Linking Perception, Confidence and Accuracy in MLLMs

Yuetian Du[1]*, Yucheng Wang[1]*, Rongyu Zhang[1], Zhijie Xu[4], Boyu Yang[2],
Ming Kong[1], Jie Liu[3]†, Qiang Zhu[1]†

[1]Zhejiang University [2]Alibaba Group
[3]City University of Hong Kong [4]University of Michigan

Project: https://github.com/anotherbricki/CA-TTS

Abstract

Recent advances in Multi-modal Large Language Models (MLLMs) have predominantly focused on enhancing visual perception to improve accuracy. However, a critical question remains unexplored: ***Do models know when they do not know?*** *Through a probing experiment, we reveal a severe confidence miscalibration problem in MLLMs. To address this, we propose* ***Confidence-Driven Reinforcement Learning (CDRL)****, which uses original-noise image pairs and a novel confidence-based reward to enhance perceptual sensitivity and robustly calibrate the model's confidence. Beyond training benefits, calibrated confidence enables more effective test-time scaling as a free lunch. We further propose* ***Confidence-Aware Test-Time Scaling (CA-TTS)****, which dynamically coordinates Self-Consistency, Self-Reflection, and Visual Self-Check modules guided by confidence signals. An Expert Model acts in multiple roles (e.g., Planner, Critic, Voter) to schedule these modules and provide external verification. Our integrated framework establishes new state-of-the-art results with consistent 8.8% gains across four benchmarks. More ablation studies demonstrate the effectiveness of each module and scaling superiority.*

1. Introduction

> *"Ignorance more frequently begets confidence than does knowledge."*
>
> — Charles Darwin, *The Descent of Man* (1871)

Recent advances [3, 13, 15, 31, 58] in Multimodal Large Language Models (MLLMs) have focused on investigations into visual perception, ranging from optimizing visual data distributions [58] and visual position encodings [15] to refining visual instruction tuning [3]. These efforts share a common goal: better visual perception enable higher accuracy. However, Darwin's paradox reminds us of an equally critical yet under-explored dimension: *Does the model know when they do not know?*

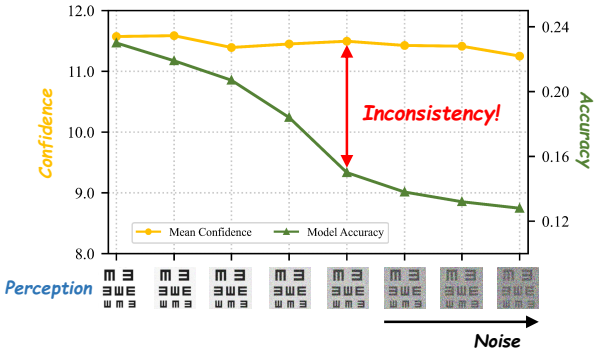

Figure 1. **Disconnection Between Model Confidence and Accuracy Under Perception Degradation.** The X-axis ('Perception') shows the input image with progressively increasing noise. The plot demonstrates that while the ***Mean Confidence*** remain highly stable (insensitive), the ***Model Accuracy*** descends sharply, revealing a ***significant gap*** *between the model's self-reported certainty and its actual performance as the visual input degrades.*

To answer this question, we conduct a key probing experiment (Figure 1). Specifically, we progressively add noise to key visual evidence containing critical information, and measure the model confidence and accuracy. If model truly relies on visual perception, its confidence and accuracy should drop substantially when visual evidence disappears. However, we observe the opposite: confidence remains surprisingly stable despite severe perception degradation. This discrepancy exposes that MLLMs suffer from severe **confidence miscalibration**, which maintaining high confidence even under perception degradation.

This confidence miscalibration problem has been well

*Equal contribution.
†Corresponding Author.

studied in LLMs through uncertainty [17, 53] and logits [10, 18] estimation for each textual output token and obtain satisfactory performance. However, these works fundamentally misalign with MLLMs' visual perception due to granularity mismatch. While LLMs calibrate confidence at *individual token granularity*, visual perception in MLLMs manifests holistically *across the entire response*. To solve this problem, we calculate the confidence for entire response as the mean negative log-probability across all output tokens. Building on this, we propose a **Confidence-Driven Reinforcement Learning (CDRL)** approach that explicitly rewards perception-confidence alignment. This enables MLLMs to develop calibrated confidence that accurately reflects their visual understanding, bridging the gap between what they see and what they claim to know.

A free lunch for this confidence calibration is its direct applicability to test-time scaling with considerable improvement. The calibrated confidence naturally serves as a reliability indicator, enabling models to identify uncertain predictions that warrant additional reasoning effort. To fully leverage this advantage, we further propose **Confidence-Aware Test-Time Scaling (CA-TTS)**, which coordinates three proposed synergistic modules guided by confidence signals. *Self-Consistency* employs confidence-weighted voting combined with expert model calibration to aggregate multiple reasoning paths. *Self-Reflection* leverages expert-generated critiques to refine low-confidence predictions. *Self-Check* validates visual grounding through contrastive decoding between original and noised images. By dynamically routing to appropriate modules based on confidence levels, CA-TTS achieves substantial accuracy gains while maintaining computational efficiency.

Our contributions are summarized as follows:

- We present **the first systematic investigation** of visual perception-aware confidence calibration in MLLMs.
- We propose **Confidence-Driven RL from Vision (CDRL)**, a calibration training method with specialized confidence rewards to enhance perceptual sensitivity.
- We show CDRL's calibrated confidence enables test-time scaling as a free lunch, enabling our **Confidence-Aware Test-Time Scaling (CA-TTS).**
- Our integrated framework achieves new state-of-the-art results, significantly outperforming baselines with consistent **8.8%** overall gain on four benchmarks.

2. Related Work

2.1. Visual Perception Studies of MLLMs

The perceptual capabilities of MLLMs initially benefited from the integration of independently pre-trained models. CLIP [37] achieved alignment between visual and textual representations via contrastive learning. Subsequent models, such as LLaVA [25] and Qwen-VL [2], leveraged more powerful architectures and higher-quality data to enhance visual understanding and instruction-following abilities.

Key factors influencing perceptual capabilities have been identified. In the realm of visual processing, MLLMs demonstrate a tendency to over-focus on a few visual tokens [39]. Regarding in-context learning, they still show difficulty in effectively leveraging visual cues for fine-grained reasoning [23]. Positional encoding methods also present persistent challenges, such as modal confusion and inadequate multi-scale representation [44]. Furthermore, reinforcement learning based on preference optimization has been shown to significantly enhance performance on vision-intensive tasks compared to SFT, while also strengthening the vision encoder's representational capabilities [40].

