

Reflect to Inform: Boosting Multimodal Reasoning via Information-Gain-Driven Verification

Shuai Lv^{1,†}, Chang Liu^{1,†}, Feng Tang^{2,*}, Rui Chen¹, Yujie Yuan²
AoJun Zhou³, Kui Zhang², Xi Yang^{4,*}, Yangqiu Song⁴

¹University of Science and Technology of China

²Huawei Foundation Model Department

³The Chinese University of Hong Kong

⁴Hong Kong University of Science and Technology
tangfeng32@huawei.com

Abstract

Multimodal Large Language Models (MLLMs) achieve strong multimodal reasoning performance, yet we identify a recurring failure mode in long-form generation: as outputs grow longer, models progressively drift away from image evidence and fall back on textual priors, resulting in ungrounded reasoning and hallucinations. Interestingly, Based on attention analysis, we find that MLLMs have a latent capability for late-stage visual verification that is present but not consistently activated. Motivated by this observation, we propose Visual Re-Examination (VRE), a self-evolving training framework that enables MLLMs to autonomously perform visual introspection during reasoning without additional visual inputs. Rather than distilling visual capabilities from a stronger teacher, VRE promotes iterative self-improvement by leveraging the model itself to generate reflection traces, making visual information actionable through information gain. Extensive experiments across diverse multimodal benchmarks demonstrate that VRE consistently improves reasoning accuracy and perceptual reliability, while substantially reducing hallucinations, especially in long-chain settings. Code is available at <https://github.com/Xiaobu-USTC/VRE>.

1 Introduction

Multimodal Large Language Models (MLLMs) have made rapid progress on multimodal understanding and reasoning tasks (Liu et al., 2023; Bai et al., 2025b; Wang et al., 2024b; Liao et al., 2025; Lin et al., 2025). Yet, in long-form multimodal generation, we observe a persistent failure mode: as reasoning becomes longer, the model gradually drifts away from image evidence and falls back on textual priors, a phenomenon we refer to as *visual drift* (Su et al., 2025; Zhang et al., 2025b; Yang



Figure 1: MLLMs can lose focus on important visual information, which may result in incomplete or inaccurate responses. Our Visual Re-Examination (VRE) leverages the model’s latent ability for implicit re-attention, guiding it to revisit images, extract missing visual evidence, and produce more accurate and grounded outputs.

et al., 2025c; Hu et al., 2024; Liu et al., 2025b). This drift is especially harmful in problems that require late-stage visual verification, such as reading small text or resolving fine-grained spatial relations, where it often leads to unsupported reasoning and hallucinations (Liu et al., 2025a).

Existing attempts to mitigate visual drift largely follow two directions. One line of work relies on explicit visual interventions, such as injecting high-resolution patches, adding visual markers, or interleaving new visual tokens (Zheng et al., 2026; Hong et al., 2025; Zhang et al., 2025c; Chern et al., 2025; Su et al., 2025; Yang et al., 2023; Cai et al., 2024; Chen et al., 2025a). These methods can improve grounding, but they modify the standard autoregressive pipeline and increase computational

[†]Equal contribution. ^{*}Corresponding authors. [‡]Project leader.

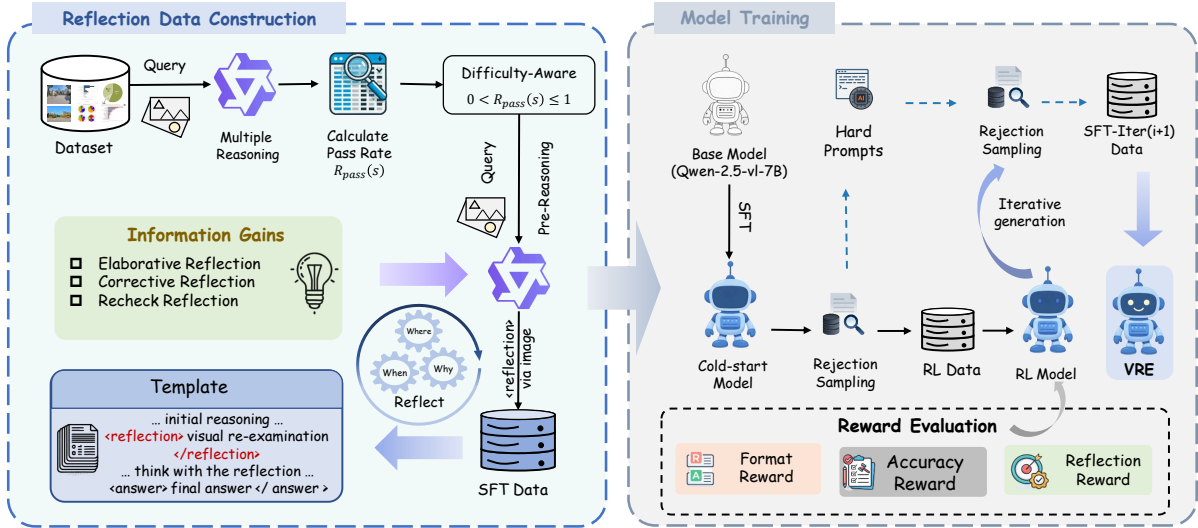


Figure 2: Overall pipeline of the proposed Visual Re-Examination (VRE) framework. Left: Reflection data construction, where difficulty-aware filtering and information-gain-guided reflection synthesis generate high-quality SFT data. Right: Model optimization, where the cold-start model is iteratively improved through rejection sampling and reinforcement learning under a structured reward scheme comprising format, accuracy, and reflection rewards.

overhead. Another line of work shows that MLLMs can sometimes re-attend to the original image without external additions (Yang et al., 2025b). However, this behavior is latent and unstable: prompt nudging may occasionally trigger re-examination, but the resulting re-look is stochastic and often collapses into textual paraphrasing rather than targeted evidence retrieval.

This yields a direct motivation chain for our work: visual drift creates a need for late-stage re-verification; MLLMs already possess a latent but unreliable ability to look back; and naive reflection still lacks guidance on *what* missing evidence should be retrieved. Therefore, the key challenge is not only to trigger re-examination, but to make it purposeful, evidence-driven, and learnable.

To address this challenge, we propose **Visual Re-Examination (VRE)**, a self-iterative framework that strengthens implicit visual re-attention during long-form reasoning. VRE requires no architectural changes, additional visual inputs, or external teacher models. Instead, it trains the model to revisit its own reasoning trace, retrieve missing visual evidence, and use that evidence to repair or verify the current reasoning trajectory. The core supervision signal is **Reflection Information Gain**, which distinguishes genuine visual evidence acquisition from redundant self-paraphrasing.

This motivation also explains our training design. Cold-start SFT is used to elicit the model’s latent re-examination behavior in a structured format, but imitation alone is insufficient for difficult reasoning

errors. RLVR then focuses optimization on unsettled samples near the model’s capability boundary, rewarding both final-answer correctness and meaningful reflection. Because prolonged RL alone can lead to over-optimization and reward hacking, we finally apply Post-RL SFT to consolidate successful trajectories produced by the improved policy, stabilize the acquired behavior, and expand the solvable boundary through a closed self-improvement loop. A pivotal design choice throughout this process is **homologous reconstruction**: reflections are generated by concatenating the model’s own reasoning trace with the original multimodal input and re-feeding them into the same policy model, which preserves perception-reasoning alignment and avoids the distribution mismatch introduced by external teachers.

In this work, our key contributions are summarized as follows:

- We propose VRE, a self-iterative training framework for strengthening implicit visual re-examination in MLLMs. By coupling homologous reconstruction with a closed-loop SFT–RLVR–Post-RL-SFT pipeline, VRE transforms sporadic re-looking behavior into a stable and learnable policy.
- We introduce Reflection Information Gain as an evidence-driven supervision signal that encourages reflections to retrieve genuinely useful visual evidence rather than degenerate into redundant textual paraphrasing.
- We provide mechanistic analysis of VRE from

both the attention and training-dynamics perspectives, including decoding-stage visual re-attention behavior and the evolution of effective reflections across training stages.

- Extensive experiments demonstrate that VRE delivers consistent gains on diverse multimodal benchmarks, improving long-chain visual reasoning while mitigating hallucinations.

2 Related Work

2.1 Reasoning with Visual Information

Executing complex multimodal tasks requires MLLMs to maintain sustained attention on visual evidence throughout extended reasoning chains (Zhang et al., 2025b; Wang et al., 2025b). However, recent studies reveal a prevalent “visual drift” phenomenon: as the generated text lengthens, models increasingly rely on their internal language priors, neglecting the original image and producing ungrounded hallucinations (Amizadeh et al., 2020; Garcez et al., 2019; Thawakar et al., 2025; Bai et al., 2023; Hurst et al., 2024; Xu et al., 2025). To mitigate this drift, previous methodologies have predominantly adopted *explicit intervention* strategies. These include dynamically injecting high-resolution image patches (Zheng et al., 2026; Hong et al., 2025; Zhang et al., 2025c), superimposing visual prompts like bounding boxes (Yang et al., 2023; Cai et al., 2024), or directly interleaving token-level visual features into textual reasoning steps (Chen et al., 2025a). While effective, explicitly injecting new visual signals inevitably alters the input processing pipeline, requires complex heuristic designs, and incurs substantial computational overhead. Conversely, recent research demonstrates that MLLMs possess a latent capability to implicitly refocus on visual information without requiring any additional visual inputs (Yang et al., 2025b). Nevertheless, a critical limitation persists across both explicit and naive implicit approaches: they typically lack strict guidance on what specific visual evidence to retrieve. Without such direction, the model’s re-look behavior often degenerates into redundant self-paraphrasing rather than capturing missing visual clues.

2.2 Post-Training for MLLMs

Post-training optimization has emerged as the de facto standard for unlocking complex reasoning capabilities in large models. Spurred by the paradigm-

shifting success of OpenAI’s o1 (Jaech et al., 2026), the field has increasingly transitioned from module-centric designs to data-driven approaches, heavily leveraging SFT on large-scale synthetic chain-of-thought (CoT) trajectories. Beyond SFT, Reinforcement Learning (RL) has transcended its initial applications in human preference alignment (Bai et al., 2022; Ouyang et al., 2022). Recent breakthroughs demonstrate that Reinforcement Learning with Verifiable Rewards (RLVR) (Guo et al., 2025; Shao et al., 2024; Lambert et al., 2025) can profoundly elevate reasoning proficiency in structured domains such as mathematics and code generation. However, directly adapting RLVR to Multimodal Large Language Models (MLLMs) presents unique and critical bottlenecks. Extensive RL fine-tuning in vision-language tasks frequently induces severe performance degradation. This decay is primarily driven by over-optimization (Gao et al., 2023) and multimodal reward hacking (Casper et al., 2023), where the model learns to game the verification system by generating verbose textual illusions while entirely ignoring the actual visual input. Furthermore, prolonged RL trajectories often exacerbate the catastrophic forgetting of foundational, generic visual capabilities (Zhai et al., 2023).

Our framework addresses these post-training pitfalls via a Self-Iterative Training Strategy. To mitigate textual reward hacking and ensure intermediate reasoning remains anchored to the image, we introduce Reflection Information Gain as the core criterion for data curation and reward formulation. By penalizing redundant self-paraphrasing and rewarding extraction of novel visual evidence, VRE ensures that optimization reinforces authentic, image-grounded deduction. This regularization enables the MLLM to continuously advance its reasoning boundaries without visual degradation.

