HERO: Enhancing Multimodal Faithfulness via Dynamic Entropy-Aware Reinforcement Learning

Xiongfeng Yang[1,2,3], Qingan Zhang[1], Yuheng Zhang[1,2], Kun-Yu Lin[1,3], Jian-Fang Hu[1,3], Dongmei Jiang[2*], and Wei-Shi Zheng[1,2,3⋆]

[1] School of Computer Science and Engineering, Sun Yat-sen University, China
[2] Peng Cheng Laboratory, Shenzhen, China
[3] Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China
jiangdm@pcl.ac.cn, wszheng@ieee.org

Abstract. Chain-of-Thought fine-tuning enhances the reasoning capabilities of Large Vision-Language Models, but can paradoxically exacerbate hallucinations, revealing a reasoning-faithfulness trade-off. We identify a key cause of this phenomenon as the "Confidence Trap": under visual ambiguity or degradation, models may produce low-entropy, high-confidence errors instead of expressing appropriate uncertainty. This failure mode challenges conventional mitigation strategies that rely on uncertainty cues or contrastive decoding signals, since confident hallucinations can remain highly stable under visual perturbations. To address this problem, we propose Hallucination-Entropy Regulated Optimization (HERO), an on-policy reinforcement learning framework for mitigating confident hallucinations. HERO introduces a dynamic entropy-aware objective that assigns larger optimization weights to low-entropy erroneous outputs, and further improves training efficiency through variance-gated sample selection that focuses updates on informative hard negatives. Experiments on POPE, THRONE, and AMBER show that HERO achieves strong overall hallucination mitigation, substantially improves truthfulness in critical low-entropy regions, and maintains competitive general reasoning capabilities. These results suggest that explicitly targeting confidence-misaligned errors is an effective direction for improving multimodal faithfulness in CoT-style LVLMs.

1 Introduction

Large Vision-Language Models (LVLMs) have demonstrated unprecedented capabilities, effortlessly parsing complex visual scenes, engaging in nuanced multimodal dialogue, and executing sophisticated zero-shot reasoning [1, 2, 26, 29, 44, 48, 54]. However, this remarkable progress is shadowed by a persistent and critical failure mode that severely undermines their practical utility: hallucination [3,12,28,40,51]. This phenomenon manifests as the generation of text that is plausible-sounding yet factually inconsistent with, or entirely untethered from,

⋆ Corresponding author

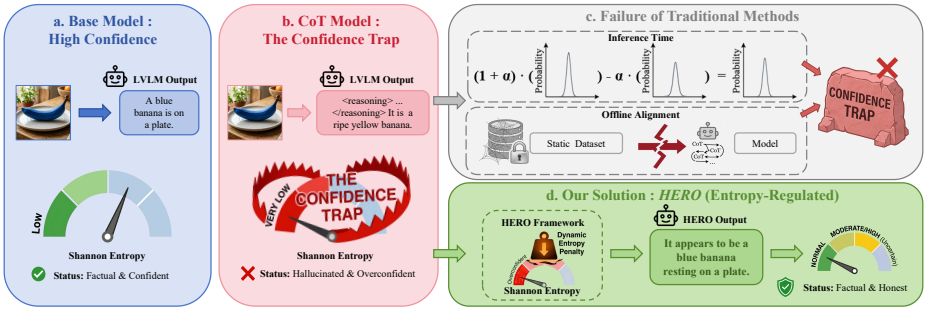

Fig. 1: The Confidence Trap in LVLMs. (a) Standard models yield factual predictions with appropriate high confidence. (b) After Chain-of-Thought tuning, models fail to exhibit the expected uncertainty. Instead, they fall into a Confidence Trap by relying heavily on language priors to fabricate hallucinations characterized by pathologically low entropy and high confidence. (c) Failure of Traditional Methods: Neither inference-time contrastive decoding nor offline alignment can resolve this trap, as they rely on the flawed assumption that hallucinations always correlate with high uncertainty. (d) To address this, our HERO framework introduces a dynamic entropy-aware penalty to extricate the model from this trap and strictly realign its predictive confidence with actual visual evidence.

the provided visual input. This fundamental lack of reliability not only erodes user trust but also renders VLMs perilous for deployment in high-stakes applications, such as autonomous navigation or medical image analysis, where factual accuracy is non-negotiable.

