Improving Multimodal Reasoning via Worst Dimension Optimization

Haocheng Lv, **Huaping Zhang**, **Qiuchi Li**, **Lei Li** and **Chunxiao Gao***
Beijing Institute of Technology
{3120255822, kevinzhang, liqiuchi,lilei,gao_chunxiao}@bit.edu.cn

Abstract

Multimodal reasoning requires a path that retains integrity over a wide range of constraints, from visual grounding to logic consistency. However, the current Process Reward Models (PRMs) focus on heuristically defined rewards that equally weigh these factors, which may lead to the concealment of individual dimension failures (such as visual hallucinations) by the dominating factors, without guaranteeing the validity of the reasoning process in general. Therefore, to overcome the limitation, the paper proposes the concept of Multimodal Multi-Dimensional Scalarization Process Reward Modeling (MMS-PRM), a paradigm specifically developed to enforce the worst dimension's robustness in multimodal reasoning. Specifically, a hierarchical fine-grained reward space is developed to represent the multimodal risks in the reasoning tasks, and a Chebyshev-based Monte Carlo Tree Search (MCTS) algorithm is introduced, in which the primary focus during the path searching is given to the worst-performing dimension. Moreover, a curriculum-based Direct Preference Optimization (DPO) approach is developed to gradually learn the balanced reasoning skills in the policy. The experimental results show that, without the dimension collapse issue, the MMS-PRM approach significantly improves the reliability of the multimodal reasoning performance and reaches competitive results in various challenging tasks. The code is available at https://github.com/leibniz-Man/MMS-PRM.

1 Introduction

Multimodal Large Language Models (MLLMs) are shown to perform well on complex reasoning tasks such as mathematical diagrams and scientific figures. In contrast to the reasoning task on pure-text reasoning, multimodal reasoning requires satisfying multiple constraints simultaneously, including visual grounding and logical correctness, where violation of any single aspect invalidates the entire reasoning trajectory.

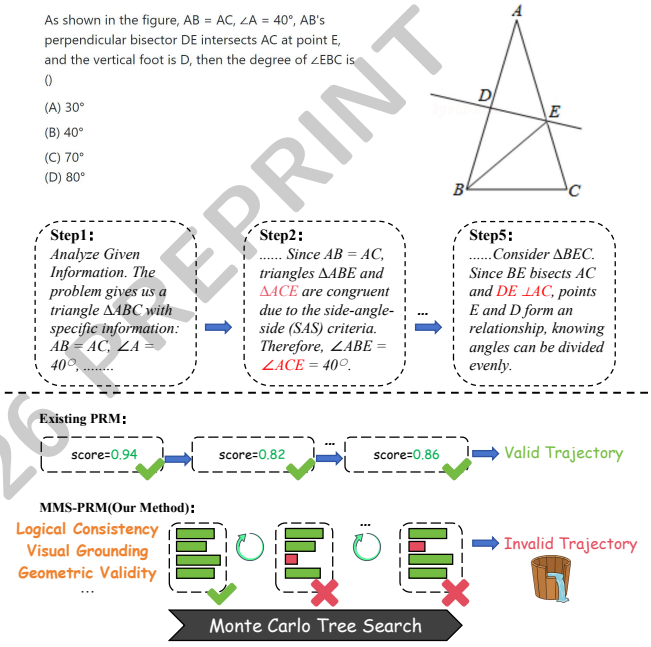

Figure 1: Comparison Between Existing Methods and Our Method.

To supervise such processes, prior work introduces Process Reward Models (PRMs) that provide step-level feedback. However, existing PRMs [Wang *et al.*, 2025; Luo *et al.*, 2025; Ong *et al.*, 2025; Gao *et al.*, 2025] typically collapse multiple quality dimensions into a single scalar reward. This makes it possible for good performance in some factors to compensate bad performance in other factors, which causes incorrect reasoning paths to be reinforced.

The problem caused by the compensation mechanism is even more severe in multimodal reasoning tasks. As shown in Figure 1, the reasoning process can have a mathematically well-organized reasoning chain with the use of the hallucinated visual relations. Because the text logic is well-organized, the reasoning process will normally be assigned with a high confidence score by the scalar PRM, without punishing the reasoning process for the factual error. The failure indicates that there is a severe weakness in the current averaging approach for designing the reward.

Such findings lead to the primary tenet in this paper: *the*

*Corresponding author.

integrity of an active multimodal reasoning trajectory is measured not by average quality but by the worst active dimension. An active reasoning step that is valid but unground in images is inherently invalid and should not be reinforced. Therefore, successful multimodal alignment should progress from optimizing for expectation maximization rewards to non-compensatory goals that punish worst-dimension degeneration.

To operationalize this principle, we propose MMS-PRM, a search-enhanced, multi-dimensional process reward framework for multimodal reasoning. We reformulate multimodal process rewarding as a multi-objective trajectory optimization problem, where the quality of each reasoning step and consequently the whole reasoning trajectory is governed by its weakest relevant dimension rather than an aggregated score. First, we construct a hierarchical, fine-grained reward space that decomposes multimodal reasoning quality into interpretable dimensions and sub-dimensions, allowing rewards to be dynamically activated based on the evolving reasoning context. Second, we introduce a Chebyshev-guided Monte Carlo Tree Search (MCTS) that explicitly prioritizes the worst-performing reward dimension during trajectory exploration, preventing compensation across conflicting criteria and promoting balanced reasoning paths. Finally, we integrate a curriculum-style Direct Preference Optimization (DPO) strategy that progressively trains MLLMs with our obtained balanced reasoning trajectories, from short, high-confidence trajectories to long-horizon multimodal reasoning chains.

