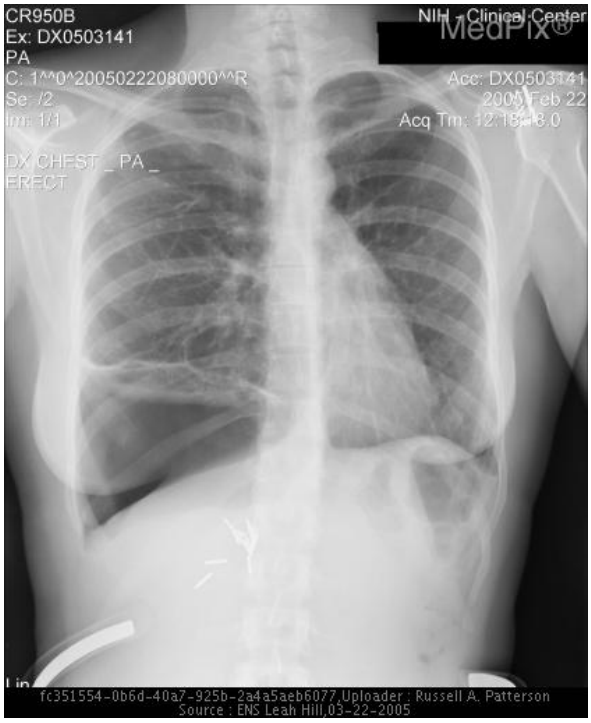
Qualitative Analysis. As illustrated in the provided examples, direct visual inspection reveals consistent behavioral distinctions. V-ERR can highlight regions that contain critical visual evidence required for problem-solving, whereas the GRPO-tuned model often attends to visually salient but task-irrelevant areas or distributes attention more uniformly. This qualitative pattern aligns with our hypothesis.

Quantitative Validation. Given the inherent subjectivity and scalability limits of human evaluation, we further employed an LLM-as-a-judge protocol to assess region-level task relevance. We prompted GPT-5.4 to perform blind pairwise comparisons, determining which of the two highlighted regions was more directly relevant to answering the given question. Across the evaluated sample set, GPT-5.4 preferred the regions identified by V-ERR in **80%** of cases. This quantitative result corroborates our qualitative observations and confirms that V-ERR's recurrent attention mechanism consistently directs the model toward task-critical visual evidence.

D CASE STUDY

As emphasized throughout our analysis, V-ERR is fundamentally designed to enhance multimodal reasoning through more effective evidence utilization, rather than treating attention optimization as an isolated objective. While attention metrics and TFR curves reveal how V-ERR alters processing dynamics, they do not directly capture improvements in problem-solving capability. To empirically demonstrate that this mechanism translates into tangible reasoning gains, we present a set of qualitative case studies. These examples illustrate how V-ERR's evidence-rechecking behavior enables the model to maintain alignment between visual cues and textual logic to improve reasoning.

Case 1



Question:
Is the heart wider than more than half the width of the thorax?

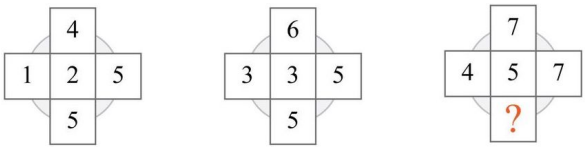
Options:
A. Yes
B. No

Answer: B

Original Response:
<think>
To determine if the heart is wider than more than half the width of the thorax, we need to analyze the image provided. The thoracic cavity is the area enclosed by the ribcage. If the heart's width is greater than half the width of the thoracic cavity, then the heart is wider than more than half the width of the thorax. In this case, the heart appears to be wider than half the width of the thoracic cavity based on the image provided.
</think>
So the answer is A.

New Response:
<think>
To determine if the heart is wider than more than half the width of the thorax, we need to compare the size of the heart with the size of the thorax. In the image, the heart appears as a dark area located centrally within the thoracic cavity. We can measure the distance between the outer edges of the ribs on either side of the spine.
</think>
So the answer is B.

Case 2



Question:
Which number is missing?

