

Reinforcing Multimodal Reasoning via Token-Level Perception-Grounded Advantage Estimation

Zhihan Zhang

Singapore Management University
Singapore, Singapore
zhihanzhang.2024@phdcs.smu.edu.sg

Lizi Liao

Singapore Management University
Singapore, Singapore
lzliao@smu.edu.sg

Abstract

Reinforcement Learning with Verifiable Rewards (RLVR) has improved the reasoning capabilities of Multimodal Large Language Models (MLLMs), yet existing frameworks rely on coarse, sequence-level reward signals that lack the fine-grained supervision over the visually-grounded steps within a multimodal reasoning chain. We investigate this gap through the lens of two token-level metrics: *visual dependency* (i.e. how much a token's prediction relies on the input image features) and *predictive entropy*. Our empirical analysis reveals two key findings: (1) correct reasoning chains exhibit a markedly sharper entropy reduction as visual grounding intensifies, compared to incorrect ones; (2) pivotal tokens, those whose misprediction triggers reasoning collapse, are statistical outliers in the joint distribution of visual dependency and predictive entropy derived from correct chains. Motivated by these findings, we propose **Token-level Perception-grounded Advantage Estimation (TPAE)**, which estimates token-level advantages by measuring each token's statistical consistency with the vision-entropy patterns of correct rollouts. TPAE leverages this granular score to modulate the sequence-level advantage, producing a fine-grained supervision signal that can be integrated into various RLVR frameworks. Extensive experiments on seven benchmarks show that TPAE consistently outperforms leading strong baselines, yielding more stable and efficient optimization for multimodal reasoning. The code is publicly available at <https://github.com/Zhihan72/TPAE>.

CCS Concepts

• **Computing methodologies** → **Reinforcement learning**; **Computer vision tasks**; **Natural language generation**.

Keywords

Multimodal Large Language Model; Multimodal Reasoning; Reinforcement Learning

ACM Reference Format:

Zhihan Zhang and Lizi Liao. 2026. Reinforcing Multimodal Reasoning via Token-Level Perception-Grounded Advantage Estimation. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26)*, November 10–14, 2026, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3767308.3835936>



This work is licensed under a Creative Commons Attribution 4.0 International License. *MM '26, Rio de Janeiro, Brazil*

© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2213-4/2026/11
<https://doi.org/10.1145/3767308.3835936>

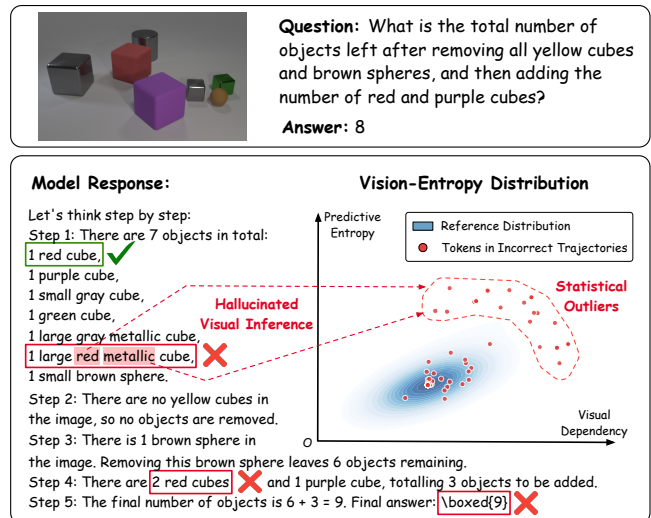


Figure 1: Token-level joint distribution of visual dependency and predictive entropy, where outlier tokens from incorrect trajectories emerge as the triggers of reasoning collapse.

1 Introduction

Reinforcement Learning from Verifiable Rewards (RLVR), powered by efficient online algorithms such as Group Relative Policy Optimization (GRPO) [24], has fundamentally shifted the frontier of reasoning capabilities in Large Language Models (LLMs). By optimizing models using verifiable signals, such as structured thinking formats and final answer accuracy, RLVR has demonstrated strong empirical success in models like DeepSeek-R1 [4]. Translating this success to Multimodal Large Language Models (MLLMs), however, introduces unique complexities where fine-grained perception is tightly synthesized with multi-step logical deduction [19, 31, 43]. This challenge has motivated an array of research aimed at adapting RLVR to multimodal reasoning, primarily through data-centric enhancements [4, 5, 9, 38], reward modelling [6, 7, 18, 34], and rollout-level algorithmic adjustments [17, 30, 39].

Despite this progress, existing RLVR frameworks for multimodal reasoning are fundamentally limited by the coarse granularity of their reward signals [8, 33, 35, 45]. Mainstream reinforcement learning (RL) algorithms such as GRPO [24] and DAPO [40] derive a single, rollout-level advantage from outcome-based rewards, assigning a uniform advantage value to every token within a rollout. This indiscriminating assignment treats all tokens—whether reliably perception-grounded or responsible for reasoning failures—as equally contributing to the final answer, introducing gradient noise

that hinders policy convergence [14, 26]. A natural remedy is to introduce finer-grained supervision via process-level reward models [6, 35]; however, this approach requires training a separate learned scorer, which is data-demanding, computationally expensive, and vulnerable to reward hacking. This motivates a key research question: *Can we obtain fine-grained, perception-grounded advantage estimates directly from the policy’s own reasoning behavior, without the overhead of auxiliary reward models?*

To answer this question, we begin by examining two intrinsic metrics that are readily available from the policy model itself: a token’s visual dependency and its predictive entropy. Through an empirical analysis of 12,800 reasoning trajectories generated by Qwen2.5-VL-7B across 100 questions, we uncover a striking pattern illustrated in Figure 2: *correct reasoning chains exhibit sharply lower entropy as visual dependency increases, whereas incorrect chains maintain high entropy despite strong visual dependency*—a phenomenon we term *non-resolving grounding*, where the model attends to the visual input but fails to extract the information needed to disambiguate its next-token predictions. This widening gap between correct and incorrect chains establishes the coupling of visual dependency and predictive entropy as an effective, model-intrinsic indicator of multimodal reasoning quality.

To pinpoint the mechanistic drivers of this divergence, we zoom into the token-level distribution. Using the joint distribution of visual dependency and entropy from correct reasoning chains as a reference, we measure the statistical alignment of each token in failed trajectories against the reference. While the majority of tokens in incorrect reasoning chains fall within the reference distribution, a critical 7.64% emerge as severe outliers (e.g., the red-highlighted tokens in Figure 1). A rigorous human evaluation of 200 randomly sampled outliers confirms that *these divergent tokens are not random noise but the semantic pivots that trigger visual mis-grounding and logical reasoning errors*—precisely the tokens that a fine-grained advantage estimator should down-weight.

Building on these findings, we propose **Token-level Perception-grounded Advantage Estimation (TPAE)**, a novel algorithm that refines coarse, rollout-level reward signals into dense, token-level advantage estimates within multimodal RLVR frameworks—without training any auxiliary reward model. TPAE jointly tracks each token’s visual dependency and predictive entropy—its *vision-entropy state*—and leverages correct rollouts to build reference distributions over these states. Each token is scored by its statistical alignment with the corresponding reference distribution through a hypothesis testing perspective: tokens consistent with the reference are deemed trustworthy, while severe outliers are penalized. These per-token scores then modulate the rollout-level advantage, directing gradient updates toward perceptually grounded reasoning and away from the triggers of reasoning failure. As a lightweight, model-intrinsic module, TPAE can be seamlessly integrated into existing RL algorithms such as GRPO and DAPO.

