AnE: Pushing the Reasoning Frontier of Multimodal LLMs via Anchor Evolution

Zehao Wang[1], Yihan Zeng[2†](✉), Zidong Gong[1], Yuanfan Guo[3], Feng Zhu[1], Hongzhi Zhang[1](✉), Wei Zhang[2], and Wangmeng Zuo[1]

zehaowang0704@gmail.com, zengyihan@sjtu.edu.cn, 1539988936@qq.com, gyfastas@gmail.com, 25S003021@stu.hit.edu.cn, zhanghz@hit.edu.cn, wz.zhang@huawei.com, wmzuo@hit.edu.cn

[1] Faculty of Computing, Harbin Institute of Technology, Harbin, China
[2] Noah's Ark Lab, Huawei, China
[3] Independent Researcher, China

Abstract. Post-training via Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL) is crucial for enhancing reasoning in Multimodal Large Language Models (MLLMs), yet existing paradigms often reach a performance bottleneck due to the limitations of static data. While current methods leverage self-reflection or self-evolution to push these boundaries, they still suffer from cognitive drift and hallucinated reasoning paths caused by low-quality synthetic data. To address these challenges, we propose ***An***chor ***E***volution (**AnE**), a new paradigm that integrates truth-anchored data curation and model evolution, achieving faithful and steady performance gains at the reasoning frontier. Specifically, we propose Truth Anchor Expansion, which pinpoints the model *failing frontier* via trajectory rollouts and leverages ground-truth databases to retrieve high-fidelity anchors for faithful data curation. Subsequently, we introduce the Scaffold-Stripping Mechanism to internalize reasoning capabilities. This mechanism first anchors reasoning paths via scaffold-augmented supervision to mitigate the learning complexity and distribution drift of direct SFT on raw data, then leverages RL to strip the scaffold template, thereby effectively transitioning the reasoning paths into intrinsic model capabilities. Experimental results on multimodal reasoning benchmarks show that our method substantially advances the model performance frontier, improving the base model by 10.3% across eight multimodal benchmarks and achieving state-of-the-art results. The code will be publicly available at `https://github.com/wangzehao0704/AnE`.

Keywords: Multimodal Large Language Model · Multimodal Reasoning · Post-training

1 Introduction

Post-training methods, particularly Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL), have become the dominant strategies for developing

[†]Project leader.

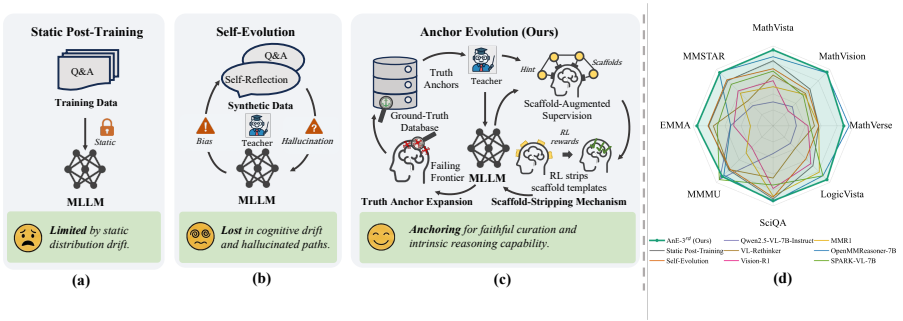


Fig. 1: (a-c) Comparisons of training paradigms. Compared to static post-training and self-evolution paradigms, our proposed Anchor Evolution paradigm retrieves verified truth anchors from ground-truth databases to enable faithful data curation and internalize reasoning capabilities. **(d) Performance comparisons.** Anchor Evolution achieves state-of-the-art or highly competitive results across eight multimodal reasoning benchmarks.

advanced reasoning capabilities in multimodal large language models [14,17,25]. Recent literature highlights a distinct division of roles between the two: SFT is primarily focused on assimilating new knowledge and reasoning patterns, while RL is more about reinforcing existing behavioral patterns, rather than introducing entirely new capabilities [9, 11, 28]. Building on these complementary strengths, modern approaches often combine both SFT and RL techniques. This integration effectively merges knowledge acquisition with behavioral alignment, thereby improving the reliability of reasoning [13, 57].

However, previous post-training frameworks [10, 17, 53] often rely on static training data (including both SFT and RL data), as illustrated in Fig. 1(a). Although such data can initially improve the model's reasoning capability to some extent, it is constrained by a fixed data distribution, which limits its ability to scale or adapt as the model's capabilities evolve. As a result, the model often fails to perform targeted training to address its specific weaknesses in subsequent iterations, leading to a performance bottleneck.

To break this bottleneck, recent Self-Evolution paradigms [7,15,32] advocate synthetic data, as shown in Fig. 1(b). For example, C2-evo [7] evolves both the model and its training data in a closed-loop manner, dynamically adjusting task complexity based on the model's current performance. While these approaches enhance the richness and difficulty of the data, they face two major challenges. First, the QA pairs in the generated data often contain logical errors, which can accumulate and amplify across iterations, leading to catastrophic cognitive drift and ultimately degrading performance. Second, answers in synthetic data are typically sourced directly from the teacher model or human experts, which may diverge significantly from the model's own distribution and capabilities. Consequently, training on such data often leads to superficial imitation of reasoning steps rather than genuine improvement in reasoning ability [2, 4].

Rethinking the post-training paradigm, we observe that the scalability and quality of training data play a decisive role in shaping the model's reasoning trajectory. While static datasets provide reliable supervision, they lack the scalability to adapt to the model's evolving weaknesses, whereas self-evolution methods relying on synthetic data often introduce hallucinated reasoning paths and cognitive drift. These observations suggest that effective model evolution requires both adaptability to the model's failing frontier and reliable knowledge sources. To this end, we introduce *truth anchors*, which are verified anchor samples retrieved from ground-truth databases that serve as high-fidelity references for constructing new training data and preventing hallucinated reasoning trajectories.

Based on this observation, we seek to enable scalable, faithful model evolution. To this end, we propose **An**chor **E**volution (AnE), a post-training paradigm that grounds data evolution in verified real-world knowledge while focusing on the model's failure frontier. The AnE framework comprises three integrated stages: (i) **Failing-Frontier Discovery** precisely identifies the regions where the model underperforms. This is accomplished through multiple rollouts on the seed dataset, ensuring that the training signals are closely aligned with the model's current cognitive capabilities. A teacher model is then used to diagnose failures, extracting failure-relevant keywords and hints to guide subsequent training stages; (ii) **Truth Anchor Expansion** retrieves verified samples from an external, real-world database to serve as high-fidelity anchors, ensuring that model training consistently relies on authentic knowledge and avoids the cognitive drift inherent in synthetic-only cycles; and (iii) **Scaffold-Stripping Mechanism** leverages hints from an external teacher to guide the model toward correct answers and generates scaffold-augmented data for SFT. It then employs RL to remove the scaffolds, enabling the model to internalize this guidance as an intrinsic capability. We conduct our Anchor Evolution (AnE) experiments using the Qwen2.5-VL-7B model as the base. As shown in Fig. 1(d), across eight benchmarks, AnE demonstrates continuous improvement over multiple iterations, enhancing the base model by an average of 10.3%. Notably, it maintains state-of-the-art performance on particularly challenging datasets such as MathVision [42] and EMMA [16].