Inspired by Reinforcement Learning from Human Feedback (RLHF), research has begun to explore its application in MLLM visual perception. Incorporating human preference data has been shown to enhance performance [48] and mitigate hallucinations [11, 56, 59]. For instance, [26] utilized AI feedback to construct reward models, improving output faithfulness. Methods such as DPO have also been employed to capture subtle visual differences [52]. While these approaches have demonstrably improved perceptual abilities, the role of **confidence** has been seldom investigated.

2.2. Calibration for MLLMs

Multi-modal Large Language Models (MLLMs) suffer from severe **systemic miscalibration** when evaluating their own outputs, with overconfidence leading to a significant gap between reported confidence and actual accuracy. Unlike the relatively mature research on LLM calibration [12, 18, 28, 34, 41, 43, 45, 53], work on MLLM calibration is nascent, facing the core challenge of addressing the unique impact of the visual component.

This issue initially gained traction in high-risk domains like clinical diagnostics and autonomous driving [9, 19], where solutions focused on Multi-round Interrogation or RL-Prompting. Furthermore, other research [38] has explored training-free methods, applying them to in-context learning (ICL) for medical image classification. More recently, as research identifies hallucinations as an extreme form of calibration failure, advancements such as visual contrastive decoding [20, 36] and DPO-based alignment [7] have emerged. However, these efforts largely remain incremental extensions of LLM calibration methods. They fail to adequately address how core concepts like visual perception fundamentally impacts calibration outcomes—the key differentiator from LLM calibration.

2.3. Test-Time Scaling Strategies

As enthusiasm for scaling computation during pre-training wanes, Test-Time Scaling (TTS), also known as Test-Time

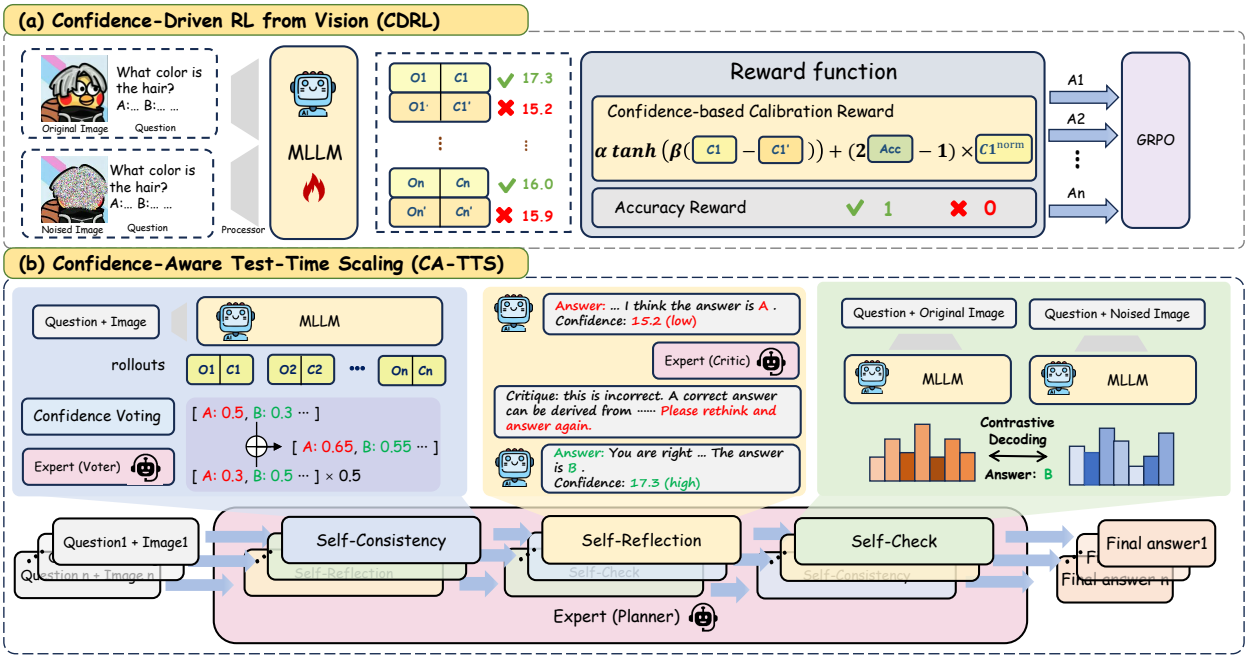

Figure 2. **Framework Overview.** The upper panel **(a)** illustrates how original-noise image pairs are used to optimize the model via Reinforcement Learning, driven by a **Confidence-based Calibration Reward** and a Accuracy Reward. The bottom panel **(b)** shows the adaptive **Confidence-Aware Test-Time Scaling (CA-TTS)** system, where an Expert Model acts as a **Planner, Voter, and Critic** to coordinate the **Self-Consistency, Self-Reflection, and Self-Check** modules, which collaborate to produce the final answer.

Compute, has emerged as a significant research focus [1, 16, 63]. TTS aims to establish a paradigm where increasing computational expenditure during inference yields consistent performance improvements. This approach has gained traction due to its potential for greater generalization and flexibility compared to the standard pre-training and fine-tuning framework. Early practices, such as CoT-prompting [49], initiated this line of inquiry. Subsequently, TTS research diverged into two primary branches: (1) **Parallel Scaling**, represented by methods like Self-Consistency [21], which combines multi-sampling and majority voting; and (2) **Sequential Scaling**, exemplified by approaches such as Self-Refine [32] and STaR [62]. The integration of these two branches has led to the development of tree-structured **Hybrid Scaling** strategies like ToT [57] and MCTS-based methods [24], which enhance policy robustness. Furthermore, large-scale RL-based inference models, spearheaded by Openai-o1 [35] and Deepseek-R1 [8], can be viewed as a novel **Internal Scaling** paradigm.