We validate the effectiveness of TPAE in advancing multimodal reasoning by integrating it into two base RL algorithms across two MLLM backbones. Our evaluation results across seven diverse benchmarks demonstrate that integrating TPAE with either GRPO

or DAPO yields consistent performance gains; notably, our TPAE-D-Qwen2.5-7B surpasses existing state-of-the-art multimodal reasoning models utilizing the same Qwen2.5-VL-7B backbone. To sum up, our **main contributions** are threefold:

- We reveal that the coupling between token-level visual dependency and predictive entropy is an effective, model-intrinsic indicator of multimodal reasoning quality, and identify statistically divergent tokens as the drivers of reasoning failures.
- We propose TPAE, a token-level perception-grounded advantage estimation algorithm that translates this insight into dense, per-token supervision requiring no auxiliary reward model.
- Extensive experiments across seven multimodal benchmarks demonstrate the superior performance of TPAE over baseline algorithms and existing state-of-the-art models.

2 Related Work

Multimodal Large Language Model. Existing advancements in MLLMs have primarily focused on the modality alignment via multimodal projectors that bridge vision encoders with pre-trained Large Language Model (LLM) backbones. Diverse alignment strategies have been explored to optimize visual perception, ranging from compressing visual tokens [1, 13] to fusing multiple vision encoders [15, 29]. Beyond the exploration of architecture of MLLMs, large-scale instruction tuning has emerged as a central technique for improving multimodal alignment and overall performance on multimodal benchmarks [1, 16]. However, MLLMs still exhibit a significant performance gap in complex multimodal reasoning tasks [19, 27, 28, 43, 44]. Drawing on the success of policy optimization in textual reasoning, recent works have begun adapting foundational algorithms such as PPO [23] and GRPO [24] to the multimodal setting to tackle increasingly complex reasoning tasks.

Multimodal Reinforcement Learning. Most strategies in multimodal reinforcement learning have focused on enhancing the RLVR framework from three perspectives: data, rollout and reward. Data-centric methods focus on the curation of large-scale, visually-grounded datasets by distilling chain-of-thought trajectories from proprietary models, typically employing a two-stage training paradigm consisting of initial supervised fine-tuning (SFT) followed by reinforcement learning via the GRPO objective [2, 4, 5, 9, 10, 38]. Works on rollout improvements focus on diversifying the rollout space by visual augmentations [17, 39] or sample replay mechanisms [12, 30]. Reward-centric works aim to evaluate the visual grounding of rollouts by either enforcing verifiable intermediate steps—such as the generation of spatial coordinates for relevant objects [7, 18] or auxiliary captioning tasks [34]—or by deploying external neural reward models for dense supervision [6, 42].

However, existing works predominantly conceptualize visual grounding as a post-hoc supervisory constraint, treating sampled rollouts as static artifacts for external verification. While more recent efforts have introduced token-level visual dependency to probe the internal dynamics of multimodal reasoning, this metric remains relegated to passive filters or heuristic penalties within the optimization objective [8, 33]. In this work, we propose TPAE, a novel perception-grounded advantage estimation algorithm that leverages the token-level joint dynamics of visual dependency and predictive entropy to modulate the coarse reward signal.

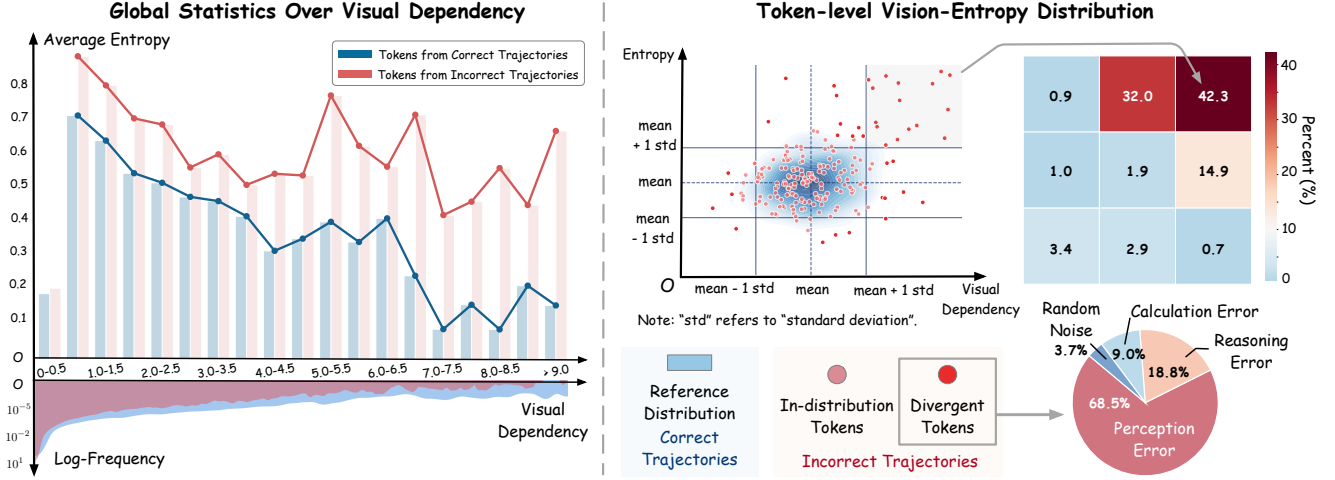


Figure 2: Empirical analysis of reasoning trajectories using Qwen2.5-VL-7B. (a) Distribution of average entropy and log-frequency across visual dependency levels (left). (b) Density of token-level vision-entropy misalignment relative to the reference distribution center (upper right). (c) Error analysis of divergent tokens through human annotation (lower right).

3 Preliminary

Existing RLVR frameworks rely on coarse, sequence-level reward signals that overlook the varying significance of individual tokens. To move beyond this simplification, we introduce token-level entropy and visual dependency as granular metrics to characterize the internal dynamics of multimodal reasoning generation. Our empirical analysis reveals that correct reasoning chains exhibit a sharper entropy reduction tied to visual grounding than the incorrect trajectories. Furthermore, a non-trivial proportion of tokens from incorrect trajectories emerge as statistical outliers relative to the correct reference distribution. Through human evaluation, these outliers are confirmed to be the pivotal triggers responsible for multimodal reasoning collapse.

3.1 Group Relative Policy Optimization

Group Relative Policy Optimization (GRPO) [24] is a variant of the Proximal Policy Optimization (PPO) [23] algorithm that removes the requirement for a separate value model by estimating advantages via group-based computation. For a given multimodal input (I, q) consisting of a visual context I and a textual query q , the old policy model $\pi_{\theta_{\text{old}}}$ samples a group of G rollouts, $\{o_i\}_{i=1}^G$.

In the RLVR framework, each rollout o_i is assigned a binary reward $R_i \in \{0, 1\}$, determined solely on whether its final answer matches the ground truth. The token-level advantage $\hat{A}_{i,j}$ is then defined as the sequence-level reward normalized across the group:

$$\hat{A}_{i,j} = \hat{A}_i = \frac{R_i - \text{mean}(\{R_k\}_{k=1}^G)}{\text{std}(\{R_k\}_{k=1}^G)} \quad (1)$$

The policy π_{θ} is then updated to maximize a clipped surrogate objective, where this uniform advantage $\hat{A}_i$ is broadcast to every

token $o_{i,j}$ within the rollout o_i :

$$\mathcal{L}^{\text{GRPO}}(\theta) = \mathbb{E}_{[\{o_i\}_{i=1}^G \sim \pi_{\theta_{\text{old}}}(\cdot | I, q)]} \frac{1}{G} \sum_{i=1}^G \frac{1}{|o_i|} \sum_{j=1}^{|o_i|} \left\{ \min \left(r_{i,j}(\theta) \hat{A}_{i,j}, \text{clip}(r_{i,j}(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_{i,j} \right) - \beta \mathbb{D}_{\text{KL}}[\pi_{\theta} | \pi_{\text{ref}}] \right\} \quad (2)$$

where $r_{i,t}(\theta) = \frac{\pi_{\theta}(o_{i,j} | I, q, o_{i,<j})}{\pi_{\theta_{\text{old}}}(o_{i,j} | I, q, o_{i,<j})}$ is the probability ratio.