Through this closed-loop framework, MMS-PRM enforces balance between visual grounding, logical coherence, and semantic correctness at both step and trajectory levels. Extensive experiments on diverse and challenging multimodal reasoning benchmarks demonstrate that MMS-PRM significantly improves reasoning reliability and robustness, particularly on long-horizon and visually demanding tasks.

The contributions of this work are as follows:

- We identify a fundamental limitation of scalar process rewards in multimodal reasoning and reformulate process reward modeling as a non-compensatory, multi-dimensional trajectory optimization problem.
- We propose a worst-dimension-aware search framework that combines hierarchical process rewards with Chebyshev-scalarized MCTS to enforce balanced multimodal reasoning.
- We introduce a curriculum-based preference alignment strategy that effectively transfers search-discovered reasoning behaviors into the model, achieving strong performance and generalization across benchmarks.

2 Related Work

VLM Reasoning. As VLMs continue to be used for increasingly complex tasks in mathematics and science[Yue *et al.*, 2024], the need for improving their reasoning ability has become crucial. Previous studies aligning visual regions with reasoning steps[Shao *et al.*, 2024], or decomposing long-chain reasoning via multi-agent frameworks[Dong *et al.*, 2025b; Shi *et al.*, 2026]. These advances demonstrate the importance of modeling intermediate reasoning steps[Zhang *et al.*, 2025b]. Nevertheless, the majority of the previous studies use coarse-grained reasoning supervision[Li *et al.*, 2024; Shao *et al.*, 2024]. In contrast, our work introduces a fine-grained, structured reasoning paradigm to more precisely enhance VLM reasoning.

Process Reward Model. With more complex tasks being addressed by LLMs, there is a need to reason over longer trajectories, rendering it inadequate to check coherence and correctness of reasoning using Outcome reward model(ORM) [Snell *et al.*, 2024; Luo *et al.*, 2024; Cai *et al.*, 2025]. Early PRM work focuses on objective domains such as mathematics, where intermediate steps admit clear correctness criteria, either by directly modeling step-wise correctness [Wang *et al.*, 2024a] or estimating its likelihood [Wang *et al.*, 2024d; Guan *et al.*, 2025]. Recent studies extend PRMs to multimodal reasoning. VisualPRM [Wang *et al.*, 2025] aligns intermediate reasoning with visual evidence via step-level rewards, URSA [Luo *et al.*, 2025] integrates process supervision into policy learning through reinforcement learning, while VL-PRM300K [Ong *et al.*, 2025] and SVIP [Gao *et al.*, 2025] enable detailed multimodal supervision with large-scale annotated datasets and visual programming. The current state-of-the-art in multimodal PRM is based on scalarized rewards, which may hiding important dimensions. To address this issue, we introduce the MMS-PRM, which represents multimodal process-level supervision in a multi-dimensional reward space to capture the quality of reasoning more accurately.

Tree-based Search in LLMs. Tree structures have demonstrated significant potential in language models [Qi *et al.*, 2025; Wu *et al.*, 2024]. Recent efforts explore applying these tree search methods to identify effective reasoning paths for MLLMs. AR-MCTS [Dong *et al.*, 2025a] enhances multimodal reasoning by integrating MCTS with active retrieval, but its high computational overhead and extensive iterations limit its practicality. Similarly, Mulberry [Yao *et al.*, 2026] distills 260K long-chain reasoning samples via tree structures from powerful models like GPT-4o, but relies heavily on resource-intensive teacher models.

3 Method

Multimodal reasoning differs from pure-text reasoning in that the model must keep every intermediate step visually grounded, logically coherent, and eventually answer-correct. This requires supervising both single-step quality and the dynamics from steps to the full trajectory. Simple final-answer supervision may overlook intermediate reasoning errors, while only step-level supervision may produce locally sound steps that fail to compose into a globally valid reasoning chain. To cope with this, we build a closed-loop alignment framework consisting of three parts: (1) a hierarchical, fine-grained reward space that models multimodal reasoning quality at the single-step level via multi-dimensional rewards; (2) a reward-guided Chebyshev MCTS that dynamically searches for trajectories balancing all activated reward dimensions; and (3) a curriculum-style DPO that progressively

aligns the MLLM policy from easy/short chains to hard/long chains. The whole pipeline is shown in Figure 2.

3.1 Hierarchical Fine-Grained Reward Space

The first component turns raw model outputs into reusable, structured reward dimensions.

Candidate Criteria Generation. For each instance in the training set, we derive criteria based on the reasoning steps. Given an input x and a reasoning sequence $\hat{y}$ produced by model sampling, we apply an automated analyzer $J(\cdot)$, typically instantiated as a vision-language model, to extract at most 5 relevant criteria:

$$J(x, \hat{y}) = \{c_1, c_2, \dots\}, \tag{1}$$

These criteria assess the reasoning quality, addressing various factors such as visual alignment, semantic correctness, logical consistency, stepwise coherence, and conciseness.

Embedding and Clustering. The criteria gathered are embedded into a d-dimensional vector space:

$$V(c_i) = [v_i^{(1)}, \dots, v_i^{(d)}], \tag{2}$$

where d is the embedding dimensionality. hierarchical clustering is performed to group similar criteria together, halting when the similarity between criteria drops below a certain threshold ξ. This results in a hierarchical reward structure $\mathcal{H} = \{\mathcal{H}_1, \mathcal{H}_2, \dots\}$, where higher levels are broad (e.g., "visually grounded") and lower levels are more specific (e.g., "accurately references bar #3").