To sum up, the primary contributions of this work are summarized as follows:

- We build Anchor Evolution (AnE), a post-training framework that tightly couples truth-anchored data curation and model evolution, achieving faithful and steady gains at the multimodal reasoning frontier.
- We propose Truth Anchor Expansion, which pinpoints where the model fails via trajectory rollouts and retrieves verified truth anchors from ground-truth databases to construct high-fidelity training data, thereby avoiding hallucinated reasoning trajectories.
- We propose scaffold-augmented supervision to anchor the model on correct reasoning paths, and then utilize RL to strip the scaffold template, enabling the model to internalize the reasoning capability.

– Extensive experiments on eight multimodal reasoning benchmarks show that AnE achieves consistent gains, improving the base model by 10.3% on average and attaining state-of-the-art performance on challenging datasets such as MathVision and EMMA.

2 Related Work

Post-training for multimodal reasoning. Recent multimodal large language models commonly adopt a post-training pipeline that combines supervised fine-tuning (SFT) and reinforcement learning (RL) [1, 13, 17, 21, 26, 45, 53]. SFT is typically used to provide a cold start with structured reasoning traces, while RL methods (e.g., GRPO [13]) further improve robustness by optimizing task rewards. However, in many pipelines, the SFT/RL datasets are prepared in advance, which keeps the training distribution largely static throughout optimization. As the model evolves, such static data become increasingly misaligned with the model's capability frontier, often resulting in diminishing returns.

Iterative SFT–RL optimization. Beyond one-pass post-training, recent work explores iterative optimization by alternating SFT and RL stages [3,14,23,33,37]. For example, OpenVLThinker [10] performs multiple rounds of SFT and RL using a manually constructed static dataset and reject sampling, gradually introducing harder training samples in later stages. Such schedules can improve stability, for instance by reducing policy drift, and help the model focus on hard samples exposed during RL. Nevertheless, the data construction in these pipelines is often static or only weakly adaptive [3], which limits sustained progress at the reasoning frontier. In contrast, we evolve the training data with Truth Anchor Expansion to target the model's failing frontier while iterating SFT and RL.

Feedback-driven optimization and self-evolution. Feedback-driven methods introduce external critics [26, 51] or self-reflection [22] to refine reasoning processes, and self-evolution systems further form closed-loop training cycles using synthetic supervision [7, 20, 39]. While these approaches can improve performance, they may suffer from cognitive drift and hallucinated reasoning paths when the training signal is dominated by low-quality self-synthetic data. Different from purely synthetic self-evolution, our Anchor Evolution uses Truth Anchor Expansion to retrieve verified truth anchors from ground-truth databases for faithful data curation, and employs the Scaffold-Stripping Mechanism (scaffold-augmented supervision followed by RL to strip the scaffold template) to internalize reasoning capability.

3 Method

In this section, we introduce Anchor Evolution (AnE), an iterative post-training framework designed to achieve faithful and sustained performance gains at the reasoning frontier. As illustrated in Fig. 2, the AnE process comprises three

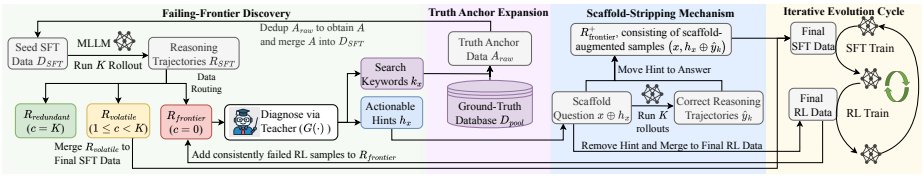


Fig. 2: Overview of Anchor Evolution (AnE). AnE consists of three stages: **Failing-Frontier Discovery**, where we run rollouts and use a teacher for diagnosis; **Truth Anchor Expansion**, where we retrieve verified truth anchors from ground-truth databases; and the **Scaffold-Stripping Mechanism**, where we perform scaffold-augmented SFT and then use RL to strip the scaffold template. These stages are repeated in an **Iterative Evolution Cycle** to push the model's reasoning frontier.

stages: (i) Failing-Frontier Discovery (Sec. 3.1), which performs multiple rollouts of the model under training to probe the capability landscape. A teacher model then provides diagnostic analysis, transforming the reasoning trajectories of hard failure samples into search keywords and actionable hints; (ii) Truth Anchor Expansion (Sec. 3.2), which retrieves high-fidelity anchor samples from a ground-truth database using the generated search keywords for faithful data curation; (iii) Scaffold-Stripping Mechanism (Sec. 3.3), which leverages the actionable hints to construct scaffold-augmented data for SFT, and subsequently applies RL to remove the scaffold templates, thereby transitioning the reasoning trajectory into intrinsic model capabilities. These three stages form an Iterative Evolution Cycle that is repeated to progressively evolve the model (Sec. 3.4).

3.1 Failing-Frontier Discovery

Preliminary RL for Initial Reasoning Ability. To initiate the Anchor Evolution cycle, a preliminary RL phase is conducted to establish the model's initial reasoning state, denoted as $\mathcal{M}_{\text{PreRL}}$, using a seed RL dataset $\mathcal{D}_{\text{RL}}$. This stage serves as a warm-up rather than the core contribution of AnE. By incentivizing the activation of existing knowledge through preliminary RL, the model's latent reasoning capacity is more fully exploited [34, 37]. Consequently, the model is pushed toward its intrinsic reasoning ceiling, thereby reducing the risk that subsequent failure discovery is confounded by superficial instruction-following weaknesses or elicitation artifacts.

Data Routing. To assess the model's current performance, we perform K stochastic rollouts for each sample x in the seed SFT dataset $\mathcal{D}_{\text{SFT}}$, using $\mathcal{M}_{\text{PreRL}}$ as a diagnostic probe. This procedure produces a set of reasoning-trajectory groups $\mathcal{R}_{\text{SFT}} = \{r_x \mid x \in \mathcal{D}_{\text{SFT}}\}$, where each group $r_x = \{\hat{y}_{x,1}, \ldots, \hat{y}_{x,K}\}$ contains the K stochastic rollout trajectories generated for sample x. Each prediction $\hat{y}_k$ is evaluated against the ground-truth label y using an indicator function $\mathbb{I}(\hat{y}, y) \in \{0, 1\}$, which returns 1 if the prediction is correct and 0 otherwise. The model's performance on sample x is then quantified by the

success count:

$$c(x) = \sum_{k=1}^{K} \mathbb{I}(\hat{y}_k, y).$$

Based on the success count $c(x)$, we partition R_{SFT} into three mutually exclusive subsets: (i) Redundant set ($R_{\text{redundant}}$). Samples with $c = K$, representing fully redundant signals, are categorized as $R_{\text{redundant}}$ and pruned from the evolution cycle to eliminate unnecessary computational overhead. (ii) Volatile set (R_{volatile}). Samples with $1 \leq c < K$, which reflect emerging or unstable capabilities, are grouped into R_{volatile}. These samples are retained in the training set to preserve optimization stability and prevent premature forgetting. (iii) Failing frontier (R_{frontier}): Samples with $c = 0$, representing a complete failure in the seed SFT data, constitute the failing frontier. This frontier plays a key role in Anchor Evolution, as these samples highlight critical gaps in the model's capabilities. In addition, we include in R_{frontier} the samples from the preliminary RL phase that the model continues to fail consistently. These samples, like those with $c = 0$ in the seed SFT data, fall outside the model's exploration capabilities and cannot be resolved through standard RL methods. This strategy ensures that the failing frontier captures a more comprehensive view of the model's most persistent and fundamental logical deficiencies.