More recently, research has built upon these foundational paradigms to conduct cutting-edge explorations. For instance, TTRL [66] ingeniously adapts these concepts to Test-Time Training (TTT), enhancing a model's ability to learn from unlabeled data. s1 [33] demonstrated that TTS techniques could achieve performance comparable to large models trained on massive datasets using only about 1k samples. Highlighting its potential, other work has shown a 3B model outperforming a 405B model through TTS [27]. Deepconf [10] also achieved substantial improvements in mathematical reasoning solely by using confidence scores for TTS. Concurrently, efforts have begun to investigate the potential of TTS in the multimodal domain. For example, researchers have proposed Test-Time Reranking (TTR) [14], which uses expert model confidence to refine the original model's probability distribution. However, robust Multimodal TTS that fundamentally addresses the role of visual components and framework robustness remains a significant research gap.

3. Method

3.1. Framework Overview

To address the Perceptual Bluntness Problem in visual reasoning, we propose an innovative framework (Figure 2) built upon two core components. **Confidence-Driven RL from Vision (CDRL)** enhances perceptual sensitivity and calibrates confidence using GRPO and original-noise image pairs. Based on the calibrated confidence, **Confidence-Aware Test-Time Scaling (CA-TTS)** employs an adaptive strategy where an Expert Planner coordinates multiple decoupled reasoning modules to ensure a robust, final answer.

3.2. Confidence-Driven RL from Vision (CDRL)

3.2.1. Preliminaries

High-Quality Data Filtering. To construct the high-quality training set D_{RL}, we first aggregate D_{source} from six public benchmarks: three for mathematical reasoning[51, 64, 65] and three general-purpose VQA datasets [29, 50, 61]. We use an LLM-based pipeline to filter this pool for quality, difficulty, and diversity, yielding D_{Filtered} with 1936 data. Detailed information about filtering procedure and dataset can be found in Appendix B.

Noised Image Generation. To sensitize the model to perturbations, we augment D_{Filtered} by using CLIP attention maps to apply a noise function $\mathcal{G}_{\text{noise}}$, generating a perturbed image i' for each original i. This creates the final training set D_{RL}, containing paired tuples of $((i, i'), q, a)$. This dataset is fundamental for the following resource-friendly GRPO algorithm, which aims to enhance the model's perceptual sensitivity and robust self-calibration capabilities.

Group Relative Policy Optimization. We employ a policy model π_θ and an initial reference model π_{ref}. In each training loop, we sample an image pair (i, i') and question q from D. The policy then generates k candidate output pairs, $(o_1, o_1'), \ldots, (o_k, o_k')$, where o_j is the reasoning trajectory for the original image i. The GRPO algorithm optimizes π_θ by maximizing the objective:

$$J_{GRPO}(\theta) = \mathbb{E}_{(i,i',q)\sim D}\left[\frac{1}{k}\sum_{j=1}^{k} R(i, i', q, o_j)\right] - \beta D_{KL}(\pi_\theta \| \pi_{\text{ref}}). \tag{1}$$

In practice, optimization uses advantage estimation. Each of the k outputs o_j is scored with a total reward r_j. Its advantage A_j—the reward normalized against the group's average—reflects its relative quality. High-advantage paths are then up-sampled, while low-advantage paths are suppressed.

Our total reward score r_j, used for the advantage estimation, comprises three components: the output accuracy reward $R_{\text{Output},j} = \mathbb{I}(\text{GT} \subseteq a_j)$ (where a_j is the final answer, GT is the ground truth, and $\mathbb{I}$ is the indicator function), a formatting reward $R_{\text{Format},j}$ to ensure structural correctness, and our novel **Confidence-based Calibration Reward** $R_{\text{Conf},j}$. The final reward is:

$$r_j = R_{\text{Conf},j} + R_{\text{Output},j} + R_{\text{Format},j} \tag{2}$$

3.2.2. Confidence-based Calibration Reward

To address both perceptual insensitivity and poor confidence calibration, we design the **Confidence-based Calibration Reward**, $R_{\text{Conf},j}$. This reward aims to enhance perceptual sensitivity using the image pair (i, i') and their outputs (o_j, o_j'), while also promoting good confidence calibration.

To compute this reward, we first need to define the model's output confidence C. At each generation step t, we compute the Negative Mean Log-Probability (NMLP) of the model's output logits L. First, the logits L are converted to log-probabilities: $\mathbf{lp} = \text{LogSoftmax}(L)$. We then select the top k highest log-probability values (denoted $\log p_{(i)}$) to define the token's confidence $\text{Conf}_{\text{token}}$. For a complete sequence o composed of T tokens, its total confidence C is the arithmetic mean of all token confidences:

$$C = \frac{1}{T}\sum_{t=1}^{T} \text{Conf}_{\text{token}_t}, \text{ where } \text{Conf}_{\text{token}} = -\frac{1}{k}\sum_{i=1}^{k} \log p_{(i)}. \tag{3}$$

A **lower** NMLP value indicates a **sharper** probability distribution, signifying higher certainty.

Our R_{Conf} reward combines these objectives. The perception goal uses the raw confidence difference $\Delta C = C_j - C_j'$ (from outputs o_j, o_j') to encourage sensitivity. The calibration goal links the binary accuracy $R_{\text{Output},j}$ (which is 0 or 1, defined previously) with the normalized confidence C_j^{norm} from the original image output. This combined mechanism encourages the model to be confident when correct and unconfident when wrong, while simultaneously being sensitive to input perturbations.

The final **Confidence-based Calibration Reward** $R_{\text{Conf},j}$ is formulated as:

$$R_{\text{Conf},j} = \underbrace{\alpha \tanh(\beta * \Delta C)}_{\text{Perception Term}} + \underbrace{(2 \cdot R_{\text{Output},j} - 1) \cdot C_j^{norm}}_{\text{Calibration Term}}, \tag{4}$$

where the first term rewards a large confidence change ΔC and the second term (which simplifies to $+C_j$ if correct and $-C_j$ if incorrect) rewards proper calibration.

Note: To avoid contaminating the training trajectory, the calibration term and its associated gradient updates are only computed based on the output o_j (from the original image) and its confidence C_j, not the perturbed output o_j'.

3.3. Confidence-Aware Test-Time Scaling (CA-TTS)

In this section, we introduce our adaptive multi-module TTS framework. This framework comprises three core modules driven by confidence and enhanced by integrating visual and textual elements: **Self-Consistency, Self-Reflection,** and **Self-Check**. Furthermore, an Expert Model is incorporated to schedule the interactions between these modules or to participate deeply within specific tasks, operating under different designated roles.

3.3.1. Self-Consistency

Instead of merely relying on multi-sampling and majority voting, our self-consistency is based on a confidence-driven, Expert Model-guided approach.