Stepwise Reward Allocation. The dynamic process of reward assignment employs a reward tree T to assign stepwise reward signals. For each step $\hat{y}(t)$ in the n-step reasoning trajectory $\hat{y} = \{\hat{y}(1), \hat{y}(2), \dots, \hat{y}(n)\}$, the corresponding reward is allocated dynamically.

Specifically, for the t-th step $\hat{y}(t)$, a selection function first chooses a coarse-grained parent reward node r_{parent} from the top level of the reward tree T, representing the primary risk dimension to be evaluated at this step. Conditioned on $\hat{y}(t)$ and r_{parent}, an analysis function Φ determines whether finer-grained evaluation is required, and if so, generates a set of candidate textual evaluation criteria:

$$\mathcal{C}_t = \Phi(\hat{y}(t), r_{\text{parent}}), \quad \mathcal{C}_t = \{c_t^1, c_t^2, \dots, c_t^{K_t}\}. \tag{3}$$

Each criterion $c_t^i \in \mathcal{C}_t$ is embedded into a d-dimensional semantic space using the embedding function $V(\cdot)$. To filter out criteria that are weakly related or semantically drifting away from the selected parent reward, we compute the cosine distance between each criterion embedding and the parent reward embedding:

$$\delta_i^t = D\big(V(c_t^i), V(r_{\text{parent}})\big), \tag{4}$$

where $D(\cdot)$ denotes cosine distance.

This distance-based filtering serves as a heuristic safeguard rather than a strict semantic entailment test, aiming to remove candidate criteria that are loosely related to the intended risk dimension or shift focus away from the core evaluation objective. Accordingly, the set of activated reward nodes for step t is defined as:

$$R_t = \{r_t^i \mid c_t^i \in \mathcal{C}_t,\ \delta_i^t \le \zeta\}, \tag{5}$$

where each r_t^i denotes the reward tree node associated with criterion c_t^i, and ζ is a predefined distance threshold.

Finally, each activated reward node $r_t^i \in R_t$ is scored using a step-level scoring function $S(\cdot)$:

$$s_i(t) = S\big(\hat{y}(t), r_t^i\big), \tag{6}$$

All step-level reward scores produced by the scoring function $S(\cdot)$ are initially defined on a discrete scale of $[1, 10]$. To ensure consistency across heterogeneous reward dimensions, we linearly normalize each reward dimension into $[0, 1]$ before any trajectory aggregation or scalarization is applied. Specifically, a raw score $s \in [1, 10]$ is mapped to $\tilde{s} = (s - 1)/9$.

3.2 Dynamic Reward-Guided Chebyshev MCTS

The second component uses search to explore trajectories that are simultaneously good across all activated dimensions, where the reward space assigns a multi-dimensional reward vector to each reasoning step rather than a single scalar or final outcome.

Chebyshev scalarization. Given a reward vector $v = (v_1, \dots, v_M)$ and an adaptive ideal point $z^* = (z_1^*, \dots, z_M^*)$, we first consider the standard Chebyshev scalarization, defined as:

$$g_{\text{cheby}}(v; z^*) = -\max_j \big(z_j^* - v_j\big). \tag{7}$$

However, multimodal reasoning tasks often involve noisy reward signals, where a single isolated misclassification by the reward model could excessively penalize a valid trajectory under the strict min-max criterion. To mitigate this noise and enhance robustness, we employ the *Augmented Chebyshev Scalarization*:

$$g_{\text{aug}}(v; z^*) = g_{\text{cheby}}(v; z^*) - \rho \sum_{j=1}^{M} (z_j^* - v_j) \tag{8}$$

where ρ is a small positive constant serving as a regularization term. This formulation preserves 1. focus on the worst dimension due to the chebyshev term; 2. for trajectories that are good on all aspects, the augmented term can capture their subtle differences and identify the best candidate

MCTS Implementation. The MCTS algorithm constructs a tree of reasoning steps, where each node corresponds to a partial reasoning trajectory. In a nutshell, the algorithm dynamically searches for the best trajectory based on Chebyshev scalarized rewards:

- **Selection:** Starting from the root, the algorithm traverses the tree by selecting the child node that maximizes the Upper Confidence Bound applied to Trees (UCT) score:

$$\text{UCT}(n) = g_{\text{aug}}(R_{\text{step}}(n); z^*) + c\sqrt{\frac{\log N(\text{parent}(n))}{N(n)}}, \tag{9}$$