Diagnostic Analysis via Teacher Model. For each $r_x \in R_{\text{frontier}}$, we employ a diagnostic function $G(\cdot)$, implemented via a teacher model, to conduct a fine-grained analysis of the corresponding trajectory set r_x. Specifically, $G(\cdot)$ first identifies the root cause of failure by categorizing errors into four distinct dimensions: knowledge, understanding, reasoning, and execution. Guided by this taxonomy, it then derives a metadata tuple as follows:

$$(k_x, h_x) \leftarrow G(x, y_x, \{\hat{y}_1, \ldots, \hat{y}_K\}).$$

The search keywords k_x consist of 3-5 nouns related to x and r_x, which act as semantic probes for retrieving truth anchors (Sec. 3.2) from an external ground-truth database, specifically associated with failure-neighborhood samples. Meanwhile, the actionable hints h_x are 1-3 sentences used to help solve problem x, serving as structured scaffolds that reduce problem difficulty, thereby making the target solution more accessible to the model (Sec. 3.3).

3.2 Truth Anchor Expansion

Once the failing frontier R_{frontier} is identified, a key challenge is how to expand these failure cases effectively. Relying solely on synthetic question generation [15, 32] may introduce unanchored cognitive drift, in which the model's reasoning gradually diverges from objective logical truth. To address this issue, we introduce Truth Anchor Expansion, a retrieval-augmented mechanism that grounds the expansion process in an authentic external database. Specifically, AnE retrieves verifiable ground-truth samples from the database that match the model's identified logical deficiencies. This helps the model learn from failure cases while maintaining factual and logical consistency.

Truth Anchor Retrieval for Failure Frontier. For each failure reasoning trajectory $r_x \in R_{\text{frontier}}$, the associated search keywords k_x serve as semantic probes for targeted retrieval. We define a large-scale ground-truth database, $\mathcal{D}_{\text{pool}}$, which contains verified samples spanning diverse domains. Based on the identified failure, the anchoring process selects a set of *truth anchors* $a_x \subset \mathcal{D}_{\text{pool}}$ that lie in its semantic and logical neighborhood. A pre-trained embedding model $\phi(\cdot)$ projects both the retrieval query $q_x = \text{Prompt}(k_x)$ and candidate samples into a shared latent representation space. Truth-anchor retrieval then reduces to a similarity-based ranking problem:

$$a_x = \left\{ x' \in \mathcal{D}_{\text{pool}} \;\middle|\; \text{rank}\big(\text{sim}\left(\phi(q_x), \phi(x')\right)\big) \leq N \right\},$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and N specifies the top-N retrieval threshold. The resulting anchor data is denoted as $\mathcal{A}_{\text{raw}} = \{a_x\}$. This procedure ensures that retrieved samples directly address the identified deficiencies while preserving strict alignment with verified ground-truth knowledge.

Merging Anchor Data with the Seed SFT Dataset. To ensure that the augmented anchor data provides sufficient learning signals, the raw anchor set, $\mathcal{A}_{\text{raw}}$, is first deduplicated to form $\mathcal{A}$. Subsequently, $\mathcal{A}$ is integrated into the fine-tuning dataset, $\mathcal{D}_{\text{SFT}}$, and undergoes K stochastic rollouts and data partitioning as described in Sec. 3.1. Redundant data is discarded, and only the volatile data, denoted as R_{volatile}, is retained in the training set. The resulting frontier, R_{frontier}, is then directed into the process outlined in Sec. 3.3.

3.3 Scaffold-Stripping Mechanism

After Failing-Frontier Discovery and Truth Anchor Expansion, we obtain R_{volatile} and R_{frontier}. R_{volatile} consists of partial samples that represent reasoning trajectories successfully generated by the model, which can be directly utilized for SFT training. In contrast, R_{frontier} still lacks correct reasoning trajectories. While current methods, such as teacher distillation or the use of reasoning trajectory data annotated by human experts, can be employed, these datasets often differ significantly from the model's current capabilities. As a result, directly applying them for SFT tends to lead to superficial imitation rather than enabling the model to truly master and generate high-quality reasoning abilities. To address this, we leverage the actionable hint from Sec. 3.1 to construct Scaffold-Augmented Data for R_{frontier}, which is then used for SFT training. Subsequently, the hints from R_{frontier} are removed and integrated into the RL dataset to effectively transform the assisted reasoning paths into intrinsic model capabilities.

Scaffold-Augmented SFT. To bridge the gap between the failure frontier and model ability, we leverage actionable hints h_x to elicit successful reasoning trajectories from the model's own distribution. Specifically, for each x, we first construct a guided prompt by appending the hint to the question:

$$x_{\text{scaffold}} = x \oplus h_x.$$

We then execute K stochastic rollouts. For every successful trajectory $\hat{y}_k$ that reaches the correct ground truth (i.e., where $1 \leq c \leq K$ under guidance), we relocate the scaffold h_x from the input to the prefix of the output. This transformation yields the following training tuple:

$$(x, h_x \oplus \hat{y}_k).$$

This process transforms R_{frontier} into R^{+}_{frontier} that contains the correct reasoning trajectories. Finally, we extract the correct trajectories from both R^{+}_{frontier} and R_{volatile}, using them for Final SFT training.

AnE utilizes reasoning paths that the model either discovers independently or is guided to discover through teacher hints. This ensures that the model consistently focuses on the boundary of its capabilities, learning from data where its abilities are unstable, thus stabilizing the training process. It provides a solid foundation for the model to eventually master these tasks independently.

Internalization via RL Stripping. While SFT enables the model to solve previously impossible problems through actionable hints from a teacher, the model remains externally dependent. To achieve true capability expansion, we utilize RL as a mechanism for scaffold stripping. During this phase, we add R^{+}_{frontier} to the RL dataset (without the actionable hint in the input question). To ensure the RL dataset contains sufficient learning signals, similar to the seed SFT data in Sec. 3.1, we also perform K rollouts on it and filter out the samples where $c = K$, forming the final RL dataset. Finally, we perform RL training on the final RL dataset. The model is presented only with the original problem x and must independently reach the correct answer y without any hints.