Decoupled Clip and Dynamic Sampling Policy Optimization (DAPO) [40] is a notable extension of GRPO, incorporating several modifications such as clip-higher, dynamic sampling, and token-level policy gradient loss (see supplementary materials for details). In this work, we investigate the application of our method to both GRPO and DAPO algorithms.

3.2 Token Entropy and Visual Dependency

While GRPO mitigates reward sparsity through a group-relative advantage estimation, it relies on a coarse, sequence-level signal that ignores the varying significance of individual tokens. To move beyond this sequence-level simplification, we introduce token entropy and visual dependency as granular metrics to characterize the internal dynamics of the multimodal generation process.

3.2.1 Predictive Entropy. The token-level entropy quantifies the model’s predictive uncertainty in the reasoning chain. Specifically, for token $o_{i,j}$ in rollout o_i , the policy model generates a probability distribution $P_{i,j}$ over the vocabulary $\mathcal{V}$, i.e. $P_{i,j} = \pi_{\theta}(\cdot | I, q, o_{i,<j})$. The entropy of this prediction is then calculated as

$$\mathcal{H}_{i,j} := - \sum_{v \in \mathcal{V}} P_{i,j}(v) \log P_{i,j}(v) \quad (3)$$

where $P_{i,j}(v)$ represents the likelihood of choosing token v . A low $\mathcal{H}_{i,j}$ value indicates that the model is highly confident, with its probability concentrated on a single optimal next token. In contrast,

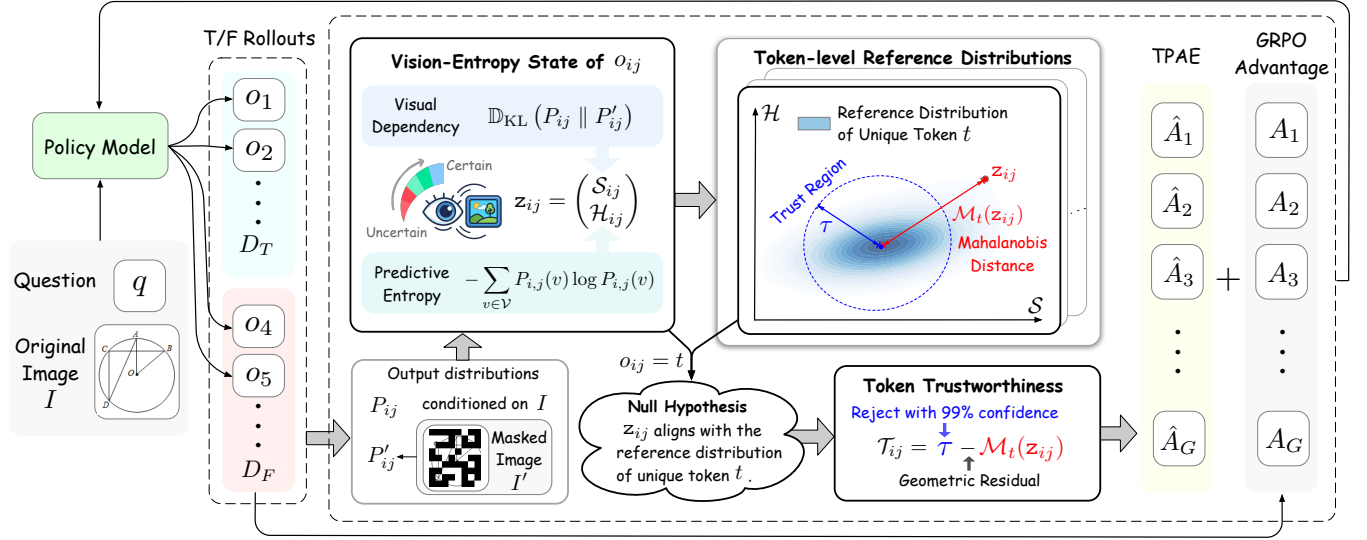


Figure 3: Overview of the TPAE algorithm. TPAE utilizes correct rollouts to establish reference distributions of vision-entropy states of unique tokens. From a hypothesis testing perspective, TPAE then applies a trust boundary τ to quantify token-level trustworthiness based on distributional alignment. This metric is then integrated with the rollout-level advantage from GRPO.

a high $\mathcal{H}_{i,j}$ value indicates a state of high uncertainty, where the model identifies multiple plausible paths for the next step.

3.2.2 Visual Dependency. The token-level visual dependency measures the degree of reliance on the visual input I during the generation of token o_{ij} [8, 33]. Formally, we define $P'_{i,j} = \pi_\theta(\cdot \mid I', q, o_{i,<})$ as the probability distribution over the vocabulary $\mathcal{V}$ conditioned on a masked, non-informative version of the image I' . From the perspective of information gain, the visual dependency is calculated as the Kullback-Leibler (KL) divergence between the original distribution $P_{i,j}$ and the vision-blind distribution $P'_{i,j}$:

$$\mathcal{S}_{ij} := \mathbb{D}_{KL}(P_{i,j} \parallel P'_{i,j}) = \sum_{v \in \mathcal{V}} P_{i,j}(v) \log \frac{P_{i,j}(v)}{P'_{i,j}(v)} \quad (4)$$

A high $\mathcal{S}_{ij}$ value indicates that the visual input provides critical information for the prediction of token o_{ij} , marking it as a key moment of perception-anchored reasoning. Conversely, a low $\mathcal{S}_{ij}$ value suggests the model is relying primarily on linguistic knowledge or the preceding textual context to determine the next token.

Following [8, 33], we implement a patch-based masking strategy for I' at a 60% masking ratio. Specifically, 60% of the image patches are randomly sampled and masked, effectively restricting the available visual cues for visual dependency quantification.

3.3 Analysis of Multimodal Reasoning

We conduct an empirical investigation of multimodal reasoning using Qwen2.5-VL-7B-Instruct [27] across 100 questions sampled from the ViRL39K dataset [30]. By generating 128 independent trajectories per question, we establish a robust distribution consisting of 5,995 correct and 6,805 incorrect trajectories. We observe a clear structural divergence: correct reasoning chains are significantly more concise, averaging 418.07 tokens compared to the 546.57 tokens of their incorrect counterparts. Moreover, as shown

in Figure 2(a), correct reasoning chains show a higher log-frequency density in regions of high visual dependency, while incorrect trajectories are heavily concentrated in the minimal dependency interval.

By integrating predictive entropy $\mathcal{H}_{ij}$ with visual dependency $\mathcal{S}_{ij}$, we investigate how visual grounding modulates the model's internal certainty across both correct and incorrect reasoning chains:

Insight 1: Correct reasoning chains exhibit a sharper entropy reduction through visual grounding. Figure 2(a) illustrates the average entropy at varying levels of visual dependency, revealing a clear divergence in how visual information impacts model certainty. In correct trajectories, we observe a monotonic decrease in entropy as visual dependency increases, suggesting that the model successfully translates visual cues into predictive certainty. Conversely, in incorrect trajectories, entropy remains at a consistently high plateau regardless of the level of visual reliance. This divergence implies that while correct reasoning chains use visual context to effectively prune the output distribution, incorrect reasoning paths suffer from non-resolving grounding—a phenomenon where the model attends to the image but fails to extract the information required to resolve its internal uncertainty.