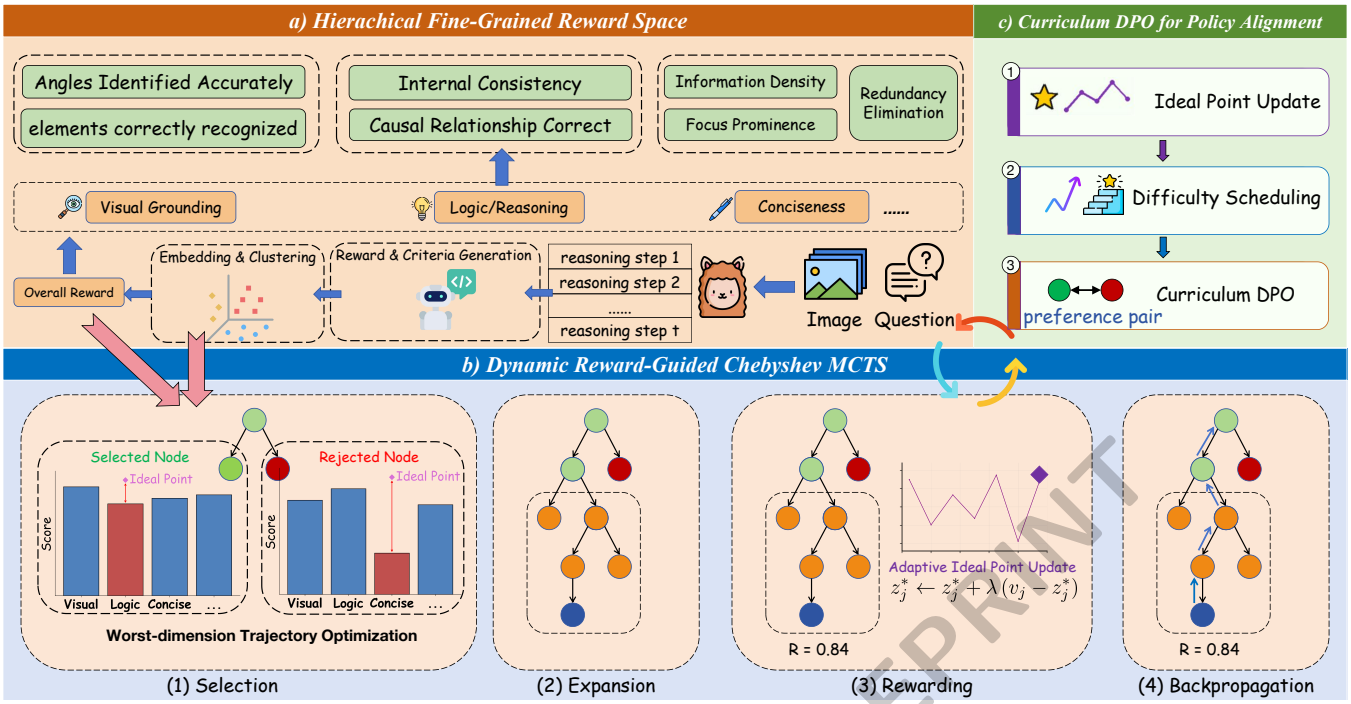

Figure 2: Overview of MMS-PRM. The framework consists of three components: (a) a hierarchical, fine-grained reward space that decomposes multimodal reasoning quality into interpretable dimensions and assigns step-level rewards; (b) a Chebyshev-guided Monte Carlo Tree Search (MCTS) that explores reasoning trajectories by explicitly prioritizing the worst-performing dimension; and (c) a curriculum-style Direct Preference Optimization (DPO) that transfers balanced reasoning behaviors discovered by search into the policy. Together, these components form a closed-loop process that enforces non-compensatory, balanced multimodal reasoning.

where $R_{\text{step}}(n)$ denote the M-dimensional reward vector associated with node n, while the scalarized step quality is always obtained via $g_{\text{aug}}(\cdot)$. c controls the exploration-exploitation trade-off. Here, N(n) represents the number of node n has been visited, while $N(\text{parent}(n))$ denotes the visit count of its parent node. The next node is selected via $n_{\text{next}} = \arg\max_{n' \in \text{children}(n)} \text{UCT}(n')$.

- **Expansion:** Upon reaching a leaf node, we expand the tree by sampling k candidate steps $\{y_1^{(t)}, \ldots, y_k^{(t)}\}$ from the current policy $\pi_\theta(y^{(t)}|x, y^{(<t)})$. Each candidate instantiates a new child node.
- **Evaluation:** For a newly expanded node n, we evaluate its immediate quality. Instead of random rollouts, we compute the fine-grained process reward vector $R_{\text{step}}(n)$. We then apply the Augmented Chebyshev Scalarization $g_{\text{aug}}(\cdot)$ to condense this vector into a scalar, strictly penalizing the worst-performing dimension while retaining discrimination via regularization.
- **Backpropagation:** The scalarized value and visit counts are propagated from the leaf to the root. For every node along the path, $N(n)$ is incremented and statistics are updated, dynamically guiding the search toward paths that are balanced across all reward dimensions.

Adaptive Ideal Point Update. In our framework, the ideal point is updated according to the worst-performing reward dimension observed in the current MCTS simulation. Given the current ideal point $z^* = (z_1^*, \ldots, z_M^*)$ and the reward vector $v = (v_1, \ldots, v_M)$, we compute the dimension-wise gap between the ideal point and the observed reward:

$$\Delta_j = z_j^* - v_j, \quad j = 1, \ldots, M. \tag{10}$$

The worst-performing dimension is selected as the dimension with the largest gap:

$$j^* = \arg \max_{j \in \{1,\ldots,M\}} \Delta_j. \tag{11}$$

The maximum gap is then used as the adaptive signal for the next-round ideal point:

$$\Delta_{\max} = \Delta_{j^*} = \max_{j \in \{1,\ldots,M\}} \left(z_j^* - v_j\right). \tag{12}$$

Finally, we broadcast this worst-dimension gap to construct the next ideal-point vector:

$$z_{\text{next}}^* = \Delta_{\max} \mathbf{1}_M. \tag{13}$$

This update directly uses the largest deviation from the ideal point as the next reference signal, thereby forcing the following MCTS iteration to concentrate on the currently weakest dimension. Unlike coordinate-wise exponential moving average updates, the proposed rule does not allow strong dimensions to dilute the failure of weak ones. Therefore, it better matches the non-compensatory objective of Chebyshev-guided search and strengthens the worst-dimension optimization principle of MMS-PRM.