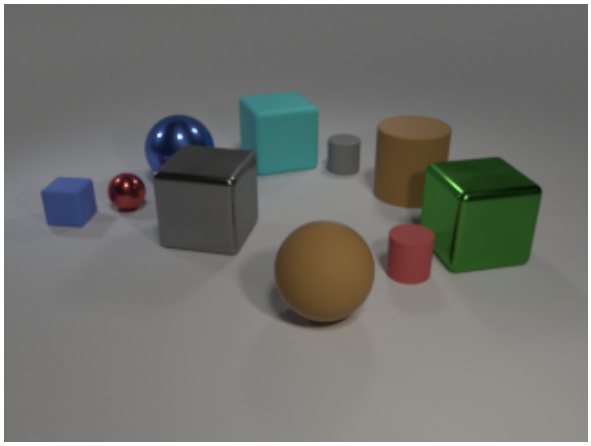
Options:
A. 7
B. 8
C. 10
D. 6

Answer: A

Original Response:
<think>
To determine the missing number, let's analyze the pattern in the given numbers:
1. The first row has the numbers 1, 2, and 5.
2. The second row has the numbers 3, 3, and 5.
3. The third row has the numbers 4, 5, and 7.
We can observe that each row contains three numbers, and the middle number is always 5. This suggests that the missing number should be placed in the same position as the other numbers in the rows.
Therefore, the missing number should be 6, as it fits the pattern of being the middle number in each row.
</think>
So the answer is D.

New Response:
<think>
To determine the missing number, let's analyze the pattern in the given numbers:
1. In the first square, the numbers are arranged in a clockwise direction starting from the top-left corner.
2. The numbers are 4, 1, 2, and 5.
3. The second square has the same arrangement but with different numbers: 6, 3, 3, and 5.
Now, let's look at the third square:
- The numbers are 4, 5, and 7.
- We need to find the missing number that fits the pattern of the other squares.
The pattern seems to be that each square contains four numbers arranged in a clockwise direction, and the numbers increase as we move clockwise.
Let's check the options:
A. 7 - This would make the sequence complete if it were placed in the correct position.
B. 8 - This does not fit the pattern of increasing numbers.
C. 10 - This is too high for the pattern.
D. 6 - This does not fit the pattern of increasing numbers.
Therefore, the missing number should be 7 to maintain the pattern.
</think>
So the answer is A.

Case 3



Question:
Subtract all large purple spheres. Subtract all small gray things. How many objects are left?

Options:
A. 7
B. 9
C. 6
D. 11

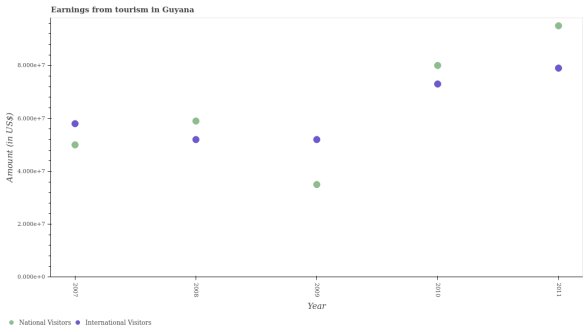
Answer: B

Original Response:
<think>
To determine how many objects are left after subtracting the specified items, let's go through each object one by one:
1. Blue sphere: This is not a purple sphere, so it remains.
2. Red sphere: This is not a purple sphere, so it remains.
3. Gray cube: This is not a purple sphere, so it remains.
4. Green cube: This is not a purple sphere, so it remains.
5. Blue cube: This is not a purple sphere, so it remains.
6. Large blue sphere: This is not a purple sphere, so it remains.
7. Large cyan cube: This is not a purple sphere, so it remains.
8. Small gray cylinder: This is not a purple sphere, so it remains.
9. Large brown cylinder: This is not a purple sphere, so it remains.
10. Red cylinder: This is not a purple sphere, so it remains.
After removing the specified items (purple spheres), we are left with all other objects. Therefore, there are 10 objects remaining.
</think>
So the answer is A.