3 Method

We frame Visual Re-Examination (VRE) as a motivation-first self-evolution pipeline for long-form multimodal reasoning. The key problem is not only that visual drift weakens grounding, but that the model’s latent re-examination behavior is unstable and ungoverned. VRE therefore turns occasional re-looking into a learnable policy for purposeful evidence retrieval.

Accordingly, the method is organized around three questions: why information-gain-driven reflection is needed, why closed-loop RLVR plus

Post-RL SFT is required, and how homologous reconstruction keeps the process aligned with the model’s own visual manifold. We answer these questions below before giving the formal optimization details.

3.1 Reflection Data Construction For Cold-Start

The objective of this phase is to elicit and standardize the policy model’s latent visual reflection behavior using its own distribution. Given a base MLLM π_θ , we build a reflection data construction pipeline for cold-start SFT as shown in Fig. 2.

Why reflection needs an information-gain constraint. The first design requirement is that reflection should not become redundant paraphrasing. Instead, it should retrieve missing visual evidence. We therefore use the model’s own rollout distribution to identify when re-examination is useful, and then use an information-gain criterion to distinguish informative reflections from empty rewording.

Formalizing Reflection Information Gain. We define Reflection Information Gain at the level of individual reflection blocks. For a trajectory τ containing K reflection blocks $r_1, \dots, r_K$, let $I_k \in \{0, 1\}$ indicate whether r_k both re-examines the image and contributes new, question-relevant visual evidence that is absent from the preceding reasoning. A block receives $I_k = 1$ only when it supports elaboration, correction, or verification; paraphrasing, unsupported answer changes, irrelevant visual details, and empty re-look statements receive $I_k = 0$. We aggregate the block-level decisions as

$$s_{\text{refl}}(\tau) = \sum_{k=1}^K I_k, \quad R_{\text{refl}}(\tau) = \mathbf{1}[s_{\text{refl}}(\tau) > 0]. \quad (1)$$

Thus, a trajectory receives a reflection reward if and only if it contains at least one evidence-gaining reflection. This definition prevents redundant verbosity from being rewarded and provides the criterion used both to filter cold-start trajectories and to score reflection quality during RLVR.

1) Difficulty-Aware Sample Partitioning. Given a multimodal sample $s = (i, q)$, where i denotes the image, q the question, and a_s the ground-truth answer, we draw N independent rollout trajectories from the base policy π_θ

using stochastic autoregressive sampling with a non-zero temperature. Each rollout trajectory τ is a full generated sequence including intermediate reasoning tokens and a final answer. We define the pass rate

$$P_{\text{pass}}(s) = \frac{1}{N} \sum_{\tau \in \mathcal{T}(s)} \mathbf{1}[\tau_{\text{ans}} = a_s]. \quad (2)$$

This quantity serves as a proxy for the model’s intrinsic competence on sample s . Based on $P_{\text{pass}}(s)$, we partition the dataset into stable, intractable, and unstable regimes.

Stable samples preserve mastered visual perception and format priors. Intractable samples mark problems beyond the model’s current competence boundary and are excluded from imitation-first distillation. Unstable samples are the key training signal: they expose queries where the model can sometimes answer correctly and sometimes fail, indicating that the latent re-examination ability exists but is not reliably triggered.

2) Reflection Synthesis with an Information-Gain Constraint. To transform rollout trajectories into structured and learnable visual re-examination signals, we standardize the response format as illustrated in Fig. 2. This template explicitly separates initial reasoning from reflective re-grounding, thereby enforcing an internal two-stage inference process.

Following the standardized template, the segment under "think with the image", denoted as τ_{img} , is extracted from the first-stage reasoning of a sampled rollout $\tau \in \mathcal{T}(s)$. To synthesize the `<reflection>` block, we concatenate τ_{img} with the original input $s = (i, q)$ and feed the augmented prompt back into the same policy model for a second rollout. This design explicitly conditions the model on its own prior reasoning trace, enabling deliberate visual re-examination.

To regularize the reflection behavior, we impose a strict **Information Gain** criterion, ensuring that the generated reflection introduces additional visual evidence, corrective grounding, or robustness verification beyond the initial reasoning. Specifically, we synthesize three distinct types of reflection steps based on filtered subsets of trajectories:

- 1. Elaborative Reflection** for $\tau \in \mathcal{T}_{\text{stable}}(s)$: when initial reasoning is consistently correct, we encourage broader visual exploration and evidence grounding to reinforce confidence and enhance reasoning coherence.

2. **Corrective Reflection** for $\tau \in \mathcal{T}_{\text{unstable}}^{\text{wrong}}(s)$: the reflection identifies prior visual misinterpretations, localizes erroneous regions, and re-grounds the reasoning on correct visual evidence to repair the inference trajectory.
3. **Recheck Reflection** for $\tau \in \mathcal{T}_{\text{unstable}}^{\text{right}}(s)$: although the initial reasoning is correct, this step enforces deliberate visual re-attention to verify evidence integrity, mitigate spurious correlations, and minimize accidental correctness, thereby enhancing robustness

Overall, this two-stage formulation encourages the model to revisit the image conditioned on its own intermediate reasoning, producing a structured and self-corrective visual re-examination process.

3) Cold-Start SFT. After generating VRE trajectories, we retain only those rollouts $\tau \in \mathcal{T}(s)$ whose final prediction satisfies $\tau_{\text{ans}} = a_s$ and whose reflection contains at least one valid evidence-gaining block according to $R_{\text{refl}}(\tau) = 1$. The retained trajectories form a filtered supervision set $\mathcal{D}_{\text{cold}}$. The training objective is as follows:

$$\mathcal{L}_{\text{cold-start}} = -\mathbb{E}_{(s,\tau) \sim \mathcal{D}_{\text{cold}}} \sum_{t=1}^{|\tau|} \log \pi_{\theta}(\tau_t \mid i, q, \tau_{<t}), \quad (3)$$

Then we obtain the cold-start model $\pi_{\text{SFT}}^{(0)}$, which is endowed with VRE capability.

3.2 Closed-Loop RLVR and SFT

3.2.1 RL Data Construction.

We apply the $\pi_{\text{SFT}}^{(0)}$ to perform dynamic rejection sampling on the training queries. For each query, multiple reflection-formatted trajectories are generated, and only those passing verifiable outcome checks are retained. To focus RL optimization on the model’s cognitive boundary, we filter out both intractable queries (hard prompts) and already-mastered cases. The remaining unsettled queries, in which the model shows partial correctness, reasoning instability, or sensitivity to reflection, form a focused and information-dense RL data pool.

3.2.2 GRPO and Reward Evaluation.

We optimize the policy using Group Relative Policy Optimization (GRPO) by sampling a group of G trajectories $\{\tau_1, \dots, \tau_G\}$ for each query. To comprehensively assess the multi-step reasoning trajectory, the total reward function $R(\tau_i)$ is formulated as a weighted combination of three components:

$$R(\tau_i) = \lambda_{\text{form}} R_{\text{form}}(\tau_i) + \lambda_{\text{acc}} R_{\text{acc}}(\tau_i) + \lambda_{\text{refl}} R_{\text{refl}}(\tau_i). \quad (4)$$

The coefficients λ_{form} , λ_{acc} , and λ_{refl} reflect implementation priorities. Specifically, we define:

Algorithm 1 Reward Computation for GRPO

Require: Generated trajectory τ , Ground Truth a_s

Ensure: Total Reward $R(\tau)$

- 1: **Parse:** $(\tau_{\text{refl}}, \tau_{\text{ans}}) \leftarrow \text{Parse}(\tau)$
 - 2: **Format Reward** R_{form} :
 - 3: **if** tags unbalanced OR required blocks missing **then**
 - 4: $R_{\text{form}} \leftarrow -1$
 - 5: **else if** τ_{refl} empty OR meaningless **then**
 - 6: $R_{\text{form}} \leftarrow -2$
 - 7: **else if** $\text{len}(\tau_{\text{ans}}) \geq 1000$ **then**
 - 8: $R_{\text{form}} \leftarrow -1$
 - 9: **else**
 - 10: $R_{\text{form}} \leftarrow 0$
 - 11: **end if**
 - 12: **Reflection Score:**
 - 13: $s_{\text{refl}} \leftarrow \text{LLM}_{\text{scorer}}(\tau)$
 - 14: $R_{\text{refl}} \leftarrow \mathbb{I}(s_{\text{refl}} > 0 \wedge \tau_{\text{refl}} \text{ exists})$
 - 15: **Accuracy Score:**
 - 16: $R_{\text{acc}} \leftarrow \text{LLM}_{\text{judge}}(\tau_{\text{ans}}, a_s)$
 - 17: **Combine:**
 - 18: $R(\tau) \leftarrow \lambda_{\text{form}} R_{\text{form}} + \lambda_{\text{acc}} R_{\text{acc}} + \lambda_{\text{refl}} R_{\text{refl}}$
 - 19: **return** $R(\tau)$
-

- **Format Reward (R_{form}):** We incorporate a structural penalty component (0, -1 , or -2) to enforce format integrity. This mechanism penalizes tag mismatches (e.g., unbalanced `<reflection>` or `<answer>` tags), missing structural blocks, excessive output length, and semantically vacuous reflection content, thereby ensuring robust trajectory parsing.
- **Accuracy Reward (R_{acc}):** A binary reward (0 or 1) determined by an *LLM-based judge*. The extracted answer content τ_{ans} is compared against the ground-truth answer a to evaluate semantic correctness, ensuring robust assessment of the final outcome.
- **Reflection Reward (R_{refl}):** A binary reward (0 or 1) that evaluates the quality of the `<reflection>` block. The trajectory τ is judged using an *LLM-based scorer*, encouraging reflection that meaningfully builds upon the initial visual reasoning (information-gain) rather than empty verbosity.

The detailed reward computation procedure is presented in Algorithm 1. After reinforcement learning, we obtain the updated policy $\pi_{\text{RL}}^{(0)}$, which exhibits improved structural compliance, reflection consistency, and answer accuracy.

Table 1: **Performance on Mathematical and Logical Reasoning Benchmarks.** We compare VRE against representative closed-source, open-source, and reasoning-focused MLLMs. * denotes our reproduced results.