Insight 2: Vision-entropy misalignment pinpoints the pivotal tokens triggering multimodal reasoning collapse. We then characterize the token-level vision-entropy state $(\mathcal{S}_t, \mathcal{H}_t)$ by constructing the empirical joint distribution of visual dependency and entropy for each unique token t across all trajectories of a given question. By establishing the distributions from correct reasoning chains as a reference baseline, we can evaluate the statistical alignment of tokens generated in failed trajectories. Our analysis reveals that while most tokens in failed trajectories align with the reference, a critical 7.64% fall into the extreme distributional tails, deviating by more than three standard deviations from reference center. To investigate these divergent tokens, we partition them into a 3×3 diagnostic grid, categorized by whether their predictive entropy

and visual dependency deviate by more than one standard deviation from the reference mean. As illustrated in Figure 2(b), over 74% of these tokens are concentrated within the upper-middle and upper-right regions of the grid. This concentration reveals a decoupling of perception and reasoning in failed trajectories, providing granular validation for the global statistics observed in Figure 2(a).

To validate these statistical trends, we conduct a human evaluation on a random subset of 200 divergent tokens, each drawn from a unique incorrect trajectory. We tasked three annotators with classifying these tokens into four distinct categories. As shown in Figure 2(c), 68.5% of these tokens are perception errors stemming from visual mis-grounding, while 18.8% are reasoning errors that misinterpret valid visual evidence during multi-step inference. Only 3.7% represent benign noise, confirming that these divergent tokens are overwhelmingly linked to reasoning failure. Further details on annotation protocol are provided in supplementary materials. Collectively, these findings demonstrate that distributional divergence from reference distributions is an effective indicator for localizing the pivotal moments of multimodal reasoning collapse.

4 Method

Building upon our empirical findings, we propose **Token-level Perception-grounded Advantage Estimation (TPAE)**, an algorithm designed to refine coarse, sequence-level signals of outcome-based rewards into dense, token-level supervision by coupling visual dependency with predictive entropy. For each question, TPAE leverages correct rollouts to establish reference distributions of vision-entropy states for unique tokens. TPAE then quantifies the trustworthiness of individual tokens by measuring their statistical alignment with these reference distributions through a hypothesis-testing perspective. This metric is integrated directly into the rollout-level advantage computed by mainstream RL algorithms to penalize tokens deviating from the established trust region.

4.1 Vision-Entropy Distribution

For a given multimodal input (I, q) , we sample a group of G rollouts, $\{o_i\}_{i=1}^G$, each with a corresponding rule-based, outcome-oriented reward $\{R_i\}_{i=1}^G$ indicating its correctness. We divide these rollouts into a correct set $D_T = \{o_i \mid R_i = 1\}$ and an incorrect set $D_F = \{o_i \mid R_i = 0\}$. In our work, we utilize the vision-entropy distribution from correct rollouts D_T as a reference baseline to capture the optimal joint dynamics of visual grounding and predictive entropy that facilitate a correct final answer.

For each unique token t in the correct set D_T , we aggregate its observed vision-entropy states $Z_t = \{z_{ij} \mid o_{ij} = t, o_i \in D_T\}$, where each state $z_{ij} = [S_{ij}, \mathcal{H}_{ij}]^\top \in \mathbb{R}^2$ captures the token’s visual dependency S and entropy $\mathcal{H}$. We then approximate the reference distribution for each unique token t through multivariate Gaussian, i.e. $P(z \mid t) \sim \mathcal{N}(\mu_t, \Sigma_t)$, defined as:

$$P(z \mid \mu_t, \Sigma_t) = \frac{1}{2\pi\sqrt{|\Sigma_t|}} \exp\left(-\frac{1}{2}(z - \mu_t)^\top \Sigma_t^{-1}(z - \mu_t)\right) \quad (5)$$

where $\mu_t = \frac{1}{|Z_t|} \sum_{z \in Z_t} z$, $\Sigma_t = \frac{1}{|Z_t|-1} \sum_{z \in Z_t} (z - \mu_t)^\top (z - \mu_t)$. The choice of a multivariate Gaussian is motivated by the observed uniform centrality of vision-entropy states in D_T (see supplementary materials for comparisons with alternative parametric models).

To ensure the robustness of the token-level estimates, we compute the covariance matrix Σ_t using the Ledoit-Wolf adaptive shrinkage approach [11]. This addresses the sample scarcity inherent to RLVR frameworks, where limited rollouts may yield unstable or singular token-level empirical estimates. Specifically, we parameterize the estimated covariance $\hat{\Sigma}_t$ as a convex combination of the empirical covariance Σ_t and a diagonal prior Σ_t^{prior} , as follows:

$$\hat{\Sigma}_t = (1 - \lambda_t)\Sigma_t + \lambda_t\Sigma_t^{\text{prior}} \quad (6)$$

where $\Sigma_t^{\text{prior}} = \text{diag}(\sigma_{S_t}^2, \sigma_{\mathcal{H}_t}^2)$ denotes the marginal variances of visual dependency and entropy for token t , as estimated from the correct rollout set D_T . The shrinkage intensity $\lambda_t \in [0, 1]$ follows the optimal shrinkage principle, $\lambda_t = \frac{3}{|Z_t|+3}$, which adaptively weights the prior based on the sample size $|Z_t|$ [11]. Further theoretical justifications are provided in the supplementary material.

4.2 Token Trustworthiness Quantification

We quantify the trustworthiness of individual tokens by measuring their statistical alignment with the reference distribution of correct reasoning paths. Formally, we frame this quantification as a hypothesis test: for a token $o_{ij} = t$, the null hypothesis H_0 posits that its vision-entropy state $z_{ij} = [S_{ij}, \mathcal{H}_{ij}]^\top$ originates from the reference distribution $P(z \mid \mu_t, \Sigma_t)$ derived from correct rollouts.

To test the null hypothesis H_0 , we calculate the likelihood of observing the specific vision-entropy state z_{ij} under the reference distribution $P(z \mid \mu_t, \Sigma_t)$. For a multivariate Gaussian model, this probability density is strictly proportional to the exponentiated negative squared Mahalanobis distance [3]:

$$P(z_{ij} \mid \mu_t, \Sigma_t) \propto \exp\left(-\frac{1}{2}\mathcal{M}_t^2(z_{ij})\right) \quad (7)$$

where $\mathcal{M}_t(z_{ij}) = \sqrt{(z_{ij} - \mu_t)^\top \Sigma_t^{-1}(z_{ij} - \mu_t)}$ denotes the Mahalanobis distance. A larger $\mathcal{M}_t(z_{ij})$ value corresponds to a lower likelihood that z_{ij} aligns with the reference distribution, providing a clear statistical signal to reject the null hypothesis H_0 .

We employ a principled boundary τ to demarcate the trust region for H_0 , leveraging the statistical properties of the Mahalanobis distance. Under H_0 where $z_{ij} \sim \mathcal{N}(\mu_t, \Sigma_t)$, the squared Mahalanobis distance strictly follows a Chi-squared distribution with two degrees of freedom, i.e. $\mathcal{M}_t^2(z_{ij}) \sim \chi_2^2$. By adopting a 99% confidence level, we reject H_0 if the squared Mahalanobis distance exceeds the value $\chi_{2,0.99}^2 \approx 9.21$. This threshold implies that a token’s vision-entropy state z_{ij} is considered misaligned with the reference distribution if its deviation from the distribution center exceeds $\sqrt{9.21} \approx 3.03$ standard deviations, i.e., $\mathcal{M}_t(z_{ij}) > \tau = 3.03$.