New Response:
<think>
To determine how many objects are left after subtracting the specified items, I'll go through each object one by one:
1. The large blue sphere is not affected.
2. The red metal ball is not affected.
3. The large gray cube is not affected.
4. The cyan cube is not affected.
5. The small gray cylinder is removed.
6. The large brown cylinder is not affected.

7. The large gray cube is not affected.
8. The small red cylinder is not affected.
9. The small blue cube is not affected.
10. The large green cube is not affected.
After removing the specified items, there are 9 objects remaining.
</think>
So the answer is B.

Case 4



Question:
In how many years, is the amount earned from national visitors greater than the average amount earned from national visitors taken over all years?

Options:
A. 3
B. 2
C. 5
D. 4

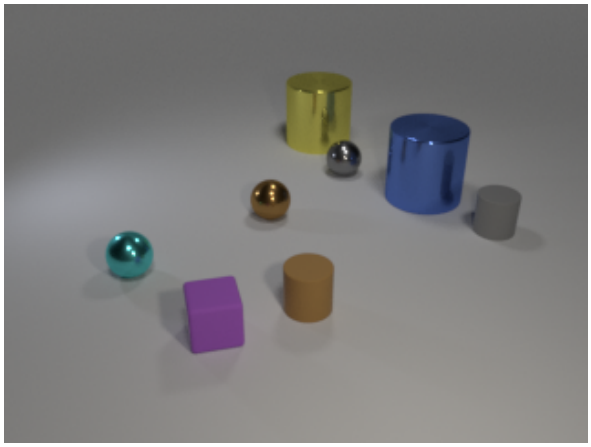
Answer: B

Original Response:
<think>
To determine how many years have earnings greater than the average, we need to calculate the average earnings from national visitors over the given years.
Average earnings = (Total earnings from national visitors) / (Number of years)
Average earnings = (5000000 + 5800000 + 3500000 + 7900000 + 9300000) / 5
Average earnings = 31500000 / 5
Average earnings = 6300000
There are 3 years where the earnings are greater than the average.
</think>
So the answer is A.

New Response:
<think>
To determine how many years have earnings greater than the average, we need to calculate the average earnings from national visitors over all years.
Average earnings = (Total earnings from national visitors) / (Number of years)
Average earnings = (5000000 + 5800000 + 3500000 + 7900000 + 9300000) / 5
Average earnings = 31500000 / 5
Average earnings = 6300000
Now, we compare each year's earnings with this average:
2007: 5000000 < 6300000 (False)
2008: 5800000 < 6300000 (False)

2009: 3500000 < 6300000 (False)
2010: 7900000 > 6300000 (True)
2011: 9300000 > 6300000 (True)
</think>
So the answer is B.

Case 5



Question:
Subtract all brown matte objects. Subtract all blue metallic objects. How many objects are left?

Options:
A. 3
B. 6
C. 8
D. 7

Answer: B

Original Response:
<think>
To determine how many objects are left after subtracting the specified items, let's list them out:
1. Yellow cylinder
2. Blue cylinder
3. Gray sphere
4. Brown sphere
5. Purple cube
6. Brown matte cylinder
7. Gray matte cylinder
Now, we need to remove the brown matte cylinder and the blue cylinder from this list.
<remove>
- Brown matte cylinder
- Blue cylinder
</remove>
This leaves us with:
1. Yellow cylinder
2. Gray sphere
3. Brown sphere
4. Purple cube

5. Gray matte cylinder
Counting these remaining objects, we have 5 objects left.
</think>
So the answer is A.

New Response:
<think>
To determine how many objects are left after subtracting the specified items, let's go through each object one by one:
1. The yellow cylinder is not affected.
2. The blue cylinder is not affected.
3. The gray sphere is not affected.
4. The brown sphere is not affected.
5. The cyan sphere is not affected.
6. The purple cube is not affected.
7. The brown matte cylinder is removed.
8. The gray matte cylinder is removed.
After removing the brown matte cylinder and the gray matte cylinder, we are left with the following objects: the yellow cylinder, the blue cylinder, the gray sphere, the brown sphere, the cyan sphere, and the purple cube.
Therefore, there are 6 objects remaining.
</think>
So the answer is B.