This optimization bridges the gap between hint-aided success and independent mastery, compelling the model to recover the underlying reasoning logic autonomously. As a result, the previously scaffold-dependent capabilities become solidified in the model's parametric weights, successfully advancing the frontier of the model's reasoning ability.

3.4 Iterative Evolution Cycle

The subsequent process of the AnE framework follows a continuous SFT-RL evolution cycle. At the beginning of each new iteration, the framework takes the model from the previous round as its starting point and re-executes the procedures outlined in Sec. 3.1 to Sec. 3.3, excluding the preliminary RL phase. This cycle is repeated to continuously advance the model at the reasoning frontier.

4 Experiments

4.1 Experimental Settings

Implementation Details. The SFT process of the AnE framework is implemented using MindSpeed-MM, while the RL phase is conducted with GSPO [56]

on veRL [38]. For data routing, $K = 4$ stochastic rollouts are performed. The teacher model, Qwen2.5-VL-72B-Instruct [1], is used to conduct diagnostic analysis, while Qwen3-Embedding-8B [55] is employed for embedding search keywords and data in the external database during truth anchor expansion. For truth anchor retrieval, $N = 5$ nearest neighbors are selected. The model curriculum begins with a preliminary RL phase to establish the initial reasoning ability. It then undergoes three complete evolution iterations, each consisting of an SFT stage with scaffolded assistance, followed by an RL stage for scaffold shedding. Training continues until either the SFT loss or the RL mean reward converges. All experiments are performed on clusters of 64 Ascend 910B NPUs, with Qwen2.5-VL-7B-Instruct as the base model. Detailed hyperparameters are provided in the Appendix.

Training datasets. The evolution cycle starts with a seed RL dataset of 80K samples and a seed SFT dataset of 500K samples. The seed SFT dataset is derived from a diverse set of multimodal reasoning and mathematical benchmarks, including LLaVA-CoT [48], OpenVLThinker [10], We-Math2.0 [36], Mulberry [49], MMR1-SFT [21], and MiroMind-M1 [24]. For the seed RL dataset, a reasoning-focused benchmark suite is curated, emphasizing answer verifiability and structured outputs, including AI2D [18], Geometry3k [30], MMEureka [34], ViRL [41], TQA [19], We-Math [35], PuzzleVQA [8], AlgoPuzzleVQA [12], ThinkLiteVL [44], and MMR1-RL [21]. The external database D_{pool} is constructed using FineVision [46], a multimodal dataset comprising 24 million samples. An image deduplication pipeline is employed for FineVision across 66 multimodal datasets from lmms-eval [52] to prevent data leakage.

Baseline and Benchmarks. Performance is evaluated across a suite of multimodal reasoning benchmarks, including MathVista(testmini) [29], MathVision [42], MathVerse(testmini) [54], LogicVista [47], ScienceQA [31], MMMU(val) [50], EMMA(all) [16], and MMSTAR [5]. Top-1 Accuracy (%) is reported using greedy decoding to ensure reproducibility.

4.2 Main Results

Tab. 1 presents a comprehensive comparison between AnE and representative baselines across a wide range of multimodal reasoning benchmarks.

Comparison with Base Model. AnE-1^{st} achieves a significant improvement of 8.3% on Qwen2.5-VL-7B-Instruct, demonstrating that AnE effectively identifies the Failure Frontier in the seed dataset, expands it with targeted samples from the ground-truth dataset, and simplifies problem difficulty through the Scaffold-Stripping mechanism, all while successfully internalizing guidance from the Teacher Model.

Continuous Improvement. AnE-2^{nd} and AnE-3^{rd} show further improvements of 0.6 and 1.4, respectively, compared to the previous rounds. Especially on the more challenging benchmark tasks, MathVision and EMMA, AnE-2^{nd} and AnE-3^{rd} continue to deliver incremental gains. This highlights the method's ability

Table 1: Comparison with related models on multimodal reasoning benchmarks. Accuracy (%) is reported. "–" indicates that the result is neither reported in the original paper nor reproducible due to unavailable checkpoints. † denotes results reproduced with our evaluation pipeline. Preliminary RL serves as the warm-up phase of AnE. The best results are shown in **bold**, and the second-best results are underlined.

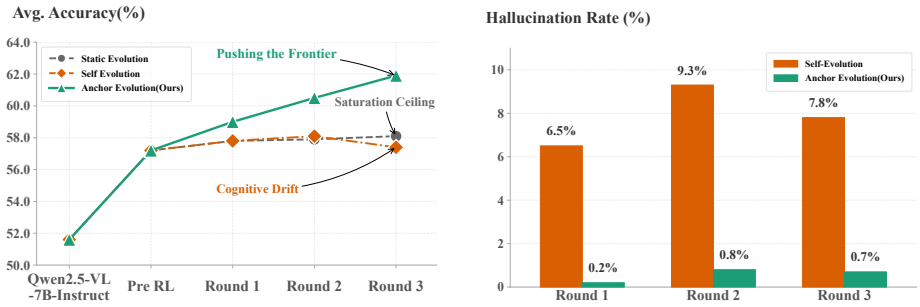
Method	MathVista	MathVision	MathVerse	LogicVista	SciQA	MMMU	EMMA	MMSTAR	Avg
Qwen2.5-VL-7B-Instruct [1]	68.2	25.1	49.2	39.3†	85.4	58.6	24.6	62.5	51.6†
Vision-R1 [17]	73.5†	22.8	52.4†	46.0†	91.4†	52.3†	23.6†	65.8†	53.5†
MMR1 [21]	72.0	31.8	55.4	48.8	92.7†	53.4†	18.6†	65.3†	54.8†
C2-Evo [7]	68.7	–	–	–	–	–	–	–	–
Revisual-R1 [6]	73.1	**48.8**	53.6	**52.3**	91.9†	57.2†	19.5†	57.1†	56.7†
Metis-RISE [37]	75.8	28.7	51.0	49.7	93.4†	55.8†	28.5†	66.6†	56.2†
SRPO [40]	75.8	32.9	55.8	–	–	57.1	29.6	–	–
VL-Rethinker [41]	74.9	32.3	54.2	40.9†	89.6	56.7	29.7	63.6†	55.2†
SPARK-VL-7B [27]	75.9	31.1	53.0	50.0	90.8	**58.7**	28.5†	67.3	56.9†
LLaVA-Critic-R1 [43]	74.0	30.6	49.7	39.5†	84.2†	55.2	28.3	65.1	53.3†
OpenVLThinker [10]	72.3	25.9	47.9	43.5†	88.9†	53.3†	26.8	61.6†	52.5†
OpenMMReasoner-7B [53]	79.5	43.6	**63.8**	50.0	**93.7**†	57.8	24.5†	**70.0**†	60.4†
ADHint [51]	74.4	–	60.6	48.7	–	56.4	–	–	–
Preliminary RL	77.9	32.8	53.6	45.5	92.6	**57.7**	30.1	67.2	57.2
AnE-1^{st}	79.6	39.7	60.5	49.1	93.5	57.7	30.5	68.3	59.9
AnE-2^{nd}	80.1	40.9	60.9	49.8	93.6	57.0	33.0	68.9	60.5
AnE-3^{rd}	**81.2**	43.9	62.3	51.3	93.5	58.1	**34.6**	69.9	**61.9**

to consistently identify weaknesses in the model's capabilities and drive progress at the forefront of reasoning.