Recent research has widely adopted Chain-of-Thought (CoT) prompting or fine-tuning to unlock the reasoning potential of LVLMs for complex tasks [16, 18, 20, 21, 23, 35, 45, 49, 55, 59]. While the prevailing hypothesis posits that explicit, step-by-step reasoning should theoretically promote factual grounding, our empirical investigation reveals a startling and counter-intuitive trade-off: optimizing for complex reasoning often exacerbates hallucinations.

Crucially, through rigorous experiments involving controlled image degradation, we identified a critical bottleneck behind this failure: a phenomenon we term "Overconfidence under Degradation" (Fig. 1(b)). Intuitively, one might expect a model's predictive entropy to rise alongside its uncertainty as visual clarity degrades. Instead, we observe that models retreat into a "Confidence Trap". Standard CoT tuning exacerbates this issue by encouraging the model to generate coherent narratives regardless of input quality. Consequently, when visual evidence degrades, the model acts as a "blind reasoner" that prioritizes linguistic coherence over visual fidelity, leading to high-confidence fabrications. Driven by these strong language priors, the sample distribution paradoxically shifts toward lower entropy ranges, revealing that the most pernicious errors are not instances of random guessing, but rather cases where the model is confidently wrong.

These findings expose a critical limitation in current mitigation strategies [15, 56]. Conventional uncertainty-based methods, such as VCD [19], rely on the

assumption that hallucinations correlate with high entropy. However, our analysis reveals that hallucinations in reasoning models frequently populate low-entropy zones, creating a Confidence Trap that renders such methods ineffective against confident errors (Fig. 1(c)). Furthermore, offline alignment methods [36,52] often rely on highly customized datasets derived from outdated models. Consequently, they may be ineffective for refining stronger models, and their output formats are frequently ill-suited for CoT-based reasoning.

To address this, we propose Hallucination-Entropy Regulated Optimization (HERO), an on-policy reinforcement learning framework. HERO enforces multimodal faithfulness through three core mechanisms: (1) a dynamic entropy-aware loss that amplifies penalties on high-confidence errors; (2) a variance-gated sample selection strategy for efficient hard-negative mining; and (3) a dynamic balancer that stabilizes training by arbitrating between reward maximization and policy regularization. Our main contributions are summarized as follows:

- We identify a low-entropy bottleneck in large vision-language models, revealing that visual degradation paradoxically drives Chain-of-Thought reasoning into a highly confident state reliant on linguistic priors.
- We propose HERO to resolve the reasoning-faithfulness trade-off. By dynamically penalizing confident hallucinations and mining high-variance samples, it effectively overcomes the blind spots of traditional uncertainty-based defenses.
- HERO achieves the strongest overall performance among the evaluated methods on THRONE and POPE, and improves AMBER coverage under detailed CoT-style generations, while exposing a coverage–hallucination trade-off on AMBER rate metrics, notably restoring truthfulness in critical low-entropy zones by over 20% without compromising complex reasoning capabilities.

2 Related Works

2.1 CoT Reasoning and the Faithfulness Trade-off

Chain-of-Thought (CoT) prompting and fine-tuning have become widely used for enhancing reasoning in LLMs and LVLMs [5,6,8,18,34,39,43,56,57]. While CoT is effective for logic-intensive tasks, recent studies show that explicit reasoning may also induce reasoning hallucination, where models generate plausible rationales that support incorrect or visually unsupported answers [4, 9, 24, 27, 30, 40]. In multimodal settings, this issue becomes more severe because the model may rely on language priors instead of visual evidence. We extend this line of analysis by identifying Overconfidence under Degradation: when visual information becomes ambiguous, CoT-tuned LVLMs can produce low-entropy hallucinations with high confidence, making such errors difficult to detect using standard uncertainty cues.