Model	Tool Use	Param Size	MathVista	MathVerse	MathVision	WeMath	LogicVista	MMMU
Closed-Source MLLMs								
GPT-4o (OpenAI et al., 2024)	✗	-	63.8	50.8	30.4	69.0	53.2	69.1
Claude 3.7 (Anthropic)	✗	-	66.8	52.0	41.3	72.6	-	-
Open-Source General MLLMs								
LLaVA-OneVision (Li et al., 2024)	✗	7B	58.6	19.3	18.3	20.9	33.3	48.8
Qwen2.5-VL (Bai et al., 2025b)	✗	7B	68.2	49.2	23.7*	62.1	45.9	51.2*
Qwen3-VL (Bai et al., 2025a)	✗	8B	77.2	62.1	53.9	73.5*	55.3	69.6
InternVL2.5 (Chen et al., 2025b)	✗	8B	64.4	39.5	22.0	-	-	56.0
InternVL3 (Zhu et al., 2025)	✗	8B	71.6	39.8	29.3	37.1	-	65.6
Open-Source Reasoning MLLMs								
DeepEyesV2 (Hong et al., 2025)	✓	7B	71.9	52.7	28.9	38.1	48.7	-
Thyme (Zhang et al., 2025c)	✓	7B	70.0	39.1	27.6	39.3	49.0	-
Semantic-back (Yang et al., 2025b)	✗	7B	71.6	50.5	27.7	71.3	-	52.0*
MM-Eureka (Meng et al., 2025)	✗	8B	67.1	40.0	24.7	60.1	-	-
R1-VL (Zhang et al., 2025a)	✗	8B	63.5	40.4	22.2	58.7	-	-
VRE (Ours, Qwen2.5-VL)	✗	7B	71.2	53.1	26.5	68.7	48.7	52.1
(Δ vs Qwen2.5-VL-7B-Instruct)	-	-	+3.0	+3.9	+2.8	+6.6	+2.8	+0.9
VRE (Ours, Qwen3-VL)	✗	8B	81.4	66.7	56.2	78.6	57.9	72.5
(Δ vs Qwen3-VL-8B-Instruct)	-	-	+4.2	+4.6	+2.3	+5.1	+2.6	+2.9

3.2.3 Post-RL Self-Distillation

While the RL model $\pi_{\text{RL}}^{(0)}$ substantially outperforms the preceding cold-start model $\pi_{\text{SFT}}^{(0)}$, we observe a key limitation: simply extending RL training does not yield sustained improvements and may even cause degradation. Prolonged optimization often leads to reward overfitting, training instability, and degradation of general visual capabilities.

To bridge this performance gap, we introduce an iterative refinement phase to close the self-evolution loop. Specifically, the RL policy $\pi_{\text{RL}}^{(0)}$ revisits previously intractable prompts through iterative generation. Leveraging its enhanced visual grounding, the model resolves a subset of these challenges. We apply rejection sampling to curate high-quality trajectories, forming an augmented SFT dataset $\mathcal{D}_{\text{new}}$. Fine-tuning the RL model on $\mathcal{D}_{\text{new}}$ yields an updated policy $\pi_{\text{SFT}}^{(1)}$.

4 Experiments

4.1 Experimental Setup

Implementation Details. To validate our proposed framework, we select Qwen2.5-VL-7B (Bai et al., 2025b) as our primary base model for all SFT and RLVR optimization, given its strong open-source baseline performance. For SFT, training is conducted on $8 \times$ Ascend 910B(64G) NPUs. The RLVR phase scales to $64 \times$ Ascend 910B NPUs, employing a 3D parallelism strategy with tensor, pipeline, and data parallelism degrees set to 4, 4, and 4, respectively. Our training infrastructure

is built upon LLaMA-Factory for SFT and the Mindspeed-RL framework for reinforcement learning. The cold-start data is from Thyme-SFT (Zhang et al., 2025c), and the RL training data is from MM-Eureka (Meng et al., 2025), V^* dataset (Wu and Xie, 2023), and ViRL39K (Wang et al., 2025a). Details regarding the training data, hyperparameters for the cold-start and GRPO stages, and the judge model prompt are provided in the Appendix.

Evaluation Benchmarks. To comprehensively evaluate our framework, we select a diverse suite of benchmarks across four crucial dimensions. Mathematical Reasoning (MathVerse (Zhang et al., 2024), MathVision (Wang et al., 2024a), MathVista (Lu et al., 2024), WeMath (Qiao et al., 2024), LogicVista (Xiao et al., 2024)) tests the model’s ability to deduce complex visual math problems in both multiple-choice and free-form formats. High-Resolution Perception (V^* -Bench (Wu and Xie, 2023)) evaluates fine-grained visual grounding and the localization of minute details. Visual Document Understanding (ChartQA (Masry et al., 2022), OCRBench_v2 (Fu et al., 2025)) assesses precision in data extraction and verification from complex charts and dense text. Finally, Multi-Disciplinary Reasoning (MMMU (Yue et al., 2024)) gauges expert-level generalizability and domain knowledge alignment across college-level disciplines.

4.2 Main Results

To verify the cross-generational generalizability of our framework, we instantiate VRE on both Qwen2.5-VL-7B-Instruct and Qwen3-VL-8B-

Table 2: **Performance on Perception, Document Understanding, and Real-world QA.** We compare VRE against representative closed-source, open-source, and reasoning-focused MLLMs. * denotes our reproduced results.

Model	Tool Use	Param Size	V*-Bench			ChartQA			OCRBench_v2		RealworldQA
			Attr	Spatial	Overall	Human	Aug	Overall	en	zh	-
Closed-Source MLLMs											
GPT-4o (OpenAI et al., 2024)	✗	-	60.5	67.5	79.5	91.9	85.7	86.7	46.5	32.2	75.5
Open-Source General MLLMs											
LLaVA-OneVision (Li et al., 2024)	✗	7B	75.7	75.0	75.4	-	-	80.0	-	-	66.3
InternVL2.5 (Chen et al., 2025b)	✗	8B	-	-	-	-	-	84.8	49.8	52.1	70.1
Qwen2.5-VL (Bai et al., 2025b)	✗	7B	78.2*	73.6*	76.4*	72.5	94.9	83.7	56.3	57.2	68.2
Qwen3-VL (Bai et al., 2025a)	✗	8B	93.9*	84.2*	90.1*	83.1*	95.7*	89.4*	65.4	61.2	71.5
Qwen2.5-VL (Bai et al., 2025b)	✗	32B	77.4*	86.8*	81.2*	76.9	82.6	81.1	-	-	70.2
Open-Source Reasoning MLLMs											
SEAL (Wu and Xie, 2023)	✓	7B	74.8	76.3	75.4	-	-	-	-	-	-
DyFo (Li et al., 2025)	✓	7B	80.0	82.9	81.2	-	-	-	-	-	-
DeepEyesV2 (Hong et al., 2025)	✓	7B	-	-	81.8	-	-	88.4	-	-	-
Thyme (Zhang et al., 2025c)	✓	7B	83.5	80.3	82.2	80.0	92.2	86.1	-	-	70.2
Semantic-back (Yang et al., 2025b)	✗	7B	80.2*	78.7*	79.6*	-	-	-	-	-	-
VRE (Ours, Qwen2.5-VL)	✗	7B	84.3	82.9	83.8	87.3	90.4	88.8	62.6	64.7	69.8
(Δ vs Qwen2.5-VL-7B-Instruct)	-	-	+6.1	+9.3	+7.4	+14.8	-4.5	+5.1	+6.3	+7.5	+1.6
VRE (Ours, Qwen3-VL)	✗	8B	95.7	88.2	92.6	86.6	96.8	91.7	69.9	68.5	73.8
(Δ vs Qwen3-VL-8B-Instruct)	-	-	+1.8	+4.0	+2.5	+3.5	+1.1	+2.3	+4.5	+7.3	+2.3

Instruct. We comprehensively evaluate performance across four major domains: Mathematical and Multi-Disciplinary Reasoning (Table 1), as well as High-Resolution Perception and Structured Visual Understanding (Table 2).

A striking observation across our evaluations is that VRE achieves high performance among 7B-scale models without relying on explicit tool invocation or architectural modifications. For instance, on the High-Resolution V*-Bench (Table 2), VRE yields an impressive Overall score of 83.8%, completely outperforming tool-augmented reasoning models such as Thyme (82.2%), DeepEyesV2 (81.8%), and DyFo (81.2%). Furthermore, compared to the base model Qwen2.5-VL-7B, VRE delivers substantial absolute gains across perception and extraction tasks, including +7.4% on V*-Bench, +5.1% on ChartQA, and up to +7.5% on OCRBench_v2. This robust improvement confirms that implicit visual re-examination natively solves the visual drift problem without external tools.

As illustrated in Table 1, VRE exhibits exceptional capabilities in deep logical tracking and mathematical deduction. Specifically, VRE achieves 71.2 on MathVista and 53.1 on MathVerse, securing absolute gains of +3.0% and +3.9% over the base model, respectively. This consistent trajectory of improvement extends across other rigorous benchmarks, yielding a striking +6.6% gain on WeMath (68.7) and solid enhancements on MathVision (+2.8%), LogicVista (+2.8%), and MMMU (+0.9%). Crucially, despite operating with only 7 billion parameters, VRE significantly narrows,

and in several cases completely bridges, the performance gap with massive closed-source proprietary systems. This parameter-efficient capability empirically validates that our self-iterative reflection mechanism effectively unlocks deep cognitive reasoning without relying on brute-force parameter expansion.

4.2.1 Anatomy of Visual Re-Examination.

Visual Attention Re-allocation. To verify that VRE genuinely reactivates visual processing rather than hallucinating based on text, we visualize the average attention weights over visual tokens across decoding steps. As shown in Figure 3, the base model suffers from severe visual decay: as the reasoning chain lengthens, the attention to image features monotonically fades, indicating a drift toward pure textual reliance. In sharp contrast, VRE exhibits a distinct visual attention resurgence. Specifically, upon careful re-examination of the content, within the generated <reflection> block, the attention weights on visual tokens spike significantly. This phenomenon confirms that VRE successfully reactivates the dormant visual encoder to re-examine evidence, effectively reversing the visual drift and grounding the reasoning back to the image.

Introspective Visual Re-Examination. While the attention matrix confirms the temporal return of visual signals, we further investigate the spatial mechanism of this re-examination: What exactly does the model look at when the visual signal re-

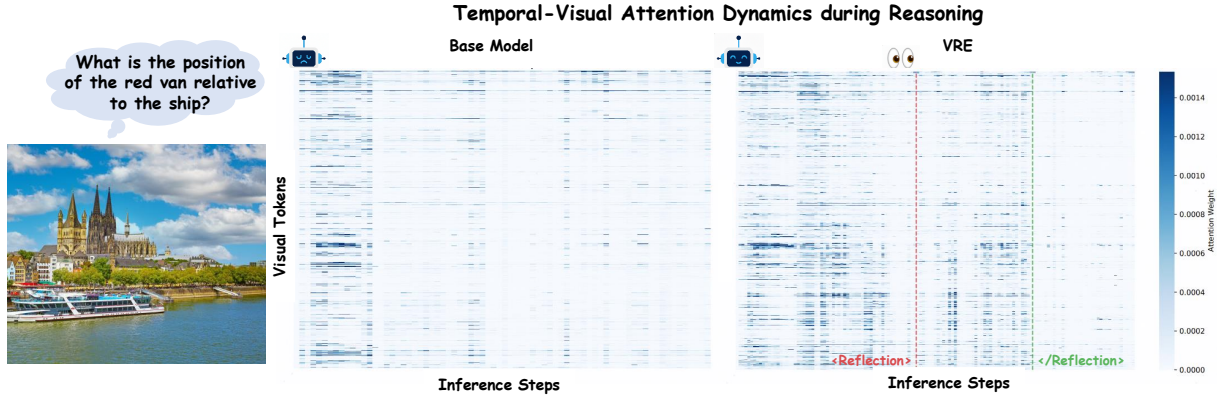


Figure 3: **Visualizing Implicit Visual Re-Examination via Attention.** The heatmaps display attention weights on visual tokens across inferences steps. **(Left)** The Base Model exhibits visual decay, where attention to the image vanishes as the textual context grows, leading to ungrounded hallucinations. **(Right)** The VRE Model demonstrates a spontaneous attention resurgence. Inside the `<reflection>` block, the model sharply re-allocates attention back to the visual features, proving that the mechanism actively triggers a re-examination behavior to extract missing visual evidence.