Building upon the above derivations, we formally quantify the trustworthiness $\mathcal{T}_{ij}$ of a token’s vision-entropy state as the residual distance to the established trust boundary:

$$\mathcal{T}_{ij} = \tau - \mathcal{M}_t(z_{ij}) \quad (8)$$

Within this formulation, a positive trustworthiness score ($\mathcal{T}_{ij} \geq 0$) confirms that the token resides safely within the 99% confidence interval of the reference distribution. Conversely, a negative score ($\mathcal{T}_{ij} < 0$) identifies a statistically significant outlier, signaling a rejection of the null hypothesis H_0 and a critical vision-entropy misalignment of token o_{ij} . For tokens that are absent from the

reference set D_T , we assign a default trustworthiness score of zero to maintain a conservative estimation.

4.3 Token-level Advantage Estimation

We integrate the token’s trustworthiness score into the existing RLVF frameworks to refine the rollout-level advantage, effectively penalizing tokens that deviate from the established trust region. Theoretically, failing to reject H_0 indicates that the token’s vision-entropy state is statistically aligned with the reference distribution at a 99% confidence level. Since this alignment is a fundamental requirement rather than an optimization target, we exclusively impose a penalty on tokens that violate this requirement—specifically those for which H_0 is rejected due to $\mathcal{M}_t > \tau$. The final token-level advantage is then defined as:

$$\hat{A}_{ij}^{\text{TPAE}} = \hat{A}_{ij}^{\text{GRPO}} + \text{sigmoid}(\min(0, \mathcal{T}_{ij})) - 0.5 \quad (9)$$

where $\mathcal{T}_{ij} = \tau - \mathcal{M}_t(z_{ij})$ and $\hat{A}_{ij}^{\text{GRPO}}$ is defined in Eq. 1. The sigmoid function $\sigma(\cdot)$ is applied to constraining the output range since $\mathcal{M}_t \in [0, +\infty)$. By centering the transformation such that the penalty vanishes when $\mathcal{T}_{ij} \geq 0$, we preserve the original advantage for all compliant tokens while providing a dense, bounded signal to suppress vision-entropy misalignment during reinforcement learning. A detailed, step-by-step implementation of the entire training procedure is provided in the supplementary material.

5 Experiment

We validate the effectiveness of TPAE in advancing multimodal reasoning by integrating it into two foundational RL algorithms across two model backbones. Our evaluation comprises a comprehensive comparative study against state-of-the-art multimodal reasoning models across seven benchmarks. Experimental results demonstrate significant performance gains attributable to TPAE over base RL algorithms, observed in both training efficiency and downstream accuracy. Furthermore, we include a sensitivity analysis on the trust boundary τ and a qualitative analysis that illustrates the precision of our method in localizing multimodal reasoning failures.

5.1 Experimental Setup

5.1.1 Models and Baselines. We validate our method using Qwen2.5-VL-7B-Instruct [27] and Qwen3-VL-8B-Instruct [28] as the base models, integrated within the standard RL algorithms GRPO and DAPO. To ensure a comprehensive comparison, we benchmark against a diverse suite of leading multimodal reasoning models: ThinkLite-VL-7B [32], OpenVLThinker-7B [5], NoisyRollout-7B [17], MM-Eureka-7B [21], Perception-R1-7B [36], VL-Rethinker-7B [30], R1-ShareVL-7B [39], PAPO-G-7B, PAPO-D-7B [33], VPPO-7B [8], and Shuffle-R1-7B [45]. Notably, all comparative baselines utilize Qwen2.5-VL-7B-Instruct as the base model.

5.1.2 Dataset, Benchmarks, and Evaluation. We train our base models on the ViRL39K dataset [30], which contains 38.9k diverse multimodal reasoning problems. We conduct a comprehensive evaluation across seven multimodal benchmarks, spanning mathematical, geometric, logical, and multi-discipline reasoning: MathVerse [43], WeMath [22], MathVision [31], DynaMath [46], Geo3k [20], LogicVista [37], and MMMU-Pro [41]. For MathVerse, we evaluate on the full set of 3.94k multimodal examples, excluding the text-only subset to

focus on tasks requiring multimodal context. For evaluation setup, we employ an exact-match scoring protocol against ground-truth answers, which eliminates the reliance on LLM-as-a-judge systems [8, 33]. We report average accuracy@8 at an inference temperature of 1.0, using a single, fixed evaluation pipeline across all models to ensure a rigorous and fair comparison.

5.1.3 Training Details. We develop four model variants using different base RL algorithms and model backbones. Specifically, TPAE-G-Qwen2.5-7B and TPAE-D-Qwen2.5-7B are initialized from Qwen2.5-VL-7B-Instruct and trained using the GRPO and DAPO algorithms, respectively. Following the same paradigm, we extend our method to Qwen3-VL-8B-Instruct to produce TPAE-G-Qwen3-8B and TPAE-D-Qwen3-8B. The training process utilizes the VerL [25] reinforcement learning framework for policy optimization. Our models are trained for 2 epochs with a learning rate of 1e-6 on 8 NVIDIA Pro6000 96G GPUs. We employ a training batch size of 384, with 16 rollouts sampled per question. We set the maximum response length to 2048, and add a small entropy penalty of 0.06 to ensure training stability [8, 17, 33]. Within our method, we set the trust region boundary at $\tau = 3.03$, corresponding to a 99% confidence level under the Chi-squared distribution χ_2^2 .

We provide more details of benchmark compositions, implementation details and computing resources in the supplementary materials and anonymous GitHub repository.

5.2 Main Results

As illustrated in Table 1, TPAE-D-Qwen2.5-7B consistently surpasses current state-of-the-art multimodal reasoning models across all seven multimodal benchmarks. By integrating our fine-grained perception-grounded advantage estimation into the DAPO framework, TPAE-D-Qwen2.5-7B attains a leading average accuracy of 48.71% over competitive baselines, outperforming Shuffle-R1-7B by 1.39% and surpassing VPPO-7B by 1.12%. Similarly, the GRPO-based variant, TPAE-G-Qwen2.5-7B, achieves an average accuracy of 47.63%, securing the second-highest performance among all models sharing the same Qwen2.5-VL-7B backbone. The performance advantage of TPAE extends to the Qwen3-VL-8B series, demonstrating its robust scalability across evolving model architectures. Both TPAE-G-Qwen3-8B and TPAE-D-Qwen3-8B variants establish new performance frontiers by achieving an average accuracy of 60.62% and 61.35% respectively. The superiority of TPAE is further evidenced by its improved training dynamics relative to base RL algorithms as discussed in Section 5.3.1.

5.3 Ablation Studies

5.3.1 Comparison with Base RL Algorithm. The effectiveness of TPAE is evidenced by its superior training dynamics and downstream performance compared with standard GRPO and DAPO implementations. In Figure 4, we display the training dynamics based on accuracy rewards on ViRL39K. The integration of TPAE yields accelerated initial convergence; specifically, models equipped with TPAE consistently surpass the peak accuracy of base RL algorithms while maintaining identical training configurations, including the training dataset, rollout space, and reward functions.