3.3 Curriculum DPO for Policy Alignment

After applying the MCTS algorithm, we obtain an ideal point z^* together with a set of candidate trajectories for each input sample. The third component then converts these searched trajectories into preference data and aligns the policy using an easy-to-hard curriculum.

Score fusion and pair construction. Our method allows us to evaluate the overall quality of any reasoning trajectory Y via a fused scalar score $G(Y)$. Based on this evaluation, preference pairs (Y^+, Y^-) are constructed from the searched trajectories produced by MCTS, satisfying $G(Y^+) - G(Y^-) \geq \delta$. Giving a reasoning trajectory $Y = \{y^{(1)}, y^{(2)}, \ldots, y^{(n)}\}$, we define the trajectory-level reward vector $\mathbf{R}_{\text{traj}}(Y) \in \mathbb{R}^M$ as:

$$\mathbf{R}_{\text{traj}}(Y) = \frac{1}{n} \sum_{t=1}^{n} \mathbf{R}_{\text{step}}(y^{(t)}) \tag{14}$$

where each $\mathbf{R}_{\text{step}}(y^{(t)})$ represents the M-dimensional reward scores (e.g., logical consistency, visual alignment) at step t. This definition ensures that the trajectory reward maintains the same dimensionality and physical meaning as the step-wise rewards. After K simulations, each trajectory Y has a fused score

$$\bar{G}(Y) = \eta\, g_{\text{aug}}(R_{\text{traj}}(Y)) + (1-\eta) R_{\text{ans}}(Y), \tag{15}$$

where $\eta \in [0, 1]$ is a balancing coefficient. $R_{ans}(n)$ denotes the outcome correctness reward, defined as 1 if the reasoning chain at terminal node n matches the ground truth answer, and 0 otherwise. For non-terminal nodes, $R_{ans}(n)$ is set to 0.

Finally, given multiple reasoning trajectories sampled for the same input x, we induce preference pairs (Y^+, Y^-) for DPO-style optimization, where the trajectory with a higher aggregated score is treated as preferred, i.e.,

$$\bar{G}(Y^+) - \bar{G}(Y^-) \geq \delta. \tag{16}$$

Warm-up and DPO. Given the constructed preference pairs (Y^+, Y^-) derived from the search process, this stage aims to align the policy with step-level reasoning preferences through a two-phase training procedure. Specifically, we first warm up the policy using supervised fine-tuning (SFT) to stabilize generation, and then apply Direct Preference Optimization (DPO) to directly optimize the policy toward preferred trajectories.

$$\mathcal{L}_{\text{SFT}}(\pi_\theta) = -\mathbb{E}_{(x,Y)} \sum_{t=1}^{T} \log \pi_\theta(y^{(t)} \mid x, y^{(<t)}). \tag{17}$$

Then we apply DPO with a reference model π_{ref}:

$$L(\theta) = -\mathbb{E}_{x, \hat{y}^{<t}, (\hat{y}^+, \hat{y}^-) \in V} \left[\log \sigma(F^+ - F^-) \right] \tag{18}$$

where,

$$F^+ = \beta \log \frac{\pi_\theta(\hat{y}^+ | x; \hat{y}^{<t})}{\pi_{\text{ref}}(\hat{y}^+ | x; \hat{y}^{<t})} \tag{19}$$

$$F^- = \beta \log \frac{\pi_\theta(\hat{y}^- | x; \hat{y}^{<t})}{\pi_{\text{ref}}(\hat{y}^- | x; \hat{y}^{<t})} \tag{20}$$

Difficulty scheduling. We define the difficulty of a reasoning trajectory based on two intuitive factors: its overall reward quality and its reasoning length.

A trajectory whose reward vector is closer to the ideal point z^* corresponds to higher overall reward, and is therefore considered easier. Similarly, shorter reasoning trajectories generally indicate simpler reasoning processes and are also treated as easier cases.

Taking both factors into account, we define the difficulty score of a trajectory Y as a linear combination of its normalized depth and its distance to the ideal point:

$$D(Y) = \alpha \frac{\text{depth}(Y)}{T_{\max}} + (1-\alpha) \frac{\|z^* - \mathbf{R}_{\text{traj}}(Y)\|_2}{\sqrt{M}}, \tag{21}$$

where $T_{\max}$ denotes the maximum allowed number of reasoning steps, M is the number of reward dimensions, and $\mathbf{R}_{\text{traj}}(Y)$ represents the aggregated reward vector of trajectory Y. The L_2 distance measures how far the trajectory performance deviates from the ideal point, and dividing by $\sqrt{M}$ normalizes the term to $[0, 1]$.

By construction, a smaller value of $D(Y)$ indicates an easier trajectory, while larger values correspond to harder cases. We use this difficulty score to perform curriculum-style difficulty scheduling during training, prioritizing easier trajectories in early stages and gradually incorporating harder ones as training progresses.

Closed-loop training. Searching with the current policy yields better-balanced trajectories; these trajectories give us preference data with graded difficulty; the updated policy, in turn, improves the quality of the search. This closes the loop and steadily pushes the model toward high-quality multimodal reasoning.