First, for each input sample (Image i, Question q), we collect n samples, gathering their corresponding Chain-of-Thought (CoT), Answer (A), and Confidence (C) sequences:

$$S = \{(CoT_i, A_i, C_i)\}_{i=1}^{n}. \tag{5}$$

Subsequently, leveraging the reliable confidence and visual perception capabilities obtained in the first stage, and inspired by prior work [10], we employ confidence-weighted majority voting. We experimented with various confidence quantification methods (e.g., tail confidence, min confidence) and found mean confidence to be the most effective.

Specifically, we aggregate the confidence scores for each candidate answer k across all samples. This serves as the aggregation and calibration of the model's internal confidence. The result of this internal voting, $V_{internal}$, is a dictionary mapping candidate options k to their weighted vote counts:

$$V_{internal}[k] = \sum_{i=1}^{n} C_i \cdot \mathbb{I}(A_i = k), \tag{6}$$

where $\mathbb{I}$ is the indicator function (1 if $A_i = k$, 0 otherwise). Next, we combine this internal calibration with an external calibration step. We first extract the set of all unique candidate answers from our initial samples to form a candidate list $L_{candidates}$. We then provide the image i, question q, and this candidate list $L_{candidates}$ to the Expert Model (now in the **Voter** role, the prompt $\mathrm{P}_{\text{voter}}$ is shown in Appendix A). The expert M_{expert}^{Voter} is designated to **verbally** output its internal confidence for each option, producing a normalized confidence list $C_{expert} = [c_1, \ldots, c_{|L|}]$ such that $\sum_j c_j = 1$. We then apply a voting weight τ_1. This weight is multiplied by the expert's normalized confidence c_k for a given answer k. This external vote is added to the normalized internal voting score $V_{internal}^{norm}[k]$ to yield the final vote dictionary V_{final}:

$$V_{final}[k] = V_{internal}^{norm}[k] + \tau_1 \cdot c_k, \quad \forall k \in L_{candidates}. \tag{7}$$

In this manner, the **Voter Expert** participates in the calibration process, providing a secondary adjustment opportunity to the base model's internal calibration, thus making the entire self-consistency process more robust.

3.3.2. Self-Reflection

We employ the Expert Model as a **Critic** (M_{expert}^{Critic}) to generate critiques that guide the base model in reconsidering its initial reasoning. In this phase, we provide the original image i and question q to the expert. Using a specific prompt ($\mathrm{P}_{\text{critique}}$, see Appendix A), the expert is directed to generate a corresponding critique ($Crit$) on the problem.

$$Crit = M_{expert}^{Critic}(i, q, \mathrm{P}_{critique}). \tag{8}$$

After obtaining the $Crit$, we use this critique to prompt the base model (M_{base}) to reconsider its reasoning and generate a new reflected Chain-of-Thought $CoT_{reflect}$ and its corresponding answer $A_{reflect}$:

$$(CoT_{reflect}, A_{reflect}) = M_{base}(i, q, Crit). \tag{9}$$

This reflected answer $A_{reflect}$ is then added to the final vote tally V_{final} by incrementing its score by a weight of τ_2. If $A_{reflect}$ is a new answer not previously in V_{final}, it is first initialized with this score.

3.3.3. Self-Check

This module shifts the focus from text-based checks to self-examination at the visual level. Using the same image pair construction method as in Section 3.2.1, we create an original-noise image pair (i, i') for the test image. As noted in Section 3.2.2, logits can be considered a precursor form of confidence. Therefore, inspired by [20], we utilize a more fundamental, confidence-driven method by applying Visual Contrastive Decoding (VCD) to adjust the output. This process decodes the answer by contrasting the log probabilities of generating answer y from the original image i and the noisy image i':

$$\begin{aligned} \log P_{VCD}(y|i, q) = (1 + \alpha) \cdot \log P_\theta(y|i, q) \\ - \alpha \cdot \log P_\theta(y|i', q), \end{aligned} \tag{10}$$

where α is a hyperparameter controlling the strength of the contrast.

This module does not directly involve the Expert Model. The answer obtained from VCD decoding, A_{check}, is added to the V_{final} dictionary by incrementing its score by a voting weight τ_3.

3.3.4. Expert Planning

In addition to its Voter and Critic roles, the Expert Model also functions as a Planner, $M_{expert}^{Planner}$, responsible for module scheduling. Before inference, the planner analyzes the input (i, q) and outputs a scheduling order π. This order is a permutation of the three modules (Self-Consistency M_{sc}, Self-Reflection M_{sr}, and Self-Check M_{sk}), ensuring that each module is used exactly once.

This adaptive scheduling is feasible because the three modules are fully decoupled and order-insensitive. The final output of every module is simply a contribution to the shared voting dictionary V_{final}. If a module is executed first, it initializes the contributions to an empty V_{final}. This design ensures the flexibility and robustness of our system.

4. Experiments

4.1. Experimental Setup

4.1.1. Model Baselines

Our framework uses Qwen2.5-VL-7B-Instruct [42] as the base model in a two-stage process. First, the **CDRL** train-

Table 1. **Evaluation Results on Visual Reasoning Benchmarks.** Abbreviations: OE (Open-Ended), MC (Multi-Choice), PE (Perception), RE (Reasoning), STEM (Science, Technology, Engineering, and Mathematics), HASS (Humanities, Arts, and Social Sciences), and ALL (Overall). Best results are **bold**.

Model/Dataset	Math Reasoning						General VQA					
	Math-Vista$_{testmini}$			Math-Vision$_{test}$			MMStar$_{test}$			MMMU$_{val}$		
	OE	MC	ALL	OE	MC	ALL	PE	RE	ALL	STEM	HASS	ALL
Training Free Baselines (Qwen2.5-VL-7B)												
Pass@1	62.3	66.8	64.7	24.2	21.7	23.0	58.4	68.2	60.2	40.1	60.7	48.8
Majority Voting	68.4	73.7	69.8	26.2	33.2	30.1	63.6	77.0	69.0	53.2	62.5	57.5
Deepconf	69.3	74.7	70.7	26.4	32.3	29.6	58.2	71.0	61.1	51.5	62.5	56.1
Training Based Framework												
DreamPRM(InternVL-2.5-8B)	-	-	68.9	-	-	22.1	-	-	62.3	-	-	61.4
R1-Onevision(Qwen2.5-VL-7B)	61.7	66.2	64.1	20.8	37.8	29.9	60.8	66.9	63.7	51.6	59.3	55.3
VL-Rethinker(Qwen2.5-VL-7B)	65.3	81.7	74.1	22.1	39.0	30.7	60.7	66.5	63.4	51.6	60.7	55.6
We-Think(Qwen2.5-VL-7B)	65.7	80.1	73.3	20.9	37.4	29.7	62.6	70.1	65.1	50.8	62.5	55.7
Ours (Qwen2.5-VL-7B)	**74.2**	**84.7**	**79.5**	**38.4**	**45.6**	**42.4**	**67.8**	**77.2**	**71.3**	**59.9**	**73.5**	**66.3**

ing stage enhances the model's perceptual sensitivity. Second, the **CA-TTS** inference stage uses this trained model as the reasoning agent. In the **CA-TTS** framework, we employ Gemini-2.5-Pro [6] as the Expert Model. Its role is to adaptively schedule the three core TTS modules and provide verification feedback.