Table 1: Main results (avg 8 acc %) across seven multimodal reasoning benchmarks. All evaluations utilize exact-match scoring on verifiable instances to ensure objective results, avoiding any LLM-as-a-judge. All comparative baselines are instantiated from the Qwen2.5-VL-7B backbone. Best performance of 7B models is marked with bold, second best with underline.

Models	Mathematical & Geometric					Logical	General	Avg.
	MathVerse	We-Math	MathVision	DynaMath	Geo3k	LogicVista	MMMU-Pro	
ThinkLite-VL-7B	42.50	65.17	27.34	47.44	37.77	39.15	28.53	41.13
OpenVLThinker-7B	41.39	65.63	25.90	51.41	38.60	43.85	32.33	42.73
NoisyRollout-7B	39.06	63.76	24.15	53.86	42.80	46.25	35.13	43.57
MM-Eureka-7B	44.56	64.18	27.89	49.80	39.62	47.32	32.11	43.64
Perception-R1-7B	42.99	69.05	25.42	52.84	45.19	44.32	34.36	44.88
VL-Rethinker-7B	44.80	67.70	29.86	52.26	39.89	45.64	36.44	45.23
R1-ShareVL-7B	44.49	69.81	27.87	53.24	42.62	46.81	34.81	45.66
PAPO-G-7B	44.82	66.79	27.76	52.82	40.25	46.07	36.63	45.02
PAPO-D-7B	45.32	68.30	28.36	55.83	44.11	46.70	36.34	46.42
Shuffle-R1-7B	45.67	69.76	29.24	54.49	<u>47.90</u>	<u>47.54</u>	36.65	47.32
VPPO-7B	<u>46.43</u>	<u>70.01</u>	<u>30.48</u>	<u>57.08</u>	45.63	46.26	37.23	47.59
TPAE-G-Qwen2.5-7B	45.50	69.07	30.02	56.59	46.73	47.20	38.27	<u>47.63</u>
TPAE-D-Qwen2.5-7B	47.15	71.70	30.56	57.39	48.14	48.55	<u>37.45</u>	48.71
TPAE-G-Qwen3-8B	54.61	80.44	47.55	66.88	66.31	59.93	48.61	60.62
TPAE-D-Qwen3-8B	55.14	81.10	48.28	66.70	68.18	61.07	48.95	61.35

Table 2: Performance (avg 8 acc %) comparison of multimodal reinforcement learning with (w/) and without (w/o) TPAE.

Models	MathVerse	We-Math	MathVision	DynaMath	Geo3k	LogicVista	MMMU-Pro	Avg.
Qwen2.5-VL-7B	28.53	46.13	18.51	45.20	36.23	42.51	25.64	34.68
+ GRPO w/o TPAE	42.42	65.88	27.27	51.12	42.12	43.71	34.53	43.86
+ GRPO w/ TPAE	45.50	69.07	30.02	56.59	46.73	47.20	38.27	47.63
+ DAPO w/o TPAE	44.22	66.18	27.14	53.12	43.11	45.86	35.19	44.97
+ DAPO w/ TPAE	47.15	71.70	30.56	57.39	48.14	48.55	37.45	48.71
Qwen3-VL-8B	40.98	65.58	27.15	60.40	53.83	51.48	33.86	47.61
+ GRPO w/o TPAE	52.79	77.32	44.92	64.48	63.39	59.10	46.24	58.32
+ GRPO w/ TPAE	54.61	80.44	47.55	66.88	66.31	59.93	48.61	60.62
+ DAPO w/o TPAE	53.08	78.30	45.44	65.06	64.35	58.98	46.76	58.85
+ DAPO w/ TPAE	55.14	81.10	48.28	66.70	68.18	61.07	48.95	61.35

Table 3: Ablation study on the trust boundary τ based on TPAE-G-Qwen2.5-7B.

Configuration	MathVerse	We-Math	MathVision	DynaMath	Geo3k	LogicVista	MMMU-Pro	Avg.
$\tau = 2.45$	45.15	67.99	29.30	55.72	45.74	46.76	<u>38.09</u>	46.96
$\tau = 3.26$	<u>45.42</u>	<u>68.84</u>	<u>29.80</u>	<u>55.83</u>	46.80	<u>46.84</u>	37.87	<u>47.34</u>
$\tau = 3.03$	45.50	69.07	30.02	56.59	<u>46.73</u>	47.20	38.27	47.63

The integration of TPAE also yields consistent improvements in downstream performance across all benchmarks and model backbones. As shown in Table 2, incorporating TPAE within the DAPO framework leads to average accuracy gains of 3.74% on Qwen2.5-VL-7B and 2.50% on Qwen3-VL-8B compared to their respective base RL baselines. Similarly, TPAE-integrated GRPO optimization achieves improvements of 3.77% on the 7B variant and 2.30% on the 8B variant. Together, these results suggest that our perception-grounded

advantage estimation facilitates a more efficient convergence toward superior reasoning capabilities.

5.3.2 Sensitivity of Trust Boundary τ . We perform a sensitivity analysis on the hyperparameter τ in Eq. 8, which determines the trust region for detecting token-level vision-entropy misalignment relative to the reference distributions. These experiments are conducted on Qwen2.5-VL-7B-Instruct utilizing GRPO as the base RL algorithm. We compare our default $\tau = 3.03$, representing a 99%

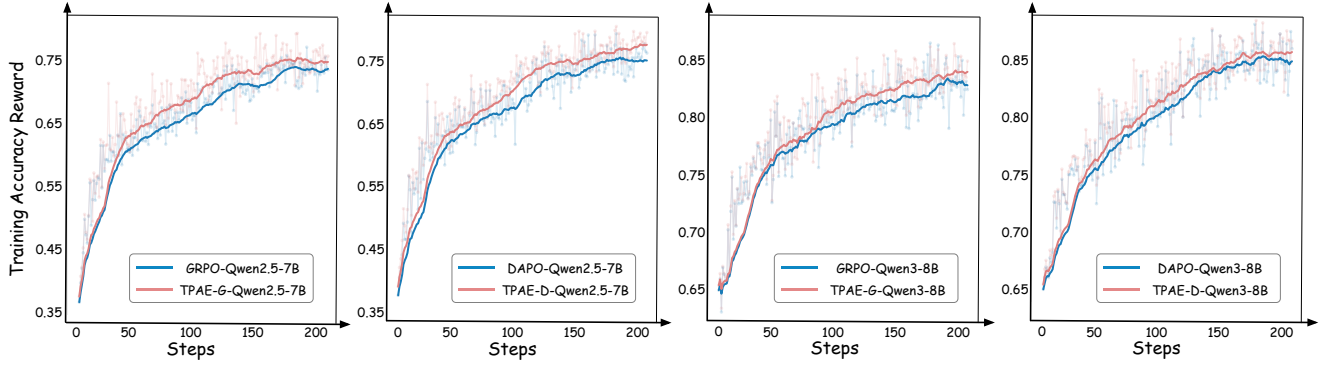


Figure 4: Comparison of RLVR training dynamics on accuracy rewards. Solid lines denote running averages with a window size of 20. TP AE achieves consistently faster learning on both GRPO and DAPO across two model backbones.