4 Experiment

4.1 Experimental Setup

Implementation details

We employ InternVL-2.5-MPO[Chen *et al.*, 2024b] as our base Vision-Language Model (VLM). The supervised fine-tuning is conducted on ShareGPT-step-300K for one epoch. Qwen2.5-VL-32B-Instruct is adopted as the backbone of the Reward and Criteria Generation Model. The BAAI/bge-en-icl model [Lee *et al.*, 2025] is used to construct the embedding space $V \in \mathbb{R}^d$, where each element corresponds to a vector representation of a generated criterion, and the dimensionality is set to $d = 4096$. To organize these criteria at multiple granularities, hierarchical clustering is performed using the BIRCH algorithm [Zhang *et al.*, 1996], resulting in a hierarchical structure H over the criterion embeddings. For the training time, SFT takes about 9 hours on a single A800 node, and three rounds of Curriculum DPO takes about 18 hours in total on a 4 * A100 node. We set branch factor $k = 3$ for a good balance between exploration and computation time. We set Search Depth $D = 10$ to limit the search space while maintaining adequate coverage of reasoning paths. We empirically set Reward Parameters $\eta = 0.5$, $\rho = 0.1$, and $\lambda = 0.2$ to optimize reasoning quality while avoiding overfitting to any single dimension.

Method	Size	MathVista	MMStar	MMMU	M3CoT	AI2D	ChartQA	Average
LLaVA-1.5 [Liu *et al.*, 2024a]	13B	27.6	32.8	-	39.5	-	-	-
IXC-2.5 [Zhang *et al.*, 2024]	7B	63.7	59.9	42.9	-	81.5	82.2	-
Ovis1.5-LLaMA3 [Lu *et al.*, 2024b]	8B	63.0	57.3	48.3	-	-	76.4	-
LLaVA-CoT [Xu *et al.*, 2025]	11B	54.8	57.6	-	-	78.7	-	-
LlamaV-o1 [Thawakar *et al.*, 2025]	11B	54.4	59.5	-	-	81.2	-	-
Vision-R1-LlamaV [Huang *et al.*, 2025]	11B	62.7	61.4	-	-	-	83.9	-
R1-VL [Zhang *et al.*, 2025a]	7B	63.5	60.0	-	-	-	83.9	-
R1-Onevision [Yang *et al.*, 2025]	7B	64.1	-	-	-	-	-	-
MiniCPM-V-2-6 [Yao *et al.*, 2024]	8B	60.6	57.5	49.8	56.0	82.1	79.4	64.2
LLaVA-OneVision [Li *et al.*, 2024]	7B	63.2	61.7	48.8	52.3	81.4	80.0	64.6
Qwen2-VL [Wang *et al.*, 2024b]	7B	58.2	60.7	53.7	57.8	83.0	77.4	65.1
Insight-V [Dong *et al.*, 2025b]	7B	59.9	61.5	50.2	61.5	79.8	81.5	65.7
InternVL-2.5 [Chen *et al.*, 2024b]	8B	62.2	59.6	54.1	62.4	84.5	84.9	68.0
LLaVA-NeXT-LLaMA3 [Liu *et al.*, 2024b]	8B	45.9	43.1	36.9	45.6	71.5	69.4	52.1
LLaVA-NeXT-LLaMA3 + SFT	8B	51.4	54.7	39.6	67.4	76.1	75.7	60.8
InternVL2.5-MPO [Wang *et al.*, 2024c]	8B	65.0	60.7	53.8	67.5	84.2	85.0	69.4
InternVL2.5-MPO + SFT	8B	65.9	61.0	53.7	75.7	81.6	**88.3**	71.0
MMS-PRM	8B	**67.5**	**65.2**	**54.2**	**79.7**	**84.2**	87.2	**73.0**

Table 1: Comparison with open-source VLMs. We evaluate our method on six benchmarks covering both general and task-specific reasoning capacities. Our method has consistently strong performances across these benchmarks, surpassing different baselines and other state-of-the-art VLMs by large margins. The items in bold and underlined respectively represent the first or second highest scores.

Benchmarks

We evaluate our method on several representative benchmarks, spanning a wide range of tasks that demand complex multimodal reasoning abilities. For most of these benchmarks, performance is measured by accuracy (or relaxed accuracy)—i.e., the ratio of correct predictions to total examples—unless specified otherwise. MathVista [Lu *et al.*, 2024a] is a well-known benchmark for multimodal mathematical reasoning, focusing on tasks such as plane geometry, functions, and visual puzzles. MMMU [Yue *et al.*, 2024] evaluates multimodal reasoning across diverse university-level disciplines. ChartQA [Masry *et al.*, 2022] aims to assess logical and numerical reasoning over chart-based visual data. MMStar [Chen *et al.*, 2024a] focuses on diverse multimodal problems that require fine-grained visual understanding. M3CoT [Shao *et al.*, 2024] evaluates multimodal chain-of-thought reasoning capabilities. AI2D [Kembhavi *et al.*, 2016] is designed to test scientific diagram understanding and reasoning.

4.2 Main Results

Table 1 reports the comparison between MMS-PRM and state-of-the-art open-source VLMs across six multimodal reasoning benchmarks. Our method consistently achieves performance gains across the majority of benchmarks Starting from InternVL2.5-MPO, supervised fine-tuning yields limited gains, indicating that performance is close to saturation under standard training. In contrast, introducing MMS-PRM leads to consistent improvements across all benchmarks, demonstrating the effectiveness of search-enhanced, multi-dimensional process supervision beyond conventional SFT. The gains are particularly pronounced on reasoning-intensive benchmarks such as M3CoT and MathVista, which require long-horizon, multi-step reasoning. This suggests that MMS-PRM is especially effective at refining complex multimodal reasoning behaviors rather than optimizing for final-answer correctness alone.