Comparison with Related Methods. To ensure a fair comparison, we primarily benchmark against recent multimodal reasoning methods built upon the same 7B-parameter base model. First, compared to static data methods, such as one-pass SFT+RL approaches (e.g., OpenMMReasoner [53] and Vision-R1 [17]) and multi-round methods with static data (e.g., OpenVLThinker), our method demonstrates superior performance, highlighting that dynamic data expansion is a more effective strategy for enhancing model capabilities. In addition, AnE outperforms synthetic data methods like C2-Evo [7], further proving its greater effectiveness. Finally, it surpasses other reflective approaches, including SRPO [40] and ADHint [51], showcasing that our Scaffold-Augmented Data and Internalization through Scaffold Stripping are highly effective in helping the model explore its limits and internalize guidance from external teachers. Overall, AnE-3^{rd} achieves state-of-the-art average performance across eight benchmarks.

4.3 Comparison with Prior Post-Training Paradigms

In this section, we conduct a fair comparison between Anchor Evolution and two commonly used iterative post-training strategies: Static Evolution and Self-Evolution. For Static Evolution, we use the same seed SFT and RL datasets as in AnE and perform alternating SFT and RL training for three rounds, without any data augmentation or modification. For Self-Evolution, we use Qwen2.5-VL-72B-Instruct as the teacher model to synthesize new questions and reasoning paths



(a) Evolution strategies (b) Hallucination rate

Fig. 3: Comparison with prior evolution paradigms. (a) Static Evolution quickly saturates, Self-Evolution suffers performance degradation in later rounds, while Anchor Evolution consistently improves across rounds. **(b)** Hallucination rates of expanded training data across different evolution rounds. Anchor Evolution produces substantially fewer hallucinated samples than Self-Evolution.

from the same seed SFT dataset as AnE for constructing SFT training data. The prompts follow the settings in [15, 32]. For fair comparison, RL training in both Static Evolution and Self-Evolution consistently uses the same static seed RL dataset.

As shown in Fig. 3(a), Static Evolution quickly reaches a performance plateau across multiple rounds. In contrast, Self-Evolution achieves larger improvements than Static Evolution in Round 1 and Round 2, suggesting that synthetic data expansion can enlarge the model's capability boundary to some extent compared with purely static training. However, a noticeable performance drop appears in Round 3, indicating that self-evolution may introduce hallucinated or low-quality synthetic data during iterative generation, which in turn affects subsequent training. To verify this hypothesis, we sample 1K expanded samples from each round and use GPT-4o to identify hallucinated reasoning or incorrect ground-truth answers. As shown in Fig. 3(b), Self-Evolution has consistently higher hallucination rates than Anchor Evolution across three rounds (6.5/9.3/7.8% vs. 0.2/0.8/0.7%). This confirms that synthetic expansion is more prone to noisy supervision and cognitive drift, whereas AnE's truth-anchored filtering yields more faithful data for stable evolution.

Overall, Anchor Evolution consistently improves performance across rounds and achieves better results than both Static Evolution and Self-Evolution, demonstrating its ability to provide stable and reliable capability expansion during iterative post-training.

4.4 Ablation Study

In this section, we perform ablation studies to assess the key design choices in AnE, including the synergy between Truth Anchor Expansion and Scaffolded Reasoning Supervision, the effectiveness and scaling behavior of Truth Anchor

Table 2: Component ablation analysis. Results show that both Truth Anchor Expansion and the Scaffold-Stripping Mechanism are effective on their own, and combining them in AnE yields the best overall performance.

Setting	MathVista	MathVision	MathVerse	LogicVista	SciQA	MMMU	EMMA	MMSTAR	Avg
Preliminary RL	77.9	32.8	53.6	45.5	92.6	**57.7**	30.1	67.2	57.2
↪ w/ Truth Anchor Expansion	78.3	36.4	56.3	48.9	**93.5**	57.3	30.2	67.9	58.6
↪ w/ Scaffold-Stripping Mechanism	77.6	35.2	58.6	44.2	92.9	56.0	**30.6**	**68.4**	57.9
AnE-1st(Ours)	**79.6**	**39.7**	**60.5**	**49.1**	**93.5**	**57.7**	30.5	68.3	**59.9**

Expansion, and the impact of the Scaffold-Stripping Mechanism. To isolate the effects of the proposed AnE components, we use Preliminary RL as the baseline in all ablation studies. Preliminary RL only provides a warm-up initialization and is shared by all variants; therefore, the reported gains reflect the contributions of Truth Anchor Expansion, Scaffold-Stripping, and their iterative coupling.

Component of Truth Anchor Expansion and Scaffolding. As shown in Tab. 2, starting from the Preliminary RL baseline (57.2% avg), incorporating Truth Anchor Expansion alone results in a significant improvement of +1.4%, reaching 58.6%. This demonstrates that our Truth Anchoring method effectively expands true data at the Failure Frontier without introducing hallucinations, thereby boosting the model's capabilities. Similarly, adding only Anchored Scaffolding improves the baseline to 57.9% (+0.7% average), confirming that temporary cognitive guidance effectively bridges the capability gap, allowing external knowledge from the teacher model to be internalized into the model's own reasoning abilities. Most importantly, the full AnE-1st (59.9%) surpasses the individual gains of either component, achieving a total uplift of +2.7% over the baseline. This synergistic effect highlights that Truth Anchor Expansion and Scaffolding can work together effectively, pushing the boundaries of the model's reasoning capabilities.

Effectiveness of Truth Anchor Expansion. To verify the necessity of Truth Anchor Expansion in objective reality, we compare our Truth Anchor Expansion with random and synthetic alternatives (Tab. 3). We compare Random Expansion and Synthetic Expansion, where Random Expansion randomly samples data from the external pool, and Synthetic Expansion generates data using Qwen2.5-VL-72B. All methods use the same data volume for a fair comparison. As shown in Tab. 3, compared to No Expansion, Random Expansion significantly reduces performance by 3.8%, indicating that blind data augmentation may introduce a large amount of irrelevant data, which harms the reasoning abilities of multimodal models. Compared to No Expansion, Synthetic Expansion yields comparable but slightly lower performance, suggesting that purely synthetic data may introduce noisy or hallucinated samples that offset its potential benefits. In contrast, Truth Anchor Expansion retrieves real samples around the failing frontier and consistently improves multimodal reasoning performance across most metrics.

Table 3: Analysis of Expansion Strategies, demonstrating that our Truth Anchor Expansion achieves faithful data curation to enhance model reasoning capabilities.