4.1.2. Evaluation Benchmarks

We evaluated our framework's effectiveness and robustness on key benchmarks covering image-based mathematical reasoning and general multimodal reasoning. The benchmarks include:

- **Math-Vista [30]:** A comprehensive mathematical vision reasoning benchmark, integrating 28 existing multimodal datasets and 3 newly created ones.
- **Math-Vision [47]:** A high-quality multimodal math reasoning benchmark containing 3040 samples, spanning 16 different mathematical disciplines.
- **MMStar [5]:** A vision-indispensable general-purpose multimodal benchmark, containing 1500 meticulously human-curated samples.
- **MMMU [60]:** A massive multidisciplinary multimodal benchmark designed to evaluate model performance across 30 subject areas and 183 subfields.

4.1.3. Baselines

We compare our method (**CDRL + CA-TTS**) against two categories of baselines to validate its superiority:

1. **Training-Free Baselines:** Includes various training-free Test-Time Scaling (TTS) strategies, such as Majority Voting [22] and Deepconf [10]. To ensure a fair comparison, all Training-Free methods were reproduced on the same base model (Qwen2.5-VL-7B-Instruct).
2. **Training Baselines:** Includes various models trained on visual reasoning tasks, such as DreamPRM [4], R1-Onevision [55], VL-Rethinker [46] and WeThink [54].

Table 2. **Ablation Study for CDRL and CA-TTS.** Best results are **bold**, second-best are underlined.

Setting	Math-Vision$_{test}$		
	OE	MC	ALL
Training-Free	24.24	21.71	22.96
CDRL	18.46	34.16	26.38
CA-TTS	<u>37.99</u>	<u>42.95</u>	<u>37.99</u>
CDRL+CA-TTS	**38.44**	**45.60**	**42.35**

4.2. Implementation Details

CDRL. We performed full-parameter fine-tuning on 8×H100 141GB GPUs using bfloat16 mixed precision and a batch size of 2. We generated 4 rollout pairs per sample, though rollouts from noised-image inputs did not participate in the gradient update.

CA-TTS. We generated 8 parallel inference samples per question, using Temperature $T = 1.0$ and *top-k* $= 40$ to encourage diversity. All module voting weights were set equally ($\tau_1 = \tau_2 = \tau_3 = 0.5$). For the Self-Check module, VCD hyperparameters were set to $\alpha = 0.5$ and $\beta = 0.1$. The Expert Voter was allowed a maximum of 3 retries to ensure reliable confidence output.

4.3. Main Results

a) Superior Performance over Training-Based Baselines. As shown in Table 1, our proposed method (Ours) achieves state-of-the-art performance across all four visual reasoning benchmarks. Specifically, our model (Ours (Qwen2.5-VL-7B)) reaches ALL scores of 79.5%, 42.4%, 71.3%, and 66.3% on Math-Vista, Math-Vision, MMStar, and MMMU, respectively. More importantly, compared to other advanced training-based methods like VL-Rethinker, our model demonstrates stronger performance on all benchmarks—for instance, achieving 0.7% higher on Math-Vista

Table 3. **Performance of Different Expert Models on Math-Vision.** OE (Open-Ended), MC (Multi-Choice). The results on other datasets can be found in Appendix C.

	Math-Vision$_{testmini}$		
Expert Model	OE	MC	ALL
Majority Voting	22.09	31.30	27.65
Qwen-2.5-VL-7B	30.91	34.03	32.57
Qwen-2.5-VL-72B	28.52	37.70	34.21
Qwen-VL-Max	35.71	36.13	35.97
GPT-5	38.94	43.93	41.45
Gemini-2.5-Pro	**46.62**	**42.41**	**43.75**

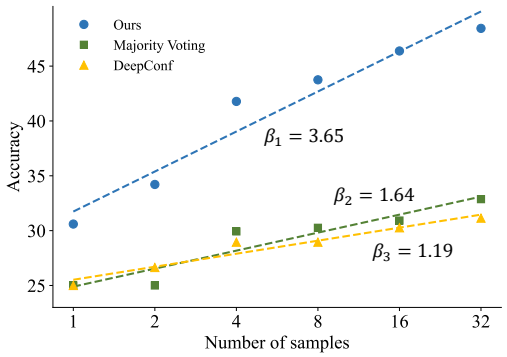

Figure 3. **Test-time scaling comparison on Math-Vision.** Accuracy vs. number of samples for our **CA-TTS** (blue), Majority Voting (green), and DeepConf (yellow). The slope of our method ($\beta_1 = 3.65$) is 2.2-3.1$\times$ steeper than baselines ($\beta_2 = 1.64$, $\beta_3 = 1.19$), demonstrating superior scaling potential with calibrated confidence.

and 0.7% higher on MMMU. This fully demonstrates the effectiveness and superior generalization capability of our proposed framework.

b) Validating the Free Lunch of Calibrated Confidence. This significantly outperforms training-free baselines, including Pass@1 and Majority Voting. This result validates the free lunch concept introduced earlier: the calibrated confidence from our training phase is directly and effectively translated into considerable performance improvements at inference time via our **Confidence-Aware Test-Time Scaling (CA-TTS)** framework.