Overall, these results show that explicitly balancing multiple reward dimensions at the process level remains a powerful mechanism for improving strong vision-language models.

4.3 Efficiency and Generalization Analysis

To further validate the practicality of MMS-PRM, we compare it against state-of-the-art multimodal PRMs that require extensive supervised fine-tuning. As shown in Table 2, while training-based methods such as Visual-PRM and SVIP achieve slightly superior accuracy by optimizing billions of parameters on large-scale step-level datasets, MMS-PRM remains highly competitive despite utilizing a training-free reward mechanism. This design avoids the need for training a separate reward model, distinguishing it from baselines that require optimizing billions of parameters for step-level supervision. The minor performance trade-off is significantly offset by our model's superior deployment efficiency and its robust generalization across diverse benchmarks, as it avoids the risk of overfitting to specific training distributions or reward hacking. Consequently, MMS-PRM offers a more resource-efficient and scalable paradigm for real-world multimodal reasoning.

4.4 Ablation Study

We perform ablation studies on the M3CoT validation set to analyze the contribution of each component in MMS-PRM. As shown in Table 3, introducing the hierarchical fine-grained process rewards consistently improves the SFT baseline, indicating the effectiveness of step-level supervision for multimodal reasoning. Further incorporating reward-guided Chebyshev MCTS leads to additional gains, demonstrating

Method	PRM Trainable Params	MathVista	MMMU
Visual-PRM	7B	68.5	60.2
SVIP	7B	61.6	61.3
URSA	8B	64.9	-
VL-PRM300K	8B	65.7	56.5
MMS-PRM	**0M**	67.5	54.2

Table 2: Comparison with state-of-the-art multimodal PRMs. Our method achieves competitive performance using a training-free process reward model, whereas baselines require extensive supervision to train their reward models.

Configuration	Acc. (%)
Baseline (SFT)	67.4
+ Hierarchical reward	70.1
+ Reward + Chebyshev MCTS	73.6
MMS-PRM (full)	**79.7**
DPO only (w/o MCTS)	71.2
Weighted-sum MCTS	75.3

Table 3: Ablation study results on the M3CoT validation set.

that explicitly searching for balanced reasoning trajectories is more effective than direct optimization. Compared with weighted-sum MCTS, Chebyshev scalarization achieves better performance, suggesting that penalizing the weakest reward dimension is crucial for multimodal reasoning. Finally, the full MMS-PRM framework, which combines hierarchical rewards, Chebyshev MCTS, and curriculum-style DPO, achieves the best performance. In contrast, removing MCTS and applying DPO alone results in a noticeable drop, highlighting the necessity of search-based trajectory optimization.

4.5 Impact of Reward Count

We further investigate the impact of supervision density by analyzing the average reward counts. In this experiment, we control the density of process supervision by explicitly constraining the number of output rewards in the prompt provided to the criteria generation model. Specifically, we adjust the instructions to request a varying number of criteria—ranging from a strict constraint (e.g., "output only the single most critical criterion") to a relaxed setting that encourages generating comprehensive criteria across multiple dimensions. As shown in Table 4, the reasoning performance consistently improves as the average number of activated process rewards increases from 1.0 to 4.22. Here, the average values are induced by prompt-level constraints on the number of generated criteria (e.g., exactly 1, 1–3, $\leq$5, and 3–5), rather than being directly set. This trend indicates that MMS-PRM benefits significantly from joint multi-dimensional constraints.

4.6 Worst-Dimension Analysis

We analyze the minimum reward dimension $v_{\min} = \min_j v_j$ of final trajectories generated by Chebyshev-guided MCTS and weighted-sum MCTS under identical rollout budgets and

Avg Rewards / Step	MathVista	M3CoT	ChartQA	Avg.
1.0	65.8	74.1	86.4	74.8
2.1	66.9	76.3	87.1	76.1
3.4	68.2	78.8	**88.4**	77.8
4.2	**70.1**	**79.7**	87.2	**79.6**

Table 4: Effect of reward dimensionality measured by the average number of activated rewards per reasoning step.

reward models. As shown in Figure 3, Chebyshev scalarization consistently shifts the distribution of $v_{\min}$ toward higher values, indicating that the weakest reward dimension is significantly improved. In contrast, weighted-sum MCTS exhibits a heavier tail in the low-$v_{\min}$ region, suggesting that severe failures in individual dimensions can be masked by strong performance in others. This confirms that Chebyshev-based optimization effectively enforces balanced multimodal reasoning by penalizing worst-dimension collapse rather than allowing compensation across dimensions.

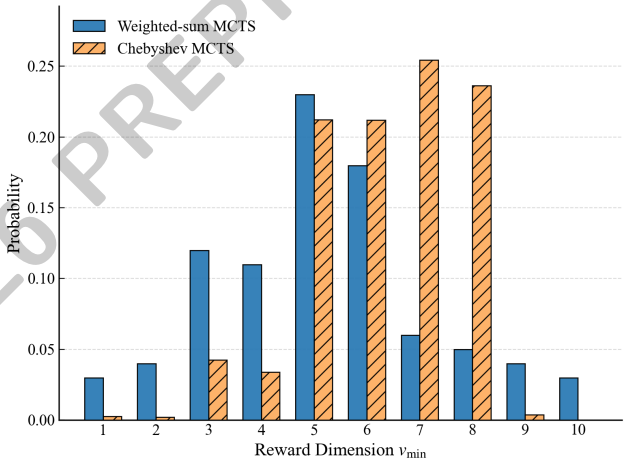

Figure 3: Distribution of the minimum reward dimension $v_{\min}$ across reasoning trajectories

5 Conclusion

This paper introduced MMS-PRM, which replaces superficial scalar rewards with fine-grained, multi-dimensional supervision via Chebyshev-scalarized MCTS. Our framework effectively balances visual and logical reasoning, achieving competitive results and robust generalization. These findings highlight the critical role of structured process rewards and search mechanisms in advancing reliable, interpretable multimodal reasoning.