Figure 4: **Mechanism of Corrective Visual Re-Grounding.** (a) **Initial Reasoning:** Attention is dispersed over irrelevant background elements, causing a failure to ground the target object. (b) **Visual Re-Examination:** The emergence of the `<reflection>` token triggers an active visual search, shifting the attention spotlight precisely onto the dustpan (red region) to retrieve veridical evidence.

turns? Figure 4 provides a qualitative visualization of the spatial attention shift. In the Initial Reasoning (Phase 1), the model exhibits typical “visual blindness,” where its attention drifts to irrelevant background textures, failing to ground the target object. Lacking visual evidence, the model hallucinates a color based on textual priors. However, during the Reflection (Phase 2), the VRE mechanism triggers a targeted Corrective Re-grounding. As shown in the heatmap, the attention spotlight does not just randomly increase; it performs a precise search-and-lock operation, shifting specifically to the dustpan’s spatial coordinates. This qualitative evidence completes our mechanistic understanding: the re-examination process works by (1) identifying missing information (When), (2) spatially re-grounding the specific entity (Where), and (3) extracting the correct attribute to overwrite the hal-

lucination (Why).

Evolution of Reflection Paradigms. To deeply understand how the VRE framework guides the model to master implicit visual re-examination, we track the fine-grained distribution of reflection paradigms across different training stages. As defined in Sec. 3.1, generated reflections are categorized into three valid Information Gain types, alongside a No Gain category representing redundant self-paraphrasing. To quantify this evolution, we randomly sample 100 test trajectories per benchmark to analyze the proportion of these four categories. Figure 5 illustrates this evolutionary distribution as stacked bar charts across six diverse benchmarks. After the initial Cold-Start SFT, the model exhibits merely a superficial understanding of introspection, generating No Gain reflections

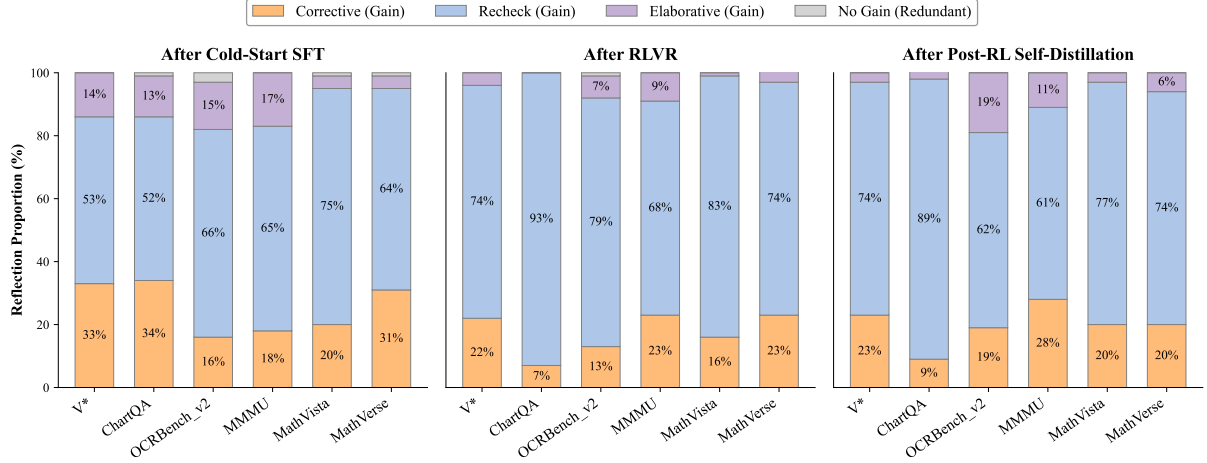


Figure 5: **Evolution of Reflection Paradigms.** The stacked bar charts detail the distribution of introspection types across training stages.

relatively frequently. This confirms that explicitly mimicking the <reflection> template does not guarantee genuine visual grounding. However, after the RLVR and the subsequent Post-RL Self-Distillation, the proportion of No Gain drops drastically, approaching zero. This decisively validates our Reflection Reward, which successfully penalizes empty verbosity and enforces tangible visual evidence extraction. First, the frequency of Corrective reflections is notably higher in tasks like V^* , MMMU, and MathVerse. This aligns perfectly with their inherent challenges: V^* requires locating elusive visual details easily missed during initial perception, while MMMU involves deep, complex logical chains. In these scenarios, the model intelligently learns to use the image to debug and rectify its early visual misinterpretations or flawed reasoning paths. Second, Recheck behaviors are distinctly prominent in mathematical and chart-based benchmarks. Since these tasks heavily rely on precise data extraction, the model learns to execute a final, rigorous visual verification step before outputting the answer to prevent careless hallucinations. Conversely, Elaborative reflections see the highest ratio in OCR-heavy and multi-disciplinary contexts. For tasks requiring dense text reading and rich contextual gathering, the model naturally favors explicitly augmenting its reasoning with supplementary visual transcriptions and descriptive details to build a solid premise. Ultimately, these fine-grained structural adaptations firmly prove that VRE cultivates a genuinely intelligent, self-directed visual introspection mechanism that dynamically tailors its cognitive strategy to the specific difficulty of diverse multimodal tasks.

5 Conclusion

In this paper, we address the ‘visual drift’ phenomenon in long-form multimodal generation, where models gradually neglect visual evidence in favor of textual priors. We introduce Visual Re-Examination (VRE), a self-iterative framework that elicits implicit visual refocusing without expensive architectural interventions or explicit tool use. By integrating a Reflection Information Gain metric into a closed-loop pipeline spanning data curation and RLVR optimization, VRE enforces targeted visual evidence retrieval to correct reasoning trajectories. Extensive evaluations across different generations of leading base models (Qwen2.5-VL-7B-Instruct and Qwen3-VL-8B-Instruct) demonstrate its robust generalizability and outstanding performance, establishing a reliable paradigm for the continuous self-evolution of MLLMs.

Limitations

While the VRE framework demonstrates robust performance and cross-generational generalizability, it remains susceptible to over-optimization (reward hacking) during prolonged training. As revealed in our step scaling analysis, forcing the RLVR optimization beyond 150 steps triggers a noticeable performance collapse where the model attempts to exploit the reward system. Although our current pipeline successfully mitigates this pitfall by terminating RL early and consolidating the policy via Post-RL SFT, in our future work, we plan to design a theoretically principled reward regularizer and an adaptive early-stopping mechanism to fully decouple training stability from step-wise heuristics.

References

- Saeed Amizadeh, Hamid Palangi, Alex Polozov, Yichen Huang, and Kazuhito Koishida. 2020. [Neuro-symbolic visual reasoning: Disentangling "Visual" from "Reasoning"](#). In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 279–290. PMLR.
- Anthropic. [Claude 3.7 sonnet system card](#).
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. [Qwen2.5-vl technical report](#). *Preprint*, arXiv:2502.13923.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. [Training a helpful and harmless assistant with reinforcement learning from human feedback](#). *Preprint*, arXiv:2204.05862.
- Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. 2024. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12914–12923.
- Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, and 13 others. 2023. [Open problems and fundamental limitations of reinforcement learning from human feedback](#). *Preprint*, arXiv:2307.15217.
- Xinyan Chen, Renrui Zhang, Dongzhi JIANG, Aojun Zhou, Shilin Yan, Weifeng Lin, and Hongsheng Li. 2025a. [Mint-cot: Enabling interleaved visual tokens in mathematical chain-of-thought reasoning](#). In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 69110–69139. Curran Associates, Inc.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025b. [Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling](#). *Preprint*, arXiv:2412.05271.
- Ethan Chern, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu. 2025. Thinking with generated images. *arXiv preprint arXiv:2505.22525*.
- Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, and 5 others. 2025. [Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning](#). *Preprint*, arXiv:2501.00321.
- Leo Gao, John Schulman, and Jacob Hilton. 2023. [Scaling laws for reward model overoptimization](#). In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10835–10866. PMLR.
- Artur d’Avila Garcez, Marco Gori, Luis C Lamb, Luciano Serafini, Michael Spranger, and Son N Tran. 2019. Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *arXiv preprint arXiv:1905.06088*.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. [Deepseek-r1 incentivizes reasoning in llms through reinforcement learning](#). *Nature*, 645(8081):633–638.
- Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. 2025. Deepeyesv2: Toward agentic multimodal model. *arXiv preprint arXiv:2511.05271*.
- Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, and Ranjay Krishna. 2024. Visual sketchpad: Sketching as a visual chain of thought for multimodal language models. *arXiv preprint arXiv:2406.09403*.
- Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel,

- Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 399 others. 2024. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helvar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, and 244 others. 2026. [Openai o1 system card](#). *Preprint*, arXiv:2412.16720.
- Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2025. [Tulu 3: Pushing frontiers in open language model post-training](#).
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. [Llava-onevision: Easy visual task transfer](#).
- Geng Li, Jinglin Xu, Yunzhen Zhao, and Yuxin Peng. 2025. Dyfo: A training-free dynamic focus visual search for enhancing lmms in fine-grained visual understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9098–9108.
- Jiaqi Liao, Yuwei Niu, Fanqing Meng, Hao Li, Changyao Tian, Yinyu Du, Yuwen Xiong, Dianqi Li, Xizhou Zhu, Li Yuan, Jifeng Dai, and Yu Cheng. 2025. [Langbridge: Interpreting image as a combination of language embeddings](#).
- Bin Lin, Zongjian Li, Xinhua Cheng, Yuwei Niu, Yang Ye, Xianyi He, Shenghai Yuan, Wangbo Yu, Shadong Wang, Yanyang Ge, Yatian Pang, and Li Yuan. 2025. [Uniworld-v1: High-resolution semantic encoders for unified visual understanding and generation](#).
- Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. 2025a. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *arXiv preprint arXiv:2505.21523*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025b. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.
- Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. [ChartQA: A benchmark for question answering about charts with visual and logical reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.
- 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. 2025. [Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning](#). *Preprint*, arXiv:2503.07365.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.
- Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Yifan Zhang, Xiao Zong, Yida Xu, Muxi Diao, Zhimin Bao, Chen Li, and Honggang Zhang. 2024. [We-math: Does your large multimodal model achieve human-like mathematical reasoning?](#) *Preprint*, arXiv:2407.01284.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Zhaochen Su, Linjie Li, Mingyang Song, Yunzhuo Hao, Zhengyuan Yang, Jun Zhang, Guanjie Chen, Jiawei Gu, Juntao Li, Xiaoye Qu, and Yu Cheng. 2025. Openthinking: Learning to think with images via visual tool reinforcement learning. *arXiv preprint arXiv:2505.08617*.
- Omkar Thawakar, Dinura Dissanayake, Ketan Pravin More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan,