4.4. Ablation Studies

We conducted a series of ablation studies and conclude four insights as shown in following:

a) CDRL and CA-TTS contribute independently and synergistically. In Table 2, we analyze the contributions of different components. Here, Training-Free corresponds to the Pass@1 baseline from Table 1. (1) Using the **CDRL** alone provides a moderate performance boost over the baseline (e.g., from 48.8% to 52.2% on MMMU). (2) Using the **CA-TTS** alone yields a significant leap in performance (e.g., jumping from 64.7% to 77.8% on Math-Vista). (3) By combining both (**CDRL+CA-TTS**), our full model achieves the best performance across all benchmarks (e.g., 42.4% on Math-Vision). This suggests that the **CDRL** stage provides a better policy or model state for the **CA-TTS** stage, and the combination is most effective.

b) This framework can be generalized to different expert models. To investigate the generalizability and scalability of our framework, we applied it to a series of different expert models. As shown in Table 3, we evaluated performance using Majority Voting as a comparison baseline and found better performance *even Qwen-2.5-VL-7B itself serves as an expert model*. The evaluation also extended to other powerful foundational models, including Qwen-2.5-VL-72B, Qwen-VL-Max, GPT-5, and Gemini-2.5-Pro. The experiments were conducted on the Math-Vision$_{testmini}$ dataset. This diversity of experts aims to verify that our method is robust and not highly dependent on the performance of any single expert model. This further highlights that our framework, particularly through its self-calibration capabilities, can consistently and effectively enhance the reasoning capabilities of models with different scales and architectures.

c) Test-time scaling is enhanced with calibrated confidence. As shown in Figure 3, our **CA-TTS** method demonstrates superior scaling properties compared to Majority Voting and DeepConf baselines. The key advantage is the substantially steeper scaling slope: our method achieves $\beta_1 = 3.65$, which is 2.2$\times$ and 3.1$\times$ higher than Majority Voting ($\beta_2 = 1.64$) and DeepConf ($\beta_3 = 1.19$), respectively. This indicates that **CA-TTS** more effectively leverages additional samples, with the performance gap widening as sample count increases from 1 to 32. While all methods start at similar accuracy with a single sample ($\sim$25-30%), our approach scales to over 45% accuracy, significantly outperforming the baselines' plateau at $\sim$35%. This robust scaling confirms that calibrated confidence enables more efficient test-time computation. Additional scaling results are provided in Appendix C.

d) CDRL enables knowing when visual evidence is insufficient. As shown in Table 4, **CDRL** training significantly enhances the model's perceptual sensitivity across multiple visual uncertainty conditions. The baseline model (Qwen2.5-VL-7B-Instruct) exhibits minimal confidence drops (CD) when visual input is compromised—near-zero or even positive CD values for Occlusion (-0.24), Viewpoint ($+0.09$), and Mosaic ($+0.11$) conditions indicate poor awareness of visual degradation. After **CDRL** training, the model demonstrates substantially larger confidence drops across all perturbation types: Noised (-1.39), Occlusion (-1.13), Viewpoint (-1.29), and Mosaic (-0.86). This represents a 4-8$\times$ enhancement in visual sensitivity on

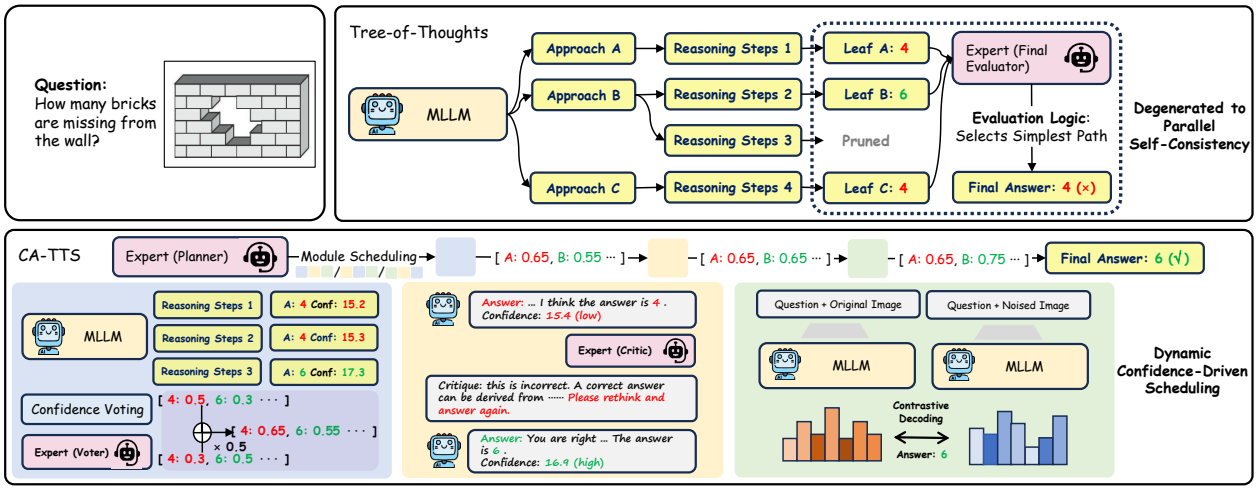

Figure 4. **A case study comparing the reasoning processes of ToT [57] and CA-TTS (Ours).** ToT (upper) conducts a complex tree search that remains vulnerable to a single-point-of-failure in its final evaluation, leading it to the incorrect answer. In contrast, Our method (bottom) demonstrates a multi-stage, resilient process: an initial error from **Self-Consistency** (Answer: 4) is corrected by **Self-Reflection** (Answer: 6) and confirmed by **Self-Check**.

Table 4. **Analysis of model confidence sensitivity under different visual uncertainty conditions on Math-Vision.** We compare the base model with our **CDRL**-trained model across five visual conditions. **CD** (Confidence Drop) measures sensitivity to visual perturbations relative to the original image . **ECE** and **AUC** measure calibration quality. Best results are in **bold**.

	Qwen2.5-VL-7B-Instruct			**+CDRL (Ours)**		
Visual Uncertainty	CD↓	ECE↓	AUC↑	CD↓	ECE↓	AUC↑
Origin	0	64.57	54.81	0	**62.24**	**59.42**
Noised	-0.32	66.19	60.00	**-1.39**	**63.75**	**60.19**
Occlusion	-0.24	66.19	53.88	**-1.13**	**65.05**	**56.34**
Viewpoint	+0.09	64.99	55.19	**-1.29**	**64.18**	**57.26**
Mosaic	+0.11	65.41	60.01	**-0.86**	**63.07**	**61.68**

average. Additionally, **CDRL** improves calibration metrics (ECE) and uncertainty quantification (AUC) across all conditions, demonstrating that the model learns to properly assess perception quality and reduce confidence when visual evidence is insufficient.