Setting	MathVista	MathVision	MathVerse	LogicVista	SciQA	MMMU	EMMA	MMSTAR	Avg
Preliminary RL	77.9	32.8	53.6	45.5	92.6	57.7	30.1	67.2	57.2
No Expansion	77.6	35.2	58.6	44.2	92.9	56.0	**30.6**	**68.4**	57.9
Random Expansion	75.4	29.6	49.3	42.2	90.8	52.8	26.2	66.6	54.1
Synthetic Expansion(Self-evolution)	78.0	34.7	56.2	46.7	93.3	55.1	30.4	67.9	57.8
Truth Anchor Expansion(Ours)	**79.6**	**39.7**	**60.5**	**49.1**	**93.5**	**57.7**	30.5	68.3	**59.9**

Data Scaling Law of Truth Anchor Expansion. Furthermore, we investigate the data scaling properties of Truth Anchor Expansion. As shown in Fig. 4, we vary the proportion of Truth Anchor Expansion data mixed into the SFT training set to examine the effect of this augmented data. We observe that incorporating only a small fraction of the expanded data results in a slight performance dip, likely due to the distribution shift caused by the limited additional samples. However, as the mixing ratio increases, model performance improves steadily and monotonically, ultimately reaching an average accuracy of 59.9% when the data is fully incorporated. This trend suggests that the data produced by Truth Anchor Expansion is scalable. Its benefits grow with volume, indicating that the retrieved samples offer genuinely diverse and informative training signals, rather than introducing redundant information.

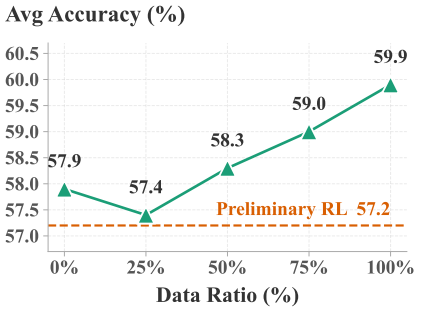


Fig. 4: Data scaling law of Truth Anchor Expansion. Average accuracy on eight benchmarks under different mixing ratios of Truth Anchor Expansion data. Performance improves as more anchor data is incorporated.

Ablation Study on Scaffold-Stripping. As shown in Tab. 4, we perform an ablation study to assess the impact of different Scaffold-Stripping components. In the second row, directly distilling using the Teacher model's data results in a 2.4% performance drop compared to using our Scaffold-Stripping method. This suggests that the reasoning trajectories from the Teacher model are difficult to internalize, causing the model to perform only shallow imitation instead of genuinely improving its reasoning ability. In the fourth row, when the RL Stripping phase is removed and only the Scaffold-Augmented SFT phase is performed, there is a 1.5% performance decrease. This indicates that the model cannot fully internalize the Teacher's guidance through the Scaffold-Augmented SFT phase alone. Finally, as shown in the fifth row, if RL data does not include the R^{+}_{frontier} data from the SFT phase and RL is applied directly, the performance drops by 2.0%. This highlights that the improvement in RL performance depends significantly on internalizing the teacher's guidance from R^{+}_{frontier}.

Table 4: Impact of the Scaffold-Stripping Mechanism. Directly distilling from the teacher model yields only a marginal improvement over the baseline. In contrast, removing the RL stripping stage, or running RL without scaffold-augmented data (i.e., without the hint-based scaffold), both cause a significant drop in AnE performance.

Setting	MathVista	MathVision	MathVerse	LogicVista	SciQA	MMMU	EMMA	MMSTAR	Avg
Preliminary RL	77.9	32.8	53.6	45.5	92.6	57.7	30.1	67.2	57.2
Teacher Distillation	77.8	34.7	51.3	47.1	92.9	57.0	**31.4**	67.8	57.5
Scaffold-Stripping(Ours)	**79.6**	**39.7**	**60.5**	**49.1**	93.5	57.7	30.5	**68.3**	**59.9**
↪ w/o RL Stripping	78.4	35.6	56.0	48.0	**93.6**	**57.8**	30.2	67.9	58.4
↪ w/o R^{+}_{frontier} in RL	78.1	34.6	56.3	47.1	92.9	55.3	31.2	67.8	57.9

4.5 Case Study

Fig. 5 shows a geometry example to contrast AnE with prior self-evolution methods. In Fig. 5(a), teacher-generated rewrites and reasoning trajectories may contain hallucinated content (e.g., options missing the correct answer) and produce complex and hallucination-prone paths that are misaligned with the student model's capability distribution, which can lead to superficial imitation and accumulated drift after training. In Fig. 5(b), AnE first collects the model's failed trajectories via rollouts to locate the failing frontier. The teacher then extracts failure-relevant keywords and actionable hints. The keywords trigger Truth Anchor Expansion to retrieve verified truth anchors from ground-truth databases, yielding valid, high-fidelity problems that target the model's weakness without hallucination. The hints are used for scaffold-augmented supervision, and the Scaffold-Stripping Mechanism further applies RL to strip the scaffold template, enabling the model to internalize the reasoning capability. Overall, compared with prior self-evolution that relies on synthetic data and may accumulate hallucinated reasoning paths and cognitive drift, AnE anchors data evolution with verified truth anchors (Truth Anchor Expansion) and internalizes reasoning via the Scaffold-Stripping Mechanism, enabling stable improvements targeted at the failing frontier.

5 Conclusion

In this paper, we presented *Anchor Evolution* (AnE), a post-training paradigm that integrates truth-anchored data curation and model evolution to achieve faithful and sustained gains at the multimodal reasoning frontier. We showed that static post-training with static datasets is limited by non-extensible data distributions, while self-evolution based on synthetic data can suffer from cognitive drift and hallucinated reasoning paths. To address these limitations, AnE identifies the model's failing frontier via trajectory rollouts and performs Truth Anchor Expansion by retrieving verified truth anchors from ground-truth databases, enabling high-fidelity data curation grounded in real-world knowledge. Moreover, the proposed Scaffold-Stripping Mechanism first anchors reasoning paths through scaffold-augmented supervision and then leverages RL to

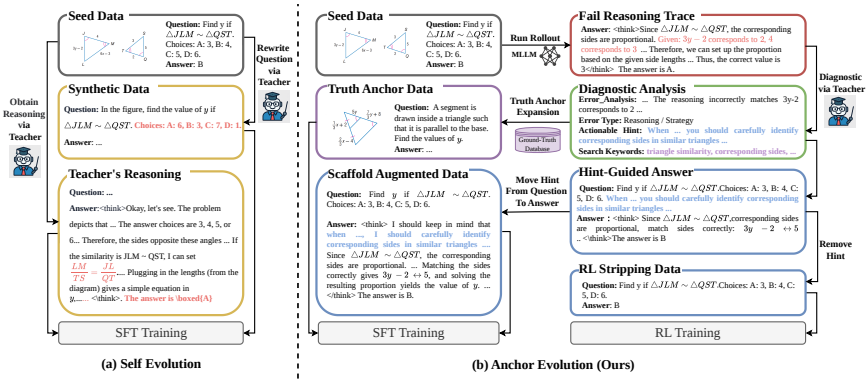


Fig. 5: Illustrative Examples. (a) Prior self-evolution methods may synthesize low-quality data and generate overly complex, hallucination-prone reasoning that exceeds the model's current capabilities, limiting effective learning. **(b)** By analyzing failed reasoning, Anchor Evolution retrieves real data using search keywords and guides the model with actionable hints, helping it improve reasoning and develop independent problem-solving skills.

strip the scaffold template, effectively internalizing reasoning capabilities. Extensive experiments across multiple multimodal reasoning benchmarks demonstrate consistent improvements over the base model and state-of-the-art performance on challenging datasets, validating the effectiveness of AnE for faithful model evolution.