Acknowledgments

This work was supported by the National Key R&D Program of China under Grant No. 2024YFC3308101, for the project "Long- and Short-Term Holographic Profiling of Bond Investors Based on Trading Behavior Features", led by Prof. Huaping Zhang.

References

[Cai *et al.*, 2025] Chengkun Cai, Haoliang Liu, Xu Zhao, Zhongyu Jiang, Tianfang Zhang, Zongkai Wu, John Lee, Jenq-Neng Hwang, and Lei Li. Bayesian Optimization for Controlled Image Editing via LLMs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025.

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

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

[Dong *et al.*, 2025a] Guanting Dong, Chenghao Zhang, Mengjie Deng, Yutao Zhu, Zhicheng Dou, and Ji-Rong Wen. Progressive multimodal reasoning via active retrieval. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3579–3602, 2025.

[Dong *et al.*, 2025b] Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9062–9072, 2025.

[Gao *et al.*, 2025] Minghe Gao, Xuqi Liu, Zhongqi Yue, Yang Wu, Shuang Chen, Juncheng Li, Siliang Tang, Fei Wu, Tat-Seng Chua, and Yueting Zhuang. Benchmarking multimodal cot reward model stepwise by visual program. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1718–1728, 2025.

[Guan *et al.*, 2025] Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. Rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *arXiv preprint arXiv:2501.04519*, 2025.

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

[Kembhavi *et al.*, 2016] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer, 2016.

[Lee *et al.*, 2025] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. In *International Conference on Learning Representations*, volume 2025, pages 79310–79333, 2025.

[Li *et al.*, 2024] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[Liu *et al.*, 2024a] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

[Liu *et al.*, 2024b] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024.

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

[Lu *et al.*, 2024b] Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye. Ovis: Structural embedding alignment for multimodal large language model. *arXiv preprint arXiv:2405.20797*, 2024.

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

[Luo *et al.*, 2025] Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinzhe Ni, Zicheng Lin, Jin Zeng, and Yujiu Yang. Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. *arXiv e-prints*, pages arXiv–2501, 2025.

[Masry *et al.*, 2022] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 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, 2022.

[Ong *et al.*, 2025] Brandon Ong, Tej Deep Pala, Vernon Toh, William Chandra Tjhi, and Soujanya Poria. Training vision-language process reward models for test-time scaling in multimodal reasoning: Key insights and lessons learned. *arXiv preprint arXiv:2509.23250*, 2025.

[Qi *et al.*, 2025] Zhenting Qi, Mingyuan Ma, Jiahang Xu, Li Lyna Zhang, Fan Yang, and Mao Yang. Mutual reasoning makes smaller llms stronger problem-solver. In *International Conference on Learning Representations*, volume 2025, pages 20788–20807, 2025.

[Shao *et al.*, 2024] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and

Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024.

[Shi *et al.*, 2026] Jingzhe Shi, Qinwei Ma, Hongyi Liu, Hang Zhao, Jenq-Neng Hwang, and Lei Li. Intrinsic entropy of context length scaling in llms. In *The Fourteenth International Conference on Learning Representations*, 2026.

[Snell *et al.*, 2024] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

[Thawakar *et al.*, 2025] Omkar Thawakar, Dinura Dissanayake, Ketan Pravin More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Ilmuz Zaman Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. Llamav-o1: Rethinking step-by-step visual reasoning in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24290–24315, 2025.

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

[Wang *et al.*, 2024b] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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

[Wang *et al.*, 2024d] Zihan Wang, Yunxuan Li, Yuexin Wu, Liangchen Luo, Le Hou, Hongkun Yu, and Jingbo Shang. Multi-step problem solving through a verifier: An empirical analysis on model-induced process supervision. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7309–7319, 2024.

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

[Wu *et al.*, 2024] Jinyang Wu, Mingkuan Feng, Shuai Zhang, Feihu Che, Zengqi Wen, Chonghua Liao, and Jianhua Tao. Beyond examples: High-level automated reasoning paradigm in in-context learning via mcts. *arXiv preprint arXiv:2411.18478*, 2024.

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

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

[Yao *et al.*, 2024] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

[Yao *et al.*, 2026] Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *Advances in Neural Information Processing Systems*, 38:29918–29952, 2026.

[Yue *et al.*, 2024] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, 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*, pages 9556–9567, 2024.

[Zhang *et al.*, 1996] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch: an efficient data clustering method for very large databases. *ACM sigmod record*, 25(2):103–114, 1996.

[Zhang *et al.*, 2024] Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*, 2024.

[Zhang *et al.*, 2025a] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. R1-vl: Learning to reason with multimodal large language models via step-wise group relative policy optimization. *arXiv preprint arXiv:2503.12937*, 2025.

[Zhang *et al.*, 2025b] Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1662, 2025.