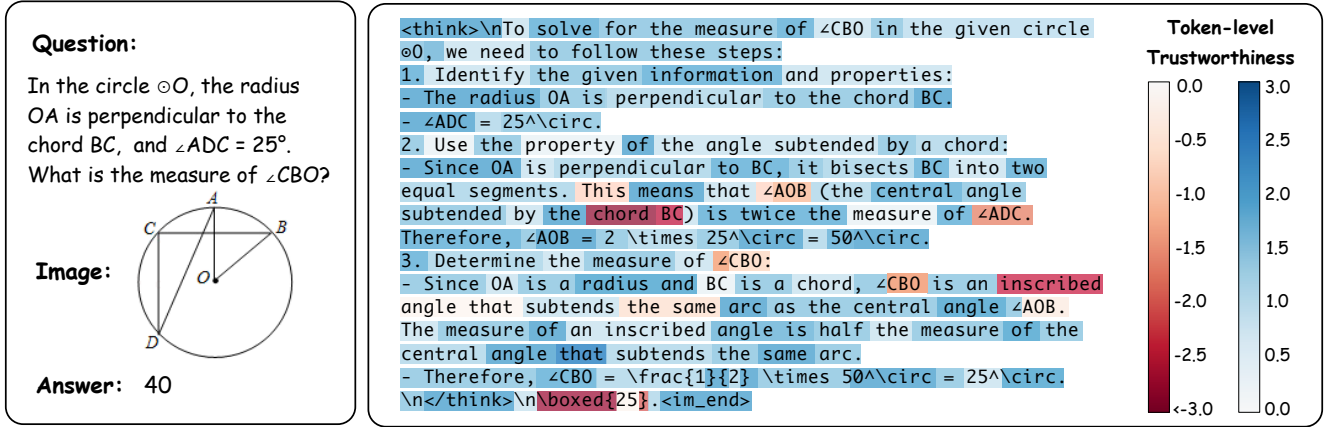


Figure 5: Qualitative analysis of token-level trustworthiness in a failed reasoning trajectory. Dark red highlights successfully localize the pivotal triggers of multimodal reasoning collapse, demonstrating the effectiveness of our TP AE algorithm.

confidence level, against a more permissive $\tau = 2.45$ at 95% confidence and a more stringent $\tau = 3.26$ at 99.5% confidence. As shown in Table 3, the default setting $\tau = 3.03$ yields the highest accuracy on six out of seven benchmarks. While increasing the threshold to $\tau = 3.26$ results in a slight decrease in average accuracy to 47.34%, a more permissive threshold of $\tau = 2.45$ leads to a more notable degradation to 46.96%. These results suggest that a 95% confidence level is insufficiently rigorous for isolating tokens responsible for reasoning failures. Conversely, our choice of $\tau = 3.03$ provides a more effective filtering mechanism that precisely targets tokens responsible for multimodal reasoning failures.

5.4 Qualitative Analysis

We provide a qualitative analysis in Figure 5 to illustrate the token-level trustworthiness defined in Eq. 8. Tokens with high untrustworthiness are highlighted in dark red and are concentrated at visually-hallucinated reasoning steps. The reasoning collapse begins when the model erroneously identifies $\angle AOB$ as the central angle subtending the arc BC . By failing to recognize that $\angle AOB$ actually subtends arc AB , the model establishes an incorrect geometric association with $\angle ADC$. This is compounded when the model incorrectly treats

$\angle CBO$ as an inscribed angle rather than an interior angle of the isosceles triangle $\triangle OBC$. Notably, the final output syntax, including the boxed terminal, also exhibits elevated untrustworthiness [26].

6 Conclusion

In this paper, we pioneer an exploration of multimodal reinforcement learning through the lens of token-level vision-entropy dynamics. Our preliminary study reveals that the joint distribution of visual dependency and predictive entropy establishes an effective indicator for multimodal reasoning collapse. Leveraging this insight, we propose Token-level Perception-grounded Advantage Estimation (TP AE), a novel algorithm that derives fine-grained advantages by statistically aligning a token’s vision-entropy state with the reference distribution observed in correct reasoning chains. TP AE not only establishes a new state-of-the-art across a diverse suite of multimodal benchmarks but also substantially accelerates convergence throughout the RL process. We believe TP AE represents a pioneering step toward perception-aware policy optimization, establishing a theoretically grounded approach for scaling the reasoning capabilities of MLLMs within the RLVR frameworks.

Acknowledgments

This research is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-002). This research is also supported by the Ministry of Education, Singapore, under its AcRF Tier 2 Funding (Proposal ID: T2EP20123-0052). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation or the Ministry of Education, Singapore.