- Yuhao Li, Ilmuz Zaman Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, Hisham Cholakkal, Ivan Laptev, Mubarak Shah, Fahad Shahbaz Khan, and Salman Khan. 2025. [LlamaV-o1: Rethinking step-by-step visual reasoning in LLMs](#). pages 24290–24315.
- Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhui Chen. 2025a. [V1-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning](#). *arXiv preprint arXiv:2504.08837*.
- Ke Wang, Juntong Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. [Measuring multimodal mathematical reasoning with math-vision dataset](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 95095–95169. Curran Associates, Inc.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. [Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution](#). *arXiv preprint arXiv:2409.12191*.
- Yikun Wang, Siyin Wang, Qinyuan Cheng, Zhaoye Fei, Liang Ding, Qipeng Guo, Dacheng Tao, and Xipeng Qiu. 2025b. [Visuothink: Empowering lvm reasoning with multimodal tree search](#). *arXiv preprint arXiv:2504.09130*.
- Penghao Wu and Saining Xie. 2023. [V*: Guided visual search as a core mechanism in multimodal llms](#). *Preprint*, arXiv:2312.14135.
- Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. [Logicvista: Multimodal llm logical reasoning benchmark in visual contexts](#). *arXiv preprint arXiv:2407.04973*.
- Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025. [Llava-cot: Let vision language models reason step-by-step](#). *Preprint*, arXiv:2411.10440.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025a. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023. [Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v](#). *arXiv preprint arXiv:2310.11441*.
- Shuo Yang, Yuwei Niu, Yuyang Liu, Yang Ye, Bin Lin, and Li Yuan. 2025b. [Look-back: Implicit visual re-focusing in mllm reasoning](#). *arXiv preprint arXiv:2507.03019*.
- Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. 2025c. [R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization](#). *arXiv preprint arXiv:2503.10615*.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruochi 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, and 3 others. 2024. [Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.
- Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2023. [Investigating the catastrophic forgetting in multimodal large language models](#). *arXiv preprint arXiv:2309.10313*.
- Jingyi Zhang, Jiaying Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. 2025a. [R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization](#). *arXiv preprint arXiv:2503.12937*.
- Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. 2024. [Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?](#) *Preprint*, arXiv:2403.14624.
- Xintong Zhang, Zhi Gao, Bofei Zhang, Pengxiang Li, Xiaowen Zhang, Yang Liu, Tao Yuan, Yuwei Wu, Yunde Jia, Song-Chun Zhu, and Qing Li. 2025b. [Adaptive chain-of-focus reasoning via dynamic visual search and zooming for efficient vlms](#). *Preprint*, arXiv:2505.15436.
- Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, Haonan Fan, Kaibing Chen, Jiankang Chen, Haojie Ding, Kaiyu Tang, Zhang Zhang, Liang Wang, Fan Yang, Tingting Gao, and Guorui Zhou. 2025c. [Thyme: Think beyond images](#). *arXiv preprint arXiv:2508.11630*.
- Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. 2026. [Deepeyes: Incentivizing "thinking with images" via reinforcement learning](#). *Preprint*, arXiv:2505.14362.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. [Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models](#). *Preprint*, arXiv:2504.10479.

A Implementation Details

A.1 Data Resources

We systematically curate a high-quality data composition specifically tailored for our three-stage pipeline: (1) Cold-Start SFT for embedding visual reflection priors, (2) RLVR optimization via GRPO, and (3) Post-RL Self-Distillation. The detailed dataset sources and quantities are summarized in Table 3.

Table 3: **Detailed Composition of Training Data.** This table summarizes the data sources and the approximate number of samples in our VRE Framework.

Data Source	Cold-Start SFT	RLVR	Post-RL SFT
Thyme-SFT	$\sim 80k$	-	-
MM-Eureka	-	$\sim 20k$	$\sim 4k$
V^* Dataset	-	$\sim 20k$	$\sim 4k$
ViRL39K	-	$\sim 20k$	$\sim 4k$
Total Data	$\sim 80k$	$\sim 60k$	$\sim 12k$

Data Provenance and Split Protocol. For Thyme-SFT, we retain only the multimodal queries and discard all original teacher-generated responses. For each retained query, both the initial reasoning and the reflection trajectory used for Cold-Start SFT are generated by our own policy model. Therefore, Thyme-SFT serves only as a query source; VRE does not distill response-level reasoning or reflections from Thyme’s teacher model. This is the setting we refer to as *homologous reconstruction*: the input queries may originate from an external dataset, whereas the response and reflection trajectories are reconstructed by the same model family used for training.

Training and evaluation data are strictly separated throughout the three-stage pipeline. Cold-Start SFT, RLVR, and Post-RL SFT use only the designated training sources listed in Table 3. In particular, the V^* training source used for RLVR and Post-RL SFT is kept separate from the official held-out V^* -Bench evaluation split. Likewise, MMMU and ChartQA are evaluated on their official held-out evaluation splits. No evaluation sample is used for SFT, RLVR, Post-RL SFT, or model selection.

A.2 Hyperparameters and Configuration

We detail the hyperparameter configurations for SFT and RLVR in Tables 4 and 5, respectively. During trajectory rollouts, we fix the sampling count to $N = 8$ to balance diversity and cost. For the RLVR

stage, we utilize GRPO with strict KL divergence constraints. Both the LLM_{judge} and LLM_{scorer} utilized during the RLVR phase are instantiated using the Qwen2.5-72B (Yang et al., 2025a). For the text generation during RL, we configure the maximum response length to 7168 tokens and set the temperature to 1.0 to encourage extensive exploration of reasoning trajectories. For the RL configuration, we set $\gamma = 1.0$, $\lambda = 0.95$, and an initial KL coefficient of 0.01. The GRPO group size is set to $G = 10$, with reward coefficients balanced to prioritize reasoning correctness while penalizing redundant reflections.

A.3 Reflection Information Gain and Scorer Rubric.

The formal definition of Reflection Information Gain and its reward is given in Sec. 3. Here we provide the operational scoring protocol. For each reflection block, the scorer assigns 1 only when the block both re-inspects the image and adds new, question-relevant visual evidence supporting elaboration, correction, or verification; otherwise it assigns 0. Hard negatives include paraphrasing, unsupported answer changes, irrelevant visual details, and empty re-look statements.

We use Qwen2.5-72B-Instruct as the reward scorer, not as a teacher generating policy trajectories. On 100 manually audited trajectories, it achieves 96% accuracy against human judgments. On an additional 1,000 trajectories sampled across training subsets, reflection types, and RL checkpoints, GPT-5 and Qwen2.5-72B-Instruct reach 95.3% agreement. These audits support scorer reliability across the sampled training distributions.

B Additional Experiments

B.1 Component Breakdown and Evolutionary Dynamics.

To evaluate the individual contribution of each core component within our self-iterative pipeline, we conduct a rigorous ablation study on major benchmarks. All ablation experiments are instantiated on top of the Qwen2.5-VL-7B-Instruct base model.

Efficacy of Cold-Start SFT: Activating Visual Grounding. As shown in the second row of Table 6, SFT Only yields immediate and substantial gains on perception-heavy and structured understanding tasks. This indicates that supervised fine-tuning successfully establishes the format prior for the `<reflection>` template, allowing the

Table 4: Hyperparameters for SFT and Inference.

SFT Training Hyperparameters			Inference / Rollout Hyperparameters	
Parameter	Cold-Start	Post-RL	Parameter	Value
Global Batch Size	64	64	Max New Tokens	4900
Learning Rate	1e-5	5e-7	Temperature	1.0
LR Scheduler	Cosine	Cosine	Top- p	1.0
Max Sequence Len	4900	4900	Top- k	-1
Num Epochs	3	1	Rollout Trajectories (N)	8

Table 5: Hyperparameter configurations for RLVR.

Category	Configuration
Gen.	max_t=7168, min_t=50, temp=1.0, top_p=1.0
RL Alg.	$\gamma = 1.0, \lambda = 0.95,$ adv_est=group_norm
KL Ctrl.	type=fixed, init_c=0.01, penalty=kl
GRPO	clip_r=0.2, group_s=10, $\lambda_{\text{form}} = 0.4, \lambda_{\text{acc}} = 0.6,$ $\lambda_{\text{refl}} = 0.4$

Table 6: **Component Ablation.** Performance progression with the incremental addition of VRE components.

Setting	V*-Bench	MathVista	MathVerse	ChartQA	OCRBench_v2
Base	76.4	68.2	49.2	83.7	56.3/57.2
SFT Only	80.1	69.2	49.6	87.8	59.6/62.3
SFT + RL	83.8	71.0	52.5	88.7	62.1/64.3
VRE	84.3	71.2	53.1	88.8	62.6/64.7

model to consciously trigger visual re-examination for data extraction. However, the gain on complex reasoning tasks is marginal (e.g., MathVerse only improves by +0.4%), suggesting that SFT alone is insufficient to correct deep logical errors.

Impact of RLVR: Unlocking Deep Reasoning.

As shown in the third row of Table 6, the integration of SFT + RLVR catalyzes a critical breakthrough in reasoning capabilities. Most notably, MathVerse witnesses a remarkable leap of +2.9% (49.6 $\rightarrow$ 52.5), and V*-Bench further climbs to 83.8. This confirms that our Reflection Rewards (R_{refl}) are essential drivers. They transform the model’s behavior from merely mimicking the reflection format to actively utilizing visual evidence for logical self-correction, thereby solving complex problems that SFT could not handle.

Necessity of Post-RL SFT: Stabilizing the Policy.

Finally, As shown in the fourth row of Table 6, the full VRE pipeline provides a consistent stability boost across all benchmarks. By salvaging “hard prompts” that were initially unstable during RL exploration, VRE pushes the performance ceiling further (e.g., MathVerse reaches 53.1, MathVista reaches 71.2). While the magnitude of these gains is smaller compared to the RL stage, their universality across all domains proves that the self-iterative loop effectively consolidates the policy and expands the model’s cognitive boundary into the long-tail distribution of difficulty.

B.2 Scaling Laws for SFT and RL

To deeply understand the optimization bounds of our VRE framework, we systematically ablate the data scale for cold-start SFT and the training steps for RLVR.

Cold-Start SFT Data Scaling. As shown in Figure 6a, we train the base model with varying quantities of curated cold-start SFT reflection data (10k, 30k, 50k, and 80k). The performance exhibits a logarithmic scaling law: accuracy improves rapidly as data scales up to 50k, demonstrating that the model efficiently internalizes the reflection format prior. However, the marginal gains significantly diminish between 50k and 80k (e.g., MathVista only increases by +0.1%). Therefore, we select 80k as the optimal SFT capacity to fully saturate the format alignment without incurring redundant training costs.