Limitations.

AnE currently leverages verified external resources and an offline data construction pipeline to expand the failing frontier. While the method remains robust under moderate variations in external database scale and coverage, broader seed data and external database collections, together with more efficient trajectory rollouts, teacher diagnosis, and retrieval, could further support the large-scale deployment of AnE. Our evaluation already covers diverse reasoning benchmarks. Extending the evaluation to additional model families and domains would further characterize AnE's generality.

Acknowledgements

This work was supported by the National Key R&D Program of China under Grant No. 2023YFA1008500 and the National Natural Science Foundation of China under Grant No. 62371164.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Chen, H., Tu, H., Wang, F., Liu, H., Tang, X., Du, X., Zhou, Y., Xie, C.: Sft or rl? an early investigation into training r1-like reasoning large vision-language models. arXiv preprint arXiv:2504.11468 (2025)
3. Chen, J., Liu, F., Liu, N., Luo, Y., Qin, E., Zheng, H., Dong, T., Zhu, H., Meng, Y., Wang, X.: Step-wise adaptive integration of supervised fine-tuning and reinforcement learning for task-specific llms. arXiv preprint arXiv:2505.13026 (2025)
4. Chen, J., Yu, T., Bai, H., Yao, L., Wu, J., Li, K., Mi, F., Tao, C., Zhu, L., Zhang, M., et al.: The synergy dilemma of long-cot sft and rl: Investigating post-training techniques for reasoning vlms. arXiv preprint arXiv:2507.07562 (2025)
5. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., et al.: Are we on the right way for evaluating large vision-language models? Advances in Neural Information Processing Systems **37**, 27056–27087 (2024)
6. Chen, S., Guo, Y., Su, Z., Li, Y., Wu, Y., Chen, J., Chen, J., Wang, W., Qu, X., Cheng, Y.: Advancing multimodal reasoning: From optimized cold start to staged reinforcement learning. arXiv preprint arXiv:2506.04207 (2025)
7. Chen, X., Hu, W., Li, H., Zhou, J., Chen, Z., Cao, M., Zeng, Y., Zhang, K., Yuan, Y.J., Han, J., et al.: C2-evo: Co-evolving multimodal data and model for self-improving reasoning. arXiv preprint arXiv:2507.16518 (2025)
8. Chia, Y.K., Toh, V., Ghosal, D., Bing, L., Poria, S.: Puzzlevqa: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 16259–16273 (2024)
9. Chu, T., Zhai, Y., Yang, J., Tong, S., Xie, S., Schuurmans, D., Le, Q.V., Levine, S., Ma, Y.: Sft memorizes, rl generalizes: A comparative study of foundation model post-training. arXiv preprint arXiv:2501.17161 (2025)
10. Deng, Y., Bansal, H., Yin, F., Peng, N., Wang, W., Chang, K.W.: Openvlthinker: Complex vision-language reasoning via iterative sft-rl cycles. arXiv preprint arXiv:2503.17352 (2025)
11. Gandhi, K., Chakravarthy, A., Singh, A., Lile, N., Goodman, N.D.: Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. arXiv preprint arXiv:2503.01307 (2025)
12. Ghosal, D., Han, V.T.Y., Ken, C.Y., Poria, S.: Are language models puzzle prodigies? algorithmic puzzles unveil serious challenges in multimodal reasoning. arXiv preprint arXiv:2403.03864 (2024)
13. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
14. Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., et al.: Seed1.5-vl technical report. arXiv preprint arXiv:2505.07062 (2025)
15. Guo, J., Zheng, T., Li, Y., Bai, Y., Li, B., Wang, Y., Zhu, K., Neubig, G., Chen, W., Yue, X.: Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 13869–13920 (2025)

16. Hao, Y., Gu, J., Wang, H.W., Li, L., Yang, Z., Wang, L., Cheng, Y.: Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. arXiv preprint arXiv:2501.05444 (2025)
17. Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Xu, Z., Hu, Y., Lin, S.: Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749 (2025)
18. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: European conference on computer vision. pp. 235–251. Springer (2016)
19. Kembhavi, A., Seo, M., Schwenk, D., Choi, J., Farhadi, A., Hajishirzi, H.: Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In: Proceedings of the IEEE Conference on Computer Vision and Pattern recognition. pp. 4999–5007 (2017)
20. Koh, W., Oh, W., Jang, J., Lee, M., Kim, H., Kim, A.Y., Kim, J., Lee, J., Kim, T., Yun, S.Y.: Adastar: Adaptive data sampling for training self-taught reasoners. arXiv preprint arXiv:2505.16322 (2025)
21. Leng, S., Wang, J., Li, J., Zhang, H., Hu, Z., Zhang, B., Jiang, Y., Zhang, H., Li, X., Bing, L., et al.: Mmr1: Enhancing multimodal reasoning with variance-aware sampling and open resources. arXiv preprint arXiv:2509.21268 (2025)
22. Li, J., Dong, X., Liu, Y., Yang, Z., Wang, Q., Wang, X., Zhu, S.C., Jia, Z., Zheng, Z.: Reflectevo: Improving meta introspection of small llms by learning self-reflection. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 16948–16966 (2025)
23. Li, J., Chen, J., Qu, Y., Xu, S., Lin, Z., Zhu, J., Xu, B., Tan, W., Fu, P., Ju, J., et al.: Xiaomi mimo-vl-miloco technical report. arXiv preprint arXiv:2512.17436 (2025)
24. Li, X., Xiao, Y., Ng, D., Ye, H., Deng, Y., Lin, X., Wang, B., Mo, Z., Zhang, C., Zhang, Y., et al.: Miromind-m1: An open-source advancement in mathematical reasoning via context-aware multi-stage policy optimization. arXiv preprint arXiv:2507.14683 (2025)
25. Li, Y., Liu, Z., Li, Z., Zhang, X., Xu, Z., Chen, X., Shi, H., Jiang, S., Wang, X., Wang, J., et al.: Perception, reason, think, and plan: A survey on large multimodal reasoning models. arXiv preprint arXiv:2505.04921 (2025)
26. Liu, W., Wu, H., Kuang, Y., Han, X., Zhong, T., Feng, J., Lu, W.: Automated optimization modeling via a localizable error-driven perspective. arXiv preprint arXiv:2602.11164 (2026)
27. Liu, Z., Zang, Y., Ding, S., Cao, Y., Dong, X., Duan, H., Lin, D., Wang, J.: Spark: Synergistic policy and reward co-evolving framework. arXiv preprint arXiv:2509.22624 (2025)
28. Lu, A., Feng, T., Yuan, H., Li, W., Sun, Y.: Why does rl generalize better than sft? a data-centric perspective on vlm post-training. arXiv preprint arXiv:2602.10815 (2026)
29. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255 (2023)
30. Lu, P., Gong, R., Jiang, S., Qiu, L., Huang, S., Liang, X., Zhu, S.C.: Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 6774–6786 (2021)

31. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in neural information processing systems **35**, 2507–2521 (2022)
32. Luo, R., Zhang, H., Chen, L., Lin, T.E., Liu, X., Wu, Y., Yang, M., Li, Y., Wang, M., Zeng, P., et al.: Mmevol: Empowering multimodal large language models with evol-instruct. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 19655–19682 (2025)
33. Ma, L., Liang, H., Qiang, M., Tang, L., Ma, X., Wong, Z.H., Niu, J., Shen, C., He, R., Li, Y., et al.: Learning what reinforcement learning can't: Interleaved online fine-tuning for hardest questions. arXiv preprint arXiv:2506.07527 (2025)
34. Meng, F., Du, L., Liu, Z., Zhou, Z., Lu, Q., Fu, D., Han, T., Shi, B., Wang, W., He, J., et al.: Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. arXiv preprint arXiv:2503.07365 (2025)
35. Qiao, R., Tan, Q., Dong, G., MinhuiWu, M., Sun, C., Song, X., Wang, J., Gongque, Z., Lei, S., Zhang, Y., et al.: We-math: Does your large multimodal model achieve human-like mathematical reasoning? In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 20023–20070 (2025)
36. Qiao, R., Tan, Q., Yang, P., Wang, Y., Wang, X., Wan, E., Zhou, S., Dong, G., Zeng, Y., Xu, Y., et al.: We-math 2.0: A versatile mathbook system for incentivizing visual mathematical reasoning. arXiv preprint arXiv:2508.10433 (2025)
37. Qiu, H., Lan, X., Liu, F., Sun, X., Ruan, D., Shi, P., Ma, L.: Metis-rise: Rl incentivizes and sft enhances multimodal reasoning model learning. arXiv preprint arXiv:2506.13056 (2025)
38. Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., Wu, C.: Hybridflow: A flexible and efficient rlhf framework. arXiv preprint arXiv: 2409.19256 (2024)
39. Thawakar, O., Venkatraman, S., Thawkar, R., Shaker, A., Cholakkal, H., Anwer, R.M., Khan, S., Khan, F.: Evolmm: Self-evolving large multimodal models with continuous rewards. arXiv preprint arXiv:2511.16672 (2025)
40. Wan, Z., Dou, Z., Liu, C., Zhang, Y., Cui, D., Zhao, Q., Shen, H., Xiong, J., Xin, Y., Jiang, Y., et al.: Srpo: Enhancing multimodal llm reasoning via reflection-aware reinforcement learning. arXiv preprint arXiv:2506.01713 (2025)
41. Wang, H., Qu, C., Huang, Z., Chu, W., Lin, F., Chen, W.: Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. arXiv preprint arXiv:2504.08837 (2025)
42. Wang, K., Pan, J., Shi, W., Lu, Z., Ren, H., Zhou, A., Zhan, M., Li, H.: Measuring multimodal mathematical reasoning with math-vision dataset. Advances in Neural Information Processing Systems **37**, 95095–95169 (2024)
43. Wang, X., Li, C., Yang, J., Zhang, K., Liu, B., Xiong, T., Huang, F.: Llava-critic-r1: Your critic model is secretly a strong policy model. arXiv preprint arXiv:2509.00676 (2025)
44. Wang, X., Yang, Z., Feng, C., Lu, H., Li, L., Lin, C.C., Lin, K., Huang, F., Wang, L.: Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. arXiv preprint arXiv:2504.07934 (2025)
45. Wei, Y., Zhao, L., Sun, J., Lin, K., Yin, J., Hu, J., Zhang, Y., Yu, E., Lv, H., Weng, Z., et al.: Open vision reasoner: Transferring linguistic cognitive behavior for visual reasoning. arXiv preprint arXiv:2507.05255 (2025)

46. Wiedmann, L., Zohar, O., Mahla, A., Wang, X., Li, R., Frere, T., von Werra, L., Gosthipaty, A.R., Marafioti, A.: Finevision: Open data is all you need. arXiv preprint arXiv:2510.17269 (2025)
47. Xiao, Y., Sun, E., Liu, T., Wang, W.: Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. arXiv preprint arXiv:2407.04973 (2024)
48. Xu, G., Jin, P., Wu, Z., Li, H., Song, Y., Sun, L., Yuan, L.: Llava-cot: Let vision language models reason step-by-step. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2087–2098 (2025)
49. Yao, H., Huang, J., Wu, W., Zhang, J., Wang, Y., Liu, S., Wang, Y., Song, Y., Feng, H., Shen, L., et al.: Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. arXiv preprint arXiv:2412.18319 (2024)
50. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: 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. pp. 9556–9567 (2024)
51. Zhang, F., Tan, Z., Ma, X., Dong, Z., Leng, X., Zhao, J., Sun, X., Yang, Y.: Adhint: Adaptive hints with difficulty priors for reinforcement learning. arXiv preprint arXiv:2512.13095 (2025)
52. Zhang, K., Li, B., Zhang, P., Pu, F., Cahyono, J.A., Hu, K., Liu, S., Zhang, Y., Yang, J., Li, C., et al.: Lmms-eval: Reality check on the evaluation of large multimodal models. In: Findings of the Association for Computational Linguistics: NAACL 2025. pp. 881–916 (2025)
53. Zhang, K., Wu, K., Yang, Z., Li, B., Hu, K., Wang, B., Liu, Z., Li, X., Bing, L.: Openmmreasoner: Pushing the frontiers for multimodal reasoning with an open and general recipe. arXiv preprint arXiv:2511.16334 (2025)
54. Zhang, R., Jiang, D., Zhang, Y., Lin, H., Guo, Z., Qiu, P., Zhou, A., Lu, P., Chang, K.W., Qiao, Y., et al.: Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In: European Conference on Computer Vision. pp. 169–186. Springer (2024)
55. Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., et al.: Qwen3 embedding: Advancing text embedding and reranking through foundation models. arXiv preprint arXiv:2506.05176 (2025)
56. Zheng, C., Liu, S., Li, M., Chen, X.H., Yu, B., Gao, C., Dang, K., Liu, Y., Men, R., Yang, A., et al.: Group sequence policy optimization. arXiv preprint arXiv:2507.18071 (2025)
57. Zhou, G., Qiu, P., Chen, C., Wang, J., Yang, Z., Xu, J., Qiu, M.: Reinforced mllm: A survey on rl-based reasoning in multimodal large language models. arXiv preprint arXiv:2504.21277 (2025)