4.5. Case Study: CA-TTS vs. ToT

As shown in Figure 4, Our method highlights a robust, decoupled reasoning process. As illustrated, the Expert (Planner) first schedules modules. In the *Self-Consistency* phase, the model may initially converge on an incorrect answer (e.g., 4), even with voter intervention. However, the process continues: in the *Self-Reflection* phase, the expert acts as a *Critique*, providing a new incentive signal that guides the model to correct its answer to 6. This is subsequently solidified in the *Self-Check* phase using visual-contrastive decoding. This demonstrates our method's key advantages: it is robust, features multiple decoupled verification stages, and benefits from continuous and varied incentive signals.

In contrast, tree-based methods like ToT [57], while exploratory, are often more cumbersome and possess a critical vulnerability: a heavy reliance on the performance of a single-pass, final evaluation model. When all paths reach their leaf nodes, this single evaluation determines the outcome. As our comparative example illustrates, a flaw in this single evaluation can cause the entire complex exploration to converge on the wrong answer (i.e., 4). Our approach avoids this single point of failure. As discussed in Section 4.4, our method shows significant gains even with self-evaluation, proving it is less sensitive to the expert model's perfection. Thus, **CA-TTS** offers a more resilient and efficient process through multi-stage validation, unlike ToT's high-stakes dependency on one-shot evaluation.

5. Conclusion

This work identifies *perceptual bluntness* as a root cause of hallucination in MLLMs: a model that reasons before perceiving will inevitably produce unreliable answers. We demonstrate that a synergistic *Perceive-then-Reason* approach is essential. The **CDRL** training stage first instills perceptual calibration, creating the necessary foundation that enables our adaptive **CA-TTS** framework to successfully orchestrate its reasoning modules. Extensive experiments across four challenging benchmarks validate our approach. This finding signals a paradigm shift for MLLM research: future efforts must move beyond text-level preference tuning and co-optimize visual grounding with confidence calibration to build truly robust, self-aware systems that know what they see and when they don't know.

6. Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 42394060 and 42394064, the Earth System Big Data Platform of the School of Earth Sciences, Zhejiang University, and the Zhejiang University - Jolly Pharmaceutical Joint R&D Center for Intelligent Empowerment in Food and Medicine.

References

[1] Mohammad Ali Alomrani, Yingxue Zhang, Derek Li, Qianyi Sun, Soumyasundar Pal, Zhanguang Zhang, Yaochen Hu, Rohan Deepak Ajwani, Antonios Valkanas, Raika Karimi, Peng Cheng, Yunzhou Wang, Pengyi Liao, Hanrui Huang, Bin Wang, Jianye Hao, and Mark Coates. Reasoning on a budget: A survey of adaptive and controllable test-time compute in llms, 2025. 3

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Geng, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and more, 2023. 2

[3] Zhipeng Cai, Ching-Feng Yeh, Hu Xu, Zhuang Liu, Gregory Meyer, Xinjie Lei, Changsheng Zhao, Shang-Wen Li, Vikas Chandra, and Yangyang Shi. Depthlm: Metric depth from vision language models. *arXiv preprint arXiv:2509.25413*, 2025. 1

[4] Qi Cao, Ruiyi Wang, Ruiyi Zhang, Sai Ashish Somayajula, and Pengtao Xie. Dreamprm: Domain-reweighted process reward model for multimodal reasoning. *arXiv preprint arXiv:2505.20241*, 2025. 6

[5] 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. 6

[6] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 6

[7] Alberto Compagnoni, Davide Caffagni, Nicholas Moratelli, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. Mitigating hallucinations in multimodal llms via object-aware preference optimization, 2025. 2

[8] DeepSeek-AI et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. 3

[9] Yuetian Du, Yucheng Wang, Ming Kong, Tian Liang, Qiang Long, Bingdi Chen, and Qiang Zhu. Confidence Calibration for Multimodal LLMs: An Empirical Study through Medical VQA . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*. Springer Nature Switzerland, 2025. 2

[10] Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence. *arXiv preprint arXiv:2508.15260*, 2025. 2, 3, 5, 6, 4

[11] Yuhan Fu, Ruobing Xie, Xingwu Sun, Zhanhui Kang, and Xirong Li. Mitigating hallucination in multimodal large language model via hallucination-targeted direct preference optimization. In *ACL*, 2025. 2

[12] J. Geng, F. Cai, Y. Wang, and et al. A survey of confidence estimation and calibration in large language models. arXiv preprint, 2023. arXiv:2311.08298. 2

[13] Junlin Han, Shengbang Tong, David Fan, Yufan Ren, Koustuv Sinha, Philip Torr, and Filippos Kokkinos. Learning to see before seeing: Demystifying llm visual priors from language pre-training. *arXiv preprint arXiv:2509.26625*, 2025. 1

[14] Hung-Chun Hsu, Yuan-Ching Kuo, Chao-Han Huck Yang, Szu-Wei Fu, Hanrong Ye, Hongxu Yin, Yu-Chiang Frank Wang, Ming-Feng Tsai, and Chuan-Ju Wang. Test-time scaling strategies for generative retrieval in multimodal conversational recommendations, 2025. 3

[15] Jie Huang, Xuejing Liu, Sibo Song, Ruibing Hou, Hong Chang, Junyang Lin, and Shuai Bai. Revisiting multimodal positional encoding in vision-language models. *arXiv preprint arXiv:2510.23095*, 2025. 1

[16] Yixin Ji, Juntao Li, Yang Xiang, Hai Ye, Kaixin Wu, Kai Yao, Jia Xu, Linjian Mo, and Min Zhang. A survey of test-time compute: From intuitive inference to deliberate reasoning, 2025. 3

[17] Ziwei Ji, Lei Yu, Yeskendir Koishekenov, Yejin Bang, Anthony Hartshorn, Alan Schelten, Cheng Zhang, Pascale Fung, and Nicola Cancedda. Calibrating verbal uncertainty as a linear feature to reduce hallucinations, 2025. 2

[18] Zhewei Kang, Xuandong Zhao, and Dawn Song. Scalable best-of-n selection for large language models via self-certainty, 2025. 2

[19] Anita Kriz, Elizabeth Laura Janes, Xing Shen, and Tal Arbel. Prompt4trust: A reinforcement learning prompt augmentation framework for clinically-aligned confidence calibration in multimodal large language models, 2025. 2