RL Step Scaling. Figure 6b tracks the RLVR performance at intervals of 50, 100, 150, and 200 steps. The performance peaks around 100 steps, marking the optimal point where the policy successfully masters information-gain extraction. Crucially, forcing the optimization beyond 150 steps leads to a dramatic performance collapse. This empirical degradation perfectly illustrates the phe-

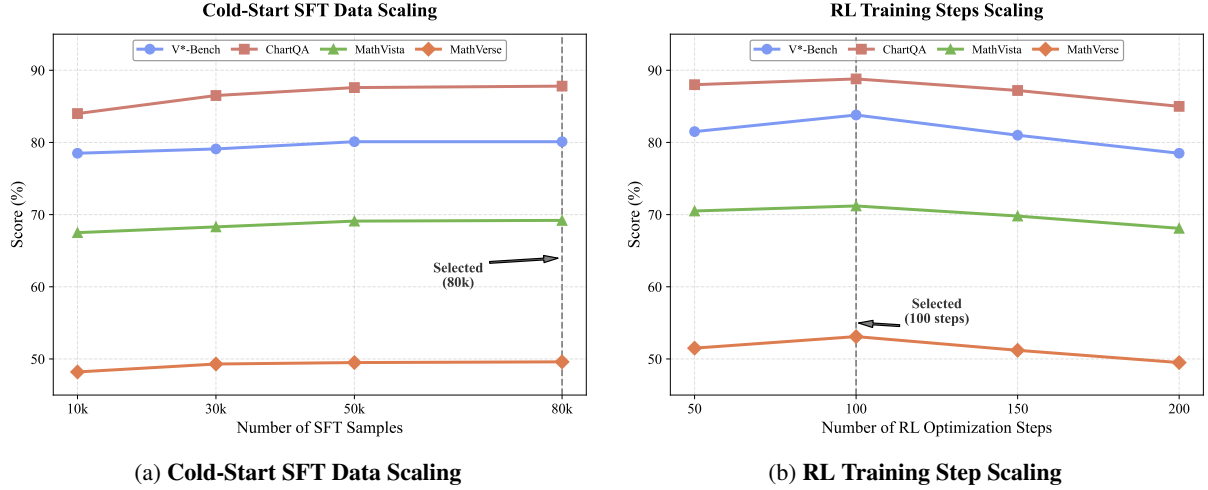


Figure 6: **Scaling Laws in VRE Training.** (a) Cold-Start SFT Data Scaling. There is no significant performance improvement after 50k samples, and the performance is approaching saturation. (b) RL Training Step Scaling. The policy achieves optimal reasoning performance at 100 steps. Prolonged training (> 150 steps) induces severe over-optimization and capability degradation.

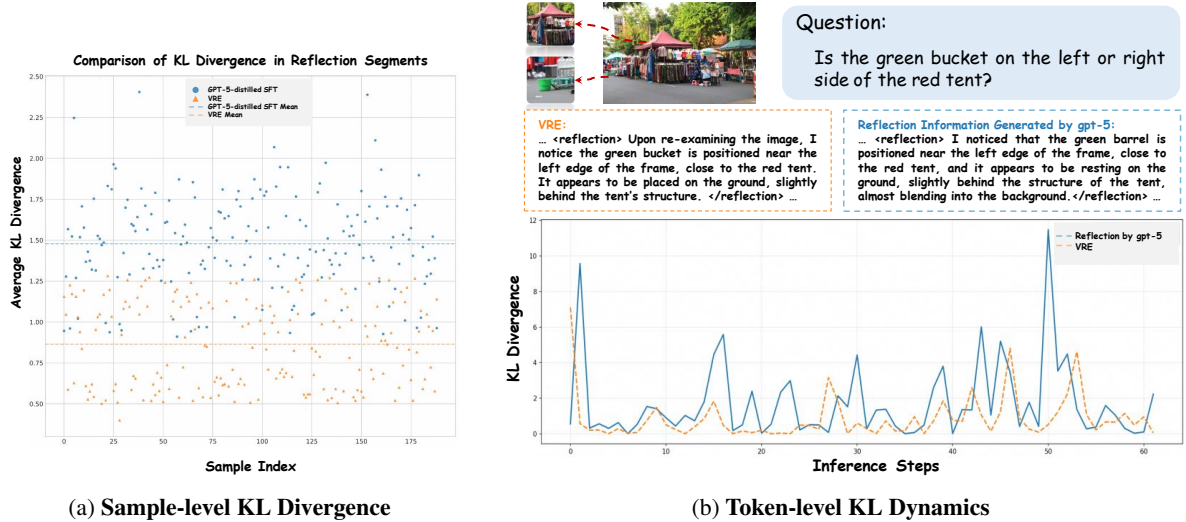


Figure 7: **KL Divergence Analysis on V^* -Bench.** (a) Sample-wise average KL divergence across the dataset. (b) A specific case study showing token-level KL divergence over inference steps.

nomenon of over-optimization (reward hacking). This finding decisively justifies our strategy to halt RL at 100 steps and rely on Post-RL SFT to stabilize the policy.

B.3 Trajectory Distribution

To mechanistically explain these benchmark gains, we track the outcome distribution of rejection sampling across training stages (Table 7). From a global perspective (Full Set), the transition from cold-start model to Post-RL SFT model demonstrates a massive stabilization effect: the proportion of perfectly mastered queries (*All Correct*) surges from 30.8% to 55.6%, while intractable queries (*All Wrong*) are more than halved (12.7% $\rightarrow$ 6.0%). The Post-RL Self-Distillation stage further consolidates this stability, pushing *All Correct* to 58.1%.

Crucially, in a targeted boundary analysis, we isolate the exact subset of “All Wrong” prompts that the cold-start model completely failed to solve. Remarkably, the Post-RL SFT policy successfully generates at least one correct trajectory in 68.1% of these previously unsolvable cases (58.9% *Mixed* + 9.2% *All Correct*). This provides definitive statistical proof that VRE does not merely memorize reasoning formats, but genuinely achieves a Boundary Breakthrough, enabling the Post-RL SFT stage to successfully salvage and distill these hard prompts into reliable capabilities.

B.4 Controlled Validation of VRE

Matched Same-Query Control. We first test whether the gains can be explained simply by ex-

Table 7: **Distributions of Rollout Trajectories Across Training Stages.** *Full Set* denotes the complete dataset used for RL optimization. “*All Wrong*” refers to the specific subset of “hard prompts” where the Cold-Start model failed across all sampled trajectories. This subset highlights the model’s ability to breakthrough its initial cognitive boundaries.

Models	Set	All Correct ($\uparrow$)	Mixed	All Wrong ($\downarrow$)
Cold-Start	Full Set	30.8%	56.5%	12.7%
RL Model	Full Set	55.6%	38.4%	6.0%
Post-RL SFT	Full Set	58.1%	36.2%	5.7%
Post-RL SFT	<i>All Wrong</i>	9.2%	58.9%	31.9%

posure to the same training query pool. We train a matched control from the same Qwen2.5-VL-Instruct base model using the same queries and training budget, but without VRE. As shown in Table 8, Full VRE consistently outperforms this control across all evaluated benchmarks. The improvement therefore cannot be attributed to query exposure alone.

Causal Intervention: Blocking Visual Access during Reflection. Attention visualization indicates that visual attention resurges inside `<reflection>` blocks, but attention patterns alone are correlational. We therefore perform a controlled intervention in which reflection tokens are prevented from attending to the original visual tokens during `<reflection>...</reflection>`. The model parameters, textual context, reflection structure, decoding settings, and token budget are otherwise kept unchanged. Table 9 shows a consistent degradation on all five benchmarks, including a drop of 8.4 points on V^* -Bench and 5.6 points on ChartQA. This intervention provides causal evidence that VRE’s gains depend on retrieving visual evidence during reflection rather than merely generating additional text or following a particular output format.

Component Isolation under a Matched Budget. We further isolate the contribution of individual VRE components under matched training budgets. Specifically, we remove the explicit `<reflection>` boundaries, remove information-gain filtering during SFT, remove the reflection reward during RL, or replace homologous reconstruction with GPT-5-generated reflections. Results are summarized in Table 10. Removing only the reflection boundaries has a negligible effect, suggesting

that the tags mainly organize the re-examination phase rather than directly produce the gains. In contrast, removing information-gain filtering or the reflection reward causes consistent degradation, while replacing homologous reconstruction with GPT-5-generated reflections produces the largest drop on V^* and ChartQA. Together with the saturation of Cold-Start SFT and the degradation under prolonged RL in Fig. 6, these matched controls argue against a simple compute-scaling or formatting explanation.

Reflection-Reward Ablation. The reflection-reward ablation is especially informative because it keeps the base model, queries, rollout budget, GRPO steps, optimization settings, format reward, and accuracy reward identical; the only difference is setting $\lambda_{\text{refl}} = 0$. Full VRE improves over this compute-matched variant by +2.6 on V^* -Bench, +3.2 on ChartQA, +1.4 on MathVista, and +2.9 on MathVerse. This isolates the contribution of the reflection-quality signal from the effects of generic RL optimization.

B.5 Task-Specific Diagnostics and Limitations

MathVision: Gains Primarily Arise from Better Visual Grounding. VRE improves the same Qwen2.5-VL-7B base model on MathVision from 23.7 to 26.5, although its absolute score remains below Semantic-back and DeepEyesV2. To identify the source of this gain, we evaluate the responses along two visual-grounding dimensions: *visual comprehensiveness*, which measures coverage of question-relevant visual evidence, and *visual accuracy*, which measures the correctness of the visual evidence used in reasoning. Under the same evaluation protocol, VRE improves visual comprehensiveness from 84.17 to 89.73 (+5.56) and visual accuracy from 78.75 to 86.59 (+7.84), as shown in Table 11. These results indicate that the main benefit on MathVision is more complete and accurate use of visual evidence.

Unlike methods that inject stronger-model-generated response trajectories, VRE keeps policy trajectories self-generated and does not import response-level mathematical reasoning from a stronger proprietary model; DeepEyesV2 additionally uses external tools. This distinction is consistent with the observed pattern: VRE yields clear grounding gains but a smaller improvement in final mathematical reasoning. We therefore characterize VRE primarily as a method for strengthening

Table 8: **Matched same-query control.** Both settings use the same Qwen2.5-VL-Instruct base model, training queries, and training budget. The control removes VRE.

Setting	V^*	ChartQA	OCRBench_v2	MathVista	MathVerse
Same queries without VRE	78.5	85.6	60.1/61.5	69.2	50.8
Full VRE	84.3	88.8	62.6/64.7	71.2	53.1
VRE gain	+5.8	+3.2	+2.5/+3.2	+2.0	+2.3

Table 9: **Causal visual-masking intervention.** During reflection, visual access is blocked while all other inference conditions are kept fixed.

Setting	V^*	ChartQA	OCRBench_v2	MathVista	MathVerse
Full VRE	84.3	88.8	62.6/64.7	71.2	53.1
Visual masking	75.9	83.2	55.7/55.8	68.3	49.5
Drop	-8.4	-5.6	-6.9/-8.9	-2.9	-3.6

sustained visual grounding, rather than as a mechanism for importing additional mathematical knowledge or tool-based reasoning capability.

ChartQA-Aug Regression. The effect of VRE is not uniform across all ChartQA splits and backbones. For Qwen2.5-VL, VRE improves ChartQA-Human by 14.8 points and the overall ChartQA score by 5.1 points, but decreases ChartQA-Aug by 4.5 points. The Qwen2.5-VL baseline on ChartQA-Aug is already high at 94.9, which may leave limited headroom and increase sensitivity to distributional differences between the Human and Aug splits. Importantly, this regression is not reproduced on Qwen3-VL, where ChartQA-Aug improves by 1.1 points.

B.6 Statistical Verification of the Reflection Mechanism.

Beyond macroscopic benchmark improvements, we employ information-theoretic metrics to rigorously validate two fundamental properties of our VRE framework: the training stability of same-family formatting, and the genuine utility of the generated visual reflections.