References

- [1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities. *ArXiv abs/2308.12966* (2023). <https://api.semanticscholar.org/CorpusID:263875678>
- [2] Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. 2025. SFT or RL? An Early Investigation into Training R1-Like Reasoning Large Vision-Language Models. *Transactions on Machine Learning Research* (2025). <https://openreview.net/forum?id=wZl5qkQeDF>
- [3] R. De Maesschalck, D. Jouan-Rimbaud, and D.L. Massart. 2000. The Mahalanobis Distance. *Chemometrics and Intelligent Laboratory Systems* 50, 1 (2000), 1–18. doi:10.1016/S0169-7439(99)00047-7
- [4] DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv:2501.12948 [cs.CL]* <https://arxiv.org/abs/2501.12948>
- [5] Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. 2025. OpenVLThinker: Complex Vision-Language Reasoning via Iterative SFT-RL Cycles. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=gFX1nqBKtu>
- [6] Kaixuan Fan, Kaituo Feng, Haoming Lyu, Dongzhan Zhou, and Xiangyu Yue. 2025. SophiaVL-R1: Reinforcing MLLMs Reasoning with Thinking Reward. *arXiv:2505.17018 [cs.CV]* <https://arxiv.org/abs/2505.17018>
- [7] Yue Fan, Xuehai He, Dijie Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Xinze Guan, and Xin Eric Wang. 2025. GRIT: Teaching MLLMs to Think with Images. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=RW3cHh3HgY>
- [8] Siyuan Huang, Xiaoye Qu, Yafu Li, Yun Luo, Zefeng He, Daizong Liu, and Yu Cheng. 2026. Spotlight on Token Perception for Multimodal Reinforcement Learning. In *The Fourteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=bRA4lVJVQ>
- [9] Wenxuan Huang, Bohan Jia, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2026. Vision-R1: Incentivizing Reasoning Capability in Multimodal Large Language Models. In *The Fourteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=UzljskfbfU>
- [10] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. 2026. Vision-R1: Incentivizing Reasoning Capability in Multimodal Large Language Models. *arXiv:2503.06749 [cs.CV]* <https://arxiv.org/abs/2503.06749>
- [11] Olivier Ledoit and Michael Wolf. 2004. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis* 88 (2004), 365–411. <https://api.semanticscholar.org/CorpusID:17238653>
- [12] Sicong Leng, Jing Wang, Jiayi Li, Hao Zhang, Zhiqiang Hu, Boqiang Zhang, Yuming Jiang, Hang Zhang, Xin Li, Lidong Bing, Deli Zhao, Wei Lu, Yu Rong, Aixin Sun, and Shijian Lu. 2025. MMRI: Enhancing Multimodal Reasoning with Variance-Aware Sampling and Open Resources. *arXiv:2509.21268 [cs.CV]* <https://arxiv.org/abs/2509.21268>
- [13] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning*. <https://api.semanticscholar.org/CorpusID:246411402>
- [14] Zicheng Lin, Tian Liang, Jiahao Xu, Qiuzhi Liu, Xing Wang, Ruilin Luo, Chufan Shi, Siheng Li, Yujie Yang, and Zhaopeng Tu. 2025. Critical Tokens Matter: Token-Level Contrastive Estimation Enhances LLM’s Reasoning Capability. In *Forty-second International Conference on Machine Learning*. <https://openreview.net/forum?id=fmz1g18Edl>
- [15] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. 2023. SPHINX: The Joint Mixing of Weights, Tasks, and Visual Embeddings for Multi-modal Large Language Models. *arXiv:2311.07575 [cs.CV]* <https://arxiv.org/abs/2311.07575>
- [16] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=w0H2xGHlkW>
- [17] Xiangyan Liu, Jinjie Ni, Zijian Wu, Chao Du, Longxu Dou, Haonan Wang, Tianyu Pang, and Michael Qizhe Shieh. 2025. NoisyRollout: Reinforcing Visual Reasoning with Data Augmentation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=9zD2iYR0t>
- [18] Yuqi Liu, Tianyuan Qu, Zhisheng Zhong, Bohao PENG, Shu Liu, Bei Yu, and Jiaya Jia. 2026. VisionReasoner: Unified Reasoning-Integrated Visual Perception via Reinforcement Learning. In *The Fourteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=QoDOWjsbAq>
- [19] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. In *International Conference on Learning Representations*, B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (Eds.), Vol. 2024. 23439–23554. https://proceedings.iclr.cc/paper_files/paper/2024/file/663bce02a0050c4a11f1eb8a7f1429d3-Paper-Conference.pdf
- [20] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-GPS: Interpretable Geometry Problem Solving with Formal Language and Symbolic Reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 6774–6786. doi:10.18653/v1/2021.acl-long.528
- [21] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. 2025. MM-Eureka: Exploring the Frontiers of Multimodal Reasoning with Rule-based Reinforcement Learning. *arXiv:2503.07365 [cs.CV]* <https://arxiv.org/abs/2503.07365>
- [22] Runqi Qiao, Quna Tan, Guanting Dong, MinhuiWu MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma GongQue, Shanglin Lei, YiFan Zhang, Zhe Wei, Miaoquan Zhang, Runfeng Qiao, Xiao Zong, Yida Xu, Peiqing Yang, Zhimin Bao, Muxi Diao, Chen Li, and Honggang Zhang. 2025. 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)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 20023–20070. doi:10.18653/v1/2025.acl-long.983
- [23] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *ArXiv abs/1707.06347* (2017). <https://api.semanticscholar.org/CorpusID:28695052>
- [24] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. DeepSeek-Math: Pushing the Limits of Mathematical Reasoning in Open Language Models. *arXiv:2402.03300 [cs.CL]* <https://arxiv.org/abs/2402.03300>
- [25] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. HybridFlow: A Flexible and Efficient RLHF Framework. In *Proceedings of the Twentieth European Conference on Computer Systems (Rotterdam, Netherlands) (EuroSys '25)*. Association for Computing Machinery, New York, NY, USA, 1279–1297. doi:10.1145/3689031.3696075
- [26] Wei Sun, Wen Yang, Pu Jian, Qianlong Du, Fuwei Cui, Shuo Ren, and Jiajun Zhang. 2025. KTAE: A Model-Free Algorithm to Key-Tokens Advantage Estimation in Mathematical Reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=yqQVRNdMkJ>
- [27] Qwen Team. 2025. Qwen2.5-VL. <https://qwenlm.github.io/blog/qwen2.5-vl/>
- [28] Qwen Team. 2025. Qwen3 Technical Report. *arXiv:2505.09388 [cs.CL]* <https://arxiv.org/abs/2505.09388>
- [29] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Kichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., 87310–87356. https://proceedings.neurips.cc/paper_files/paper/2024/file/9ee3a664ccfeabc0da16ac6f1f1cfe59-Paper-Conference.pdf
- [30] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhui Chen. 2025. VL-Rethinker: Incentivizing Self-Reflection of Vision-Language Models with Reinforcement Learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=4oXzssbVg>
- [31] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with the MATH-vision dataset. In *Proceedings of the 38th International Conference on Neural Information Processing Systems (Vancouver, BC, Canada) (NIPS '24)*. Curran Associates Inc., Red Hook, NY, USA, Article 3014, 75 pages.
- [32] Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. 2025. SoTA with Less: MCTS-Guided Sample Selection for Data-Efficient Visual Reasoning Self-Improvement.

- In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=PHu9xJeAum>
- [33] Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru WANG, Hyeonjeong Ha, Xiushi Chen, Yangyi Chen, Ming Yan, Fei Huang, and Heng Ji. 2026. Perception-Aware Policy Optimization for Multimodal Reasoning. In *The Fourteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=izbBqTL8vb>
- [34] Jiaer Xia, Yuhang Zang, Peng Gao, Sharon Li, and Kaiyang Zhou. 2025. Visionary-R1: Mitigating Shortcuts in Visual Reasoning with Reinforcement Learning. arXiv:2505.14677 [cs.CV] <https://arxiv.org/abs/2505.14677>
- [35] Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. 2025. Perception-R1: Advancing Multimodal Reasoning Capabilities of MLLMs via Visual Perception Reward. arXiv:2506.07218 [cs.LG] <https://arxiv.org/abs/2506.07218>
- [36] Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. 2026. Perception-R1: Advancing Multimodal Reasoning Capabilities of MLLMs via Visual Perception Reward. In *The Fourteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=KttCXdj4w>
- [37] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. LogicVista: Multimodal LLM Logical Reasoning Benchmark in Visual Contexts. arXiv:2407.04973 [cs.AI] <https://arxiv.org/abs/2407.04973>
- [38] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. 2025. R1-Onevision: Advancing Generalized Multimodal Reasoning through Cross-Modal Formalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2376–2385.
- [39] Huanjin Yao, Qixiang Yin, Jingyi Zhang, Min Yang, Yibo Wang, Wenhao Wu, Fei Su, Li Shen, Minghui Qiu, Dacheng Tao, and Jiaxing Huang. 2025. R1-ShareVL: Incentivizing Reasoning Capabilities of Multimodal Large Language Models via Share-GRPO. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=kMPI4XbKC>
- [40] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. 2025. DAPO: An Open-Source LLM Reinforcement Learning System at Scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=2a36EMSSTp>
- [41] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhao Chen, and Graham Neubig. 2025. MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 15134–15186. doi:10.18653/v1/2025.acl-long.736
- [42] Jingyi Zhang, Jiaxing Huang, Huanjin Yao, Shunyu Liu, Xikun Zhang, Shijian Lu, and Dacheng Tao. 2025. R1-VL: Learning to Reason with Multimodal Large Language Models via Step-wise Group Relative Policy Optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 1859–1869.
- [43] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. 2024. MATHVERSE: Does Your Multi-modal LLM Truly See the Diagrams in Visual Math Problems?. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part VIII* (Milan, Italy). Springer-Verlag, Berlin, Heidelberg, 169–186. doi:10.1007/978-3-031-73242-3_10
- [44] Zhihan Zhang, Yixin Cao, and Lizi Liao. 2025. XFinBench: Benchmarking LLMs in Complex Financial Problem Solving and Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 8715–8758. doi:10.18653/v1/2025.findings-acl.457
- [45] Linghao Zhu, Yiran Guan, Dingkan Liang, Jianzhong Ju, Zhenbo Luo, Bin Qin, Jian Luan, Yuliang Liu, and Xiang Bai. 2026. Shuffle-R1: Efficient RL framework for Multimodal Large Language Models via Data-centric Dynamic Shuffle. arXiv:2508.05612 [cs.LG] <https://arxiv.org/abs/2508.05612>
- [46] Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. 2025. DynaMath: A Dynamic Visual Benchmark for Evaluating Mathematical Reasoning Robustness of Vision Language Models. In *International Conference on Learning Representations*, Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (Eds.), Vol. 2025. 48337–48383. https://proceedings.iclr.cc/paper_files/paper/2025/file/78b248ca6f627431bba5029d92be8a3d-Paper-Conference.pdf