[20] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding, 2023. 2, 5

[21] Dacheng Li, Shiyi Cao, Chengkun Cao, Xiuyu Li, Shangyin Tan, Kurt Keutzer, Jiarong Xing, Joseph E. Gonzalez, and Ion Stoica. S*: Test time scaling for code generation, 2025. 3

[22] Yiwei Li, Ji Zhang, Shaoxiong Feng, Peiwen Yuan, Xinglin Wang, Jiayi Shi, Yueqi Zhang, Chuyi Tan, Boyuan Pan, Yao Hu, et al. Revisiting self-consistency from dynamic distributional alignment perspective on answer aggregation. *arXiv preprint arXiv:2502.19830*, 2025. 6

[23] Yue-Bei Li, Yu-Wei Ruan, Zhen-Hua Feng, Yin-Ting Lee, Xiao-Ping Zhang, Kai-Bin Jia, and Wen-Huang Cheng. What do vision-language models see in the context? investigating multimodal in-context learning, 2025. 2

[24] Qingwen Lin, Boyan Xu, Guimin Hu, Zijian Li, Zhifeng Hao, Keli Zhang, and Ruichu Cai. Cmcts: A constrained monte carlo tree search framework for mathematical reasoning in large language model, 2025. 3

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 2

[26] Haotian Liu, Chunyuan Li, and Yong Jae Lee. Llava-1.6: Improved tect-to-image generation with rlaif-v, 2024. Used as a proxy for MLLM RLAIF application. 2

[27] Runze Liu, Junqi Gao, Jian Zhao, Kaiyan Zhang, Xiu Li, Biqing Qi, Wanli Ouyang, and Bowen Zhou. Can 1b llm surpass 405b llm? rethinking compute-optimal test-time scaling, 2025. 3

[28] Y. Liu, Y. Yao, J. F. Ton, and et al. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. arXiv preprint, 2023. arXiv:2308.05374. 2

[29] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024. 4, 1

[30] 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. *arXiv preprint arXiv:2310.02255*, 2023. 6

[31] Jiayun Luo, Wan-Cyuan Fan, Lyuyang Wang, Xiangteng He, Tanzila Rahman, Purang Abolmaesumi, and Leonid Sigal. To sink or not to sink: Visual information pathways in large vision-language models. *arXiv preprint arXiv:2510.08510*, 2025. 1

[32] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023. 3

[33] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling, 2025. 3, 1

[34] S. Ni, K. Bi, J. Guo, and et al. When do LLMs need retrieval augmentation? Mitigating LLMs' overconfidence helps retrieval augmentation. arXiv preprint, 2024. arXiv:2402.11457. 2

[35] OpenAI et al. Openai o1 system card, 2024. 3

[36] Yeji Park, Deokyeong Lee, Junsuk Choe, and Buru Chang. Convis: Contrastive decoding with hallucination visualization for mitigating hallucinations in multimodal large language models, 2024. 2

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2, 1

[38] Xing Shen, Justin Szeto, Mingyang Li, Hengguan Huang, and Tal Arbel. Exposing and mitigating calibration biases and demographic unfairness in mllm few-shot in-context learning for medical image classification, 2025. 2

[39] Zeren Shen, Zuxin Liu, Yichi Zhang, Yifei Ou, Ru-Jia Wei, Haotian Liu, Fang Sun, Tiejun Huang, and Xin Wang. To sink or not to sink: Visual information pathways in large vision-language models, 2025. 2

[40] Junha Song, Sangdoo Yun, Dongyoon Han, Jaegul Choo, and Byeongho Heo. Rl makes mllms see better than sft, 2025. 2

[41] S. Tao, L. Yao, H. Ding, and et al. When to trust llms: Aligning confidence with response quality. arXiv preprint, 2024. arXiv:2404.17287. 2

[42] Qwen Team. Qwen2.5-vl, 2025. 5

[43] K. Tian, E. Mitchell, A. Zhou, and et al. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. arXiv preprint, 2023. arXiv:2305.14975. 2

[44] Aoran Wang, Zhipeng Chen, Ziyang Wang, Yufan He, Tian-Zhu Xiang, and ZENG ZHAO. Revisiting multimodal positional encoding in vision-language models, 2025. 2

[45] C. Wang, G. Szarvas, G. Balazs, and et al. Calibrating verbalized probabilities for large language models. arXiv preprint, 2024. arXiv:2410.06707. 2

[46] Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning, 2025. 6

[47] Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024. 6

[48] Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and et al. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization, 2024. 2

[49] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. 3

[50] xAI Organization. RealworldQA Dataset. https://huggingface.co/datasets/xai-org/RealworldQA, 2024. 4, 1

[51] Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts, 2024. 4, 1

[52] Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *EMNLP*, 2024. 2

[53] M. Xiong, Z. Hu, X. Lu, and et al. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. arXiv preprint, 2023. arXiv:2306.13063. 2

[54] Jie Yang, Feipeng Ma, Zitian Wang, Dacheng Yin, Kang Rong, Fengyun Rao, and Ruimao Zhang. Wethink: Toward general-purpose vision-language reasoning via reinforcement learning, 2025. 6

[55] 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. *arXiv preprint arXiv:2503.10615*, 2025. 6

[56] Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In *CVPR*, 2025. 2

[57] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models, 2023. 3, 8

[58] Keunwoo Peter Yu, Zheyuan Zhang, Fengyuan Hu, Shane Storks, and Joyce Chai. Eliciting in-context learning in vision-language models for videos through curated data distributional properties. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20416–20431, 2024. 1

[59] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *CVPR*, 2024. 2

[60] 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. 6

[61] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark, 2025. 4, 1

[62] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. Star: Bootstrapping reasoning with reasoning, 2022. 3

[63] Qiyuan Zhang, Fuyuan Lyu, Zexu Sun, Lei Wang, Weixu Zhang, Wenyue Hua, Haolun Wu, Zhihan Guo, Yufei Wang, Niklas Muennighoff, Irwin King, Xue Liu, and Chen Ma. A survey on test-time scaling in large language models: What, how, where, and how well?, 2025. 3

[64] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?, 2024. 4, 1

[65] Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models, 2025. 4, 1

[66] Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Yuchen Zhang, Xinwei Long, Ermo Hua, Biqing Qi, Youbang Sun, Zhiyuan Ma, Lifan Yuan, Ning Ding, and Bowen Zhou. Ttrl: Test-time reinforcement learning, 2025. 3