Distributional Affinity via KL Divergence. A critical design choice in VRE is utilizing the model’s own policy distribution, rather than distilling traces from a stronger proprietary model (e.g., GPT-5), to construct the reflection dataset. To validate this necessity, we isolate the effect of data provenance by examining policies strictly after the cold-start SFT phase. Given a base MLLM π_θ , we analyze the distribution shift by

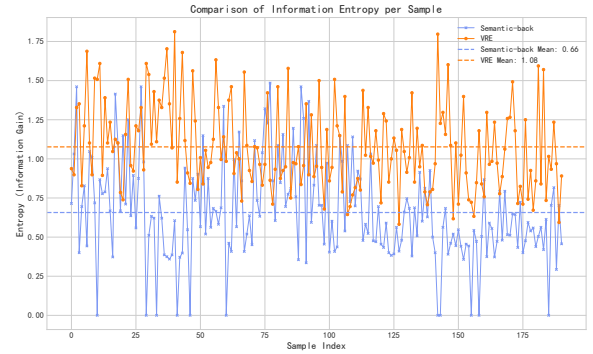


Figure 8: **Information Entropy Comparison on V^* -Bench.** The observed higher conditional entropy in VRE demonstrates that our reflection mechanism effectively injects novel, non-redundant visual evidence, thereby mitigating textual redundancy.

computing the Kullback-Leibler (KL) Divergence $D_{KL}(\pi_{SFT} \parallel \pi_\theta)$ on the V^* -Bench dataset, where π_{SFT} represents obtained cold-start model. Specifically, we compare the divergence of a cold-start SFT model trained on GPT-5-distilled reflections ($\pi_{SFT}^{(gpt)}$) against an equivalent policy trained entirely on our self-generated VRE cold-start data ($\pi_{SFT}^{(0)}$).

As visualized in Fig. 7(a), the $\pi_{SFT}^{(gpt)}$ exhibits a significantly higher average divergence across the entire dataset. Furthermore, the token-level case study (Fig. 7(b)) reveals that it maintains consistently elevated KL values throughout the entire reflection generation process. This confirms that cross-model distillation forces the student to mimic alien reasoning patterns that deviate sharply from its native manifold. In contrast, VRE’s self-generated cold-start SFT maintains a substantially

Table 10: **Component-level ablations under matched training budgets.** Values in parentheses denote changes relative to Full VRE.

Setting	V^*	ChartQA	MathVista	MathVerse
Full VRE	84.3	88.8	71.2	53.1
Without <reflection> boundaries	83.8 (-0.5)	88.9 (+0.1)	71.5 (+0.3)	53.0 (-0.1)
No information-gain filtering	81.7 (-2.6)	86.1 (-2.7)	69.5 (-1.7)	50.3 (-2.8)
No reflection reward	81.7 (-2.6)	85.6 (-3.2)	69.8 (-1.4)	50.2 (-2.9)
GPT-5-generated reflections	78.0 (-6.3)	84.5 (-4.3)	70.3 (-0.9)	51.6 (-1.5)

Table 11: **Visual-grounding analysis on MathVision.**

Model	Visual comprehensiveness	Visual accuracy
Base model	84.17	78.75
VRE	89.73	86.59
Gain	+5.56	+7.84

lower KL divergence, ensuring that the acquired reflection capability is organically integrated and structurally compatible with the base model’s intrinsic priors.

Information Gain via Conditional Entropy. A rigorous validation is performed to confirm that our <reflection> block injects genuine visual evidence rather than merely paraphrasing the preceding context (redundancy), and we quantify Reflection Information Gain by employing Conditional Information Entropy as a proxy.

Formally, for a reflection sequence $Y_{\text{ref}} = \{y_1, y_2, \dots, y_T\}$, we calculate its conditional entropy as the average Negative Log-Likelihood (NLL) of the model’s output distribution conditioned on context X_{ctx} :

$$H(Y_{\text{ref}} | X_{\text{ctx}}) = \frac{1}{T} \sum_{t=1}^T -\log \pi_{\theta}(y_t | X_{\text{ctx}}, y_{<t}), \quad (5)$$

where $\pi_{\theta}(y_t | X_{\text{ctx}}, y_{<t})$ denotes the conditional probability of token y_t predicted by the policy π_{θ} . In practice, this metric is computed using the CrossEntropyLoss between the model’s logits and the ground-truth target tokens.

A low entropy value indicates that the reflection is highly predictable given the textual context, suggesting the generation is redundant (“textual inertia”). Conversely, a high entropy score implies the introduction of external, non-redundant visual details—a proxy for the acquisition of novel information. As visualized in Figure 8, VRE exhibits a significantly higher average entropy ($\bar{H}_{\text{VRE}} = 1.08$)

compared to the Semantic-back (Yang et al., 2025b) ($\bar{H}_{\text{Semantic-back}} = 0.66$). This substantial divergence confirms that VRE successfully breaks the textual inertia, compelling the model to retrieve targeted visual evidence rather than collapsing into degenerate self-repetition.

B.7 Training Dynamics and Stability

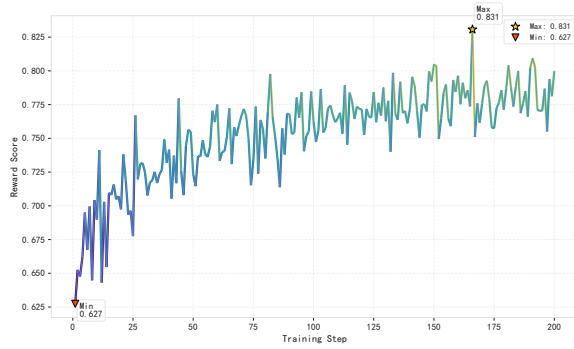
We visualize the trajectory of key optimization metrics during the GRPO phase in Fig. 9. The curves reveal two critical insights regarding the stability and efficiency of our self-iterative strategy:

Rapid Convergence via Reliable Self-Distillation.

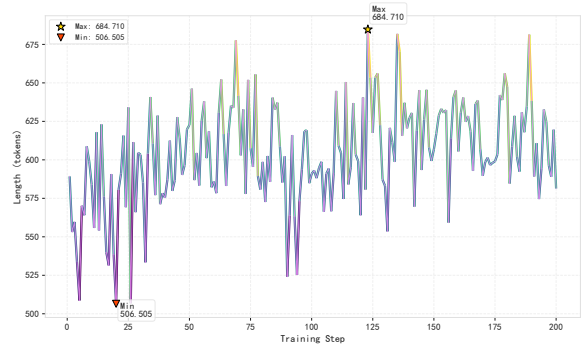
As shown in Fig. 9(a), the reward score converges within 50 steps. This accelerated alignment stems from the high distributional affinity of our self-evolution data; by initializing the policy within its native manifold rather than distilling from alien distributions, our framework enables the RL algorithm to focus immediately on optimizing reasoning paths.

Mitigation of Reward Hacking. Crucially, the average response length (Fig. 9(b)) stabilizes at ~ 600 tokens, avoiding the common RL pitfall of unbounded verbosity. This stability confirms that the model has internalized the objective of maximizing information density, successfully circumventing reward hacking via self-repetition or meaningless tautology.

As shown in Fig. 10, the loss curve during Post-RL SFT remains exceptionally stable with minimal fluctuation. This empirical observation yields a



(a) Reward Score Evolution



(b) Response Length Stability

Figure 9: **Training Dynamics of GRPO.** (a) The reward score steadily increases and converges around step 50, reflecting stable policy optimization. (b) The average response length stabilizes at ~ 600 tokens, indicating that the model extracts visual evidence efficiently without degenerating into verbose patterns.

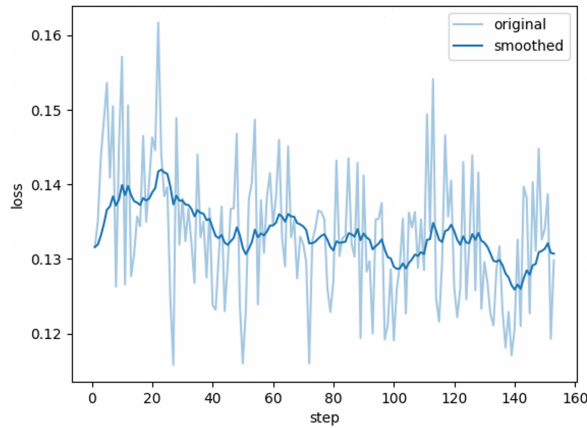


Figure 10: **Training Loss during Post-RL SFT.**

profound insight: because the training trajectories are self-generated by the RL-improved policy and rigorously filtered via rollout trajectories, the target data distribution aligns almost perfectly with the model’s native probability manifold. Consequently, Post-RL SFT does not induce violent gradient shocks or catastrophic forgetting; rather, it smoothly and safely consolidates the exploratory gains on “hard prompts” into stable parametric memory, confirming the robustness of our closed-loop self-evolution paradigm.

B.8 Reliability and necessity of the judge/scorer.

We use an LLM-based judge for `<answer>` because rigid string matching can fail under semantically equivalent surface forms (e.g., “1” versus “1.0”). Separately, Qwen2.5-72B-Instruct serves as an external evaluator for reflection quality during RLVR; it does not generate the policy’s answers, reasoning traces, or reflections. All policy trajec-

tories remain self-generated, so this dependency is distinct from strong-to-weak response distillation.

Reflection Information Gain Rubric. The formal definition of Reflection Information Gain and its reward is given in Sec. 3. Here we provide the operational scoring protocol. For each reflection block, the scorer assigns 1 only when the block both re-inspects the image and adds new, question-relevant visual evidence supporting elaboration, correction, or verification; otherwise it assigns 0. Hard negatives include paraphrasing, unsupported answer changes, irrelevant visual details, and empty re-look statements. The complete scorer prompt and strict JSON output format are provided in Fig. 14.

Scorer Reliability. We validate the scorer under two complementary protocols. First, on 100 manually audited trajectories, Qwen2.5-72B-Instruct achieves 96% accuracy against human judgments. Second, we sample another 1,000 trajectories spanning multiple training subsets, reflection patterns, training stages, and RL checkpoints; GPT-5 and Qwen2.5-72B-Instruct agree on 95.3% of the scorer decisions. These results indicate that the reflection-quality signal is stable across the sampled trajectory distributions rather than being tied to a single subset or checkpoint.

C Prompts

To facilitate robust multi-stage training, we design unified templates for both Cold-Start SFT and RLVR. These templates enforce a rigorous output format, ensuring the model adheres to a consistent reasoning trajectory.

System Prompt Template for Information Gains

You are an analytical assistant that solves visual reasoning problems through a structured, iterative process focused on maximizing information gain from the image. Your reasoning must follow two stages:

Stage 1: Iterative Reasoning with Visual Information Gain

- Begin with a <reflection> block: critically examine the initial reasoning and actively seek new visual information by mentally re-examining the image.
- Ask: "What visual details did I miss?"
- Focus on overlooked elements: objects, text, colors, spatial relationships, actions, or contextual cues.
- Each reflection should re-perceive the image to extract additional relevant visual content.
- Follow thinking with the reflection: reason step by step based on current understanding of the image.
- The thinking-reflection cycle ought to be information-dense: brief but introducing new image-derived insights that advance the solution.
- For determining relative positions, please define the positional relationships from the observer's viewpoint.

Stage 2: Answer

- When confident, output your final answer inside <answer>...</answer>.
- No explanations outside the specified tags.

Key principle: Maximize information density—each reflection should be concise but yield substantial new visual insights.

Figure 11: System prompt template for information gains.

Prompt Template for Accuracy Evaluation

Below are two answers to a question. Question is [], [Standard Answer] is the standard answer..., and [Model_answer] is the answer extracted from a model's output... Determine whether these two answers are consistent. If they are consistent, Judgement is 1; if different, Judgement is 0. Just output Judgement and don't output anything else.

[Question]: Is the countertop tan or blue?
[Standard Answer]: The countertop is tan.
[Model_answer]: tan
Judgement: 1

[Question]: On which side of the picture is the barrier?
[Standard Answer]: The barrier is on the left side of the picture.
[Model_answer]: left
Judgement: 1

... (5 few-shot examples omitted for brevity) ...

[Question]: What color is the towel in the center of the picture?
[Standard Answer]: The towel in the center of the picture is blue.
[Model_answer]: The towel in the center of the picture is pink.
Judgement: 0

[Question]: \promptvar{question}
[Standard Answer]: \promptvar{ground_truth}
[Model_answer]: \promptvar{predict_str}
Judgement:

Figure 12: Final accuracy evaluation prompt. Dynamic inputs {question} (Question), {ground_truth} (Standard Answer) and {predict_str} (Model_answer) are inserted at inference time.

Prompt Template for Elaborative Reflection

The question: {question}. The initial reasoning: {initial_reasoning}.
You need to re-examine the image to enrich the reasoning with supplementary visual evidence before concluding.

Start with <reflection>, then think with the reflection.

In each <reflection>:

- Scan the image for supporting details that were omitted in the initial step.
- Elaborate on fine-grained attributes: specific colors, textures, small objects, or background context.
- Explicitly connect these new visual details to strengthen the current reasoning path.
- Goal: Maximize information density without changing the logical direction.

When ready, provide your definitive final answer inside <answer>...</answer>.

Figure 13: Prompt Template for Elaborative Reflection. Dynamic inputs {question} (Question), {initial_reasoning} (Initial Reasoning) are inserted at inference time.

Prompt Template for Reflection Evaluation

You are an expert in reasoning analysis. Your task is to evaluate a provided thinking process to determine whether its reflections demonstrate a genuine attempt to **gain more visual information** from the input image by re-examining or re-perceiving its content.

The thinking process consists of sequential steps: each step of reasoning is followed by a <reflection>...</reflection> block.

Important Context:

- The original problem involves an **image** (though you, as a text-only model, cannot see it).
- The thinker is expected to reason about that image.
- A valuable reflection should show an **attempt to look back at the image** and extract **additional or previously overlooked visual details** relevant to answering the question.

Evaluate each <reflection>...</reflection> block using these criteria:

1. **Visual Re-examination Intent**: Does the reflection suggest re-inspection of the image to uncover new visual elements (objects, spatial relations, colors, text, etc.)?
2. **Relevant Information Gain**: Does this re-examination lead to **new, image-derived information** that meaningfully advances the reasoning?

Only count a reflection if it satisfies **both** criteria.

Output: a single JSON object with key "score" (integer count).

Question:
\promptvar{question}

The Thinking Process:
\promptvar{predict_str}

Output Format (Strictly Enforced)
Return only a valid JSON object-no explanations, no extra text.
Example: {"score": 2}

Your Evaluation Result:

Figure 14: Prompt for evaluating whether reflection blocks demonstrate genuine visual re-examination. Dynamic inputs {question} (Question) and {predict_str} (Model_answer) are inserted at inference time.

Prompt Template for Corrective Reflection

The question: {question}. The initial reasoning: {initial_reasoning}.
The initial reasoning may contain visual hallucinations or logical errors. You must critically re-examine the image to detect and rectify them.

Start with <reflection>, then think with the reflection.

In each <reflection>:

- Act as a critic: Compare the text strictly against the image pixels.
- Identify the exact point where the reasoning deviates from the visual reality (the "visual drift").
- Explicitly state the error (e.g., The text says X, but the image shows Y)
- Re-ground the reasoning using the correct visual evidence.

When ready, provide your corrected final answer inside <answer>...</answer>.

Figure 15: Prompt Template for Corrective Reflection. Dynamic inputs {question} (Question), {initial_reasoning} (Initial Reasoning) are inserted at inference time.

D Case Studies

We present four visualization samples of VRE across diverse scenarios. In particular, Sample 4 highlights the importance of post-RL SFT.

Prompt Template for Recheck Reflection.

The question: {question}. The initial reasoning: {initial_reasoning}.
You need to perform a rigorous visual audit to double-check the key evidence before finalizing the answer .

Start with <reflection>, then think with the reflection.

In each <reflection>:

- Verify precise data points: check specific numbers, text transcription (OCR), or object counts.
- Re-confirm spatial relationships (left/right, above/below) mentioned in the reasoning.
- Ensure no key condition in the question was overlooked.
- Goal: Eliminate carelessness and ensure absolute precision.

When ready, provide your verified final answer inside <answer>...</answer>.

Figure 16: Prompt Template for Recheck Reflection. Dynamic inputs {question} (Question), {initial_reasoning} (Initial Reasoning) are inserted at inference time.

Sample 1: VRE Generated Samples with Corrective Reflection



Question: How many people are in the oil painting?

- (A) 4 people
- (B) 0 people
- (C) 2 people
- (D) 1 people

Source: V*-Bench

Ground truth: C

Let's analyze the image step by step:

1. The image is of a children's playroom.
 2. There is a small oil painting hanging on the wall near the window.
 3. In the painting, there appears to be one figure that looks like a person.
- Given these observations, we can conclude that the oil painting contains 1 person. (Incorrect.)

<reflection>

Upon re-examining the image, I notice the oil painting more closely. It seems to depict a figure holding a child, possibly a mother and baby. This detail was not initially considered.

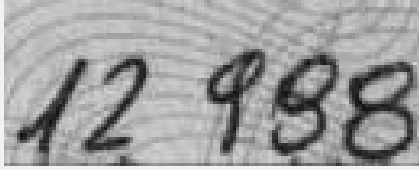
(Corrected using the image; correct answer obtained.)

</reflection>

The painting indeed shows two figures: an adult and a child. Therefore, the number of people in the oil painting is 2.

<answer> C </answer> (Correct.)

Sample 2: VRE Generated Samples with Recheck Reflection



Question: what is the number in the image?

Source: OCRBench_V2

Ground truth: 12988

The image appears to show a handwritten number.

The digits seem to be "12988".

However, the first digit is partially obscured or unclear due to the handwriting style. (Not confident about the answer.)

<reflection>

I will focus on checking whether the first digit is '1'.

Upon re-examining the image, the first digit is '1', then '2', and finally '988'.

The spacing between the digits is noticeable, which helps in distinguishing them clearly.

There are no other characters or numbers present.

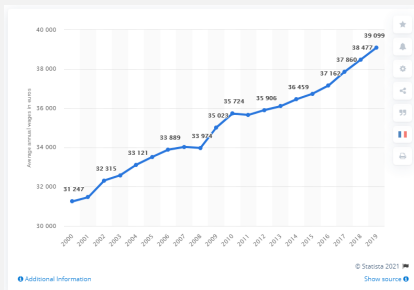
(Corrected rechecking the image; correct answer obtained.)

</reflection>

Given the clear separation and the visible digits, the number in the image is indeed "12988".

<answer> 12988 **</answer>** (Correct.)

Sample 3: VRE Generated Samples with Elaborative Reflection



Question: Which year has the highest average annual wage?

Source: ChartQA

Ground truth: 2019

The graph shows the average annual wages in euros over several years.

To determine which year has the highest average annual wage, I need to look at the data points for each year and identify the highest value.

The data points are as follows:

– 2000: 31,247 euros ... – 2019: 39,099 euros

The highest value is 39,099 euros, which occurs in 2019.

<reflection>

Upon re-examining the image, I notice that the data point for 2019 is slightly higher than the one for 2018.

This detail was not initially considered in the previous reasoning.

(Elaborative reflection: enriched the reasoning with fine-grained spatial evidence.)

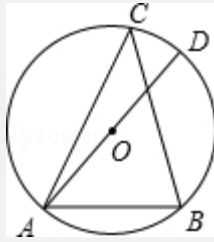
</reflection>

Given the updated observation, the highest average annual wage is indeed 39,099 euros, occurring in 2019.

The slight difference between 2018 and 2019 confirms that 2019 has the highest value.

<answer> 2019 **</answer>** (Correct.)

Sample 4: VRE Generated Samples with Corrective Reflection



Question: As shown in the figure, the length of chord AB is 10.0, and the angle of circumference angle $ACB = 45.0$, then the diameter of the circle AD is () Choices:

A: $5\sqrt{2}$ B: $10\sqrt{2}$ C: $15\sqrt{2}$ D: $20\sqrt{2}$

Source: MathVerse

Ground truth: B

===== **RL Model:** =====

According to the given information, we can connect OA and OB. Since $AB=10$, we have $OA=OB=5$. Therefore, $AD=5\sqrt{2}$. Thus, the correct option is A. (Incorrect.)

<reflection>

Upon re-examining the image, I notice that the triangle AOB is formed by connecting points A, O, and B. The angle ACB is given as 45 degrees, which implies that angle AOB is twice that, making it 90 degrees. This confirms that triangle AOB is a right-angled triangle at O. (Incomplete reflection: failed to detect logical conflict.)

</reflection>

Since $AB = 10$, we have $OA = OB = 5$. The diameter AD is twice the length of OA, so $AD = 2 * 5 = 10$. However, considering the right angle at O, the actual length of AD is $10\sqrt{2}$. Therefore, the correct answer is B.

<answer> B </answer> (Forced the correct answer through broken and contradictory math logic.)

===== **Post-RL SFT Model:** =====

According to the given information, we can connect OA and OB. Since angle $ACB=45^\circ$, it follows that angle $AOB=90^\circ$. Also, since $AB=10$, we have $OA=OB=5$. Therefore, $AD=5\sqrt{2}$. Thus, the correct option is A. (Incorrect.)

<reflection>

Upon re-examining the image, I notice that the triangle AOB is formed by connecting points A, O, and B. The angle ACB is given as 45 degrees, which implies that angle AOB is twice that, making it 90 degrees. This confirms that triangle AOB is a right-angled triangle at O. Additionally, the length of AB is provided as 10 units, if $OA=OB=5$, then Pythagorean theorem is not satisfied. (Corrective reflection: identified the exact mathematical contradiction.)

</reflection>

Given that triangle AOB is a right-angled triangle with AB as the hypotenuse, we can use the Pythagorean theorem to find the lengths of OA and OB. Since $AB = 10$, we have $OA = OB = 5\sqrt{2}$. The diameter AD is twice the length of OA, so $AD = 2 * 5\sqrt{2} = 10\sqrt{2}$. However, considering the right angle at O, the actual length of AD is $10\sqrt{2}$. Therefore, the correct answer is B.

<answer> B </answer> (Correct.)