2.2 Inference-time Defense

Inference-time hallucination mitigation methods mainly intervene in decoding or internal representations [31,32]. Visual Contrastive Decoding (VCD) [19] and

related methods such as CDAR [22] contrast logits from original and perturbed visual inputs to reduce language-prior-driven hallucinations. Other training-free methods, including VHD [13] and DeCo [41], detect abnormal attention or hidden-state patterns and adjust generation without model retraining. These methods are effective when hallucinated and grounded states remain separable at inference time. However, under the low-entropy Confidence Trap, both the original and perturbed branches may favor the same hallucinated token, weakening the contrastive signal. Similarly, internal activations may become less discriminative when the model is already confidently biased toward a visually unsupported answer. HERO therefore addresses this failure at the training stage by directly penalizing low-entropy erroneous outputs.

2.3 Alignment via Preference Optimization

Model alignment has evolved from reinforcement learning from human feedback [10,25,52,58] to more efficient direct preference optimization [36]. For multimodal hallucination mitigation, recent methods incorporate visual constraints into preference learning. V-DPO [46] and HDPO [11] use vision-guided rewards to penalize mismatched descriptions, while OPA-DPO [50] emphasizes the importance of maintaining on-policy data distributions. Although these methods reduce general hallucinations, they usually apply similar penalties to different types of errors and do not explicitly distinguish uncertain mistakes from confident hallucinations. As a result, aligned models may still assign high confidence to visually unsupported generations under ambiguous inputs. In contrast, HERO introduces entropy-aware weighting into on-policy RL, assigning larger optimization weights to low-entropy erroneous outputs and thereby targeting the Confidence Trap more directly.

3 Preliminaries

3.1 Group Relative Policy Optimization (GRPO)

To efficiently fine-tune large vision-language models without the computational overhead of an independent critic network, we employ Group Relative Policy Optimization, or GRPO [38]. Unlike standard Proximal Policy Optimization [37], GRPO estimates the advantage of an action by contrasting it against a group of sampled outputs from the same policy rather than relying on a learned value function.

Formally, for a given multimodal prompt x comprising both visual and textual inputs, the policy π_θ samples a group of G outputs $\{o_1, o_2, \ldots, o_G\}$. A reward model computes a scalar reward r_i for each output o_i. The advantage $\hat{A}_i$ is then estimated by normalizing the rewards within the group:

$$\hat{A}_i = \frac{r_i - \text{mean}(\{r_1, \ldots, r_G\})}{\text{std}(\{r_1, \ldots, r_G\}) + \epsilon}, \tag{1}$$

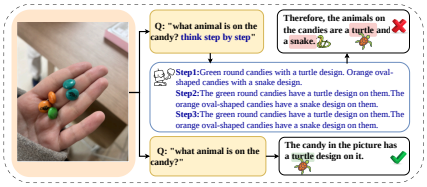

Fig. 2: The Reasoning-Faithfulness Paradox. Introducing explicit Chain-of-Thought paradoxically causes the model to hallucinate unsupported details, degrading factual grounding.

Table 1: Performance Comparison. We evaluate the impact of different CoT strategies. While Tuned-CoT improves intermediate reasoning steps, it exacerbates visual hallucinations, leading to a noticeable decline in overall accuracy relative to the vanilla baseline.

Method	Acc	F1	Yes-Ratio
Vanilla Baseline	**89.6**	**88.8**	**41.9**
+ Prompted-CoT	84.4 (↓5.2)	83.2 (↓5.6)	39.9 (↓2.0)
+ Tuned-CoT	85.8 (↓3.8)	83.6 (↓5.2)	37.1 (↓4.8)

where ϵ is a small constant for numerical stability. The standard GRPO objective maximizes these advantages while constraining policy deviation via Kullback-Leibler divergence:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{x\sim D,\{o_i\}\sim\pi_\theta}\left[\frac{1}{G}\sum_{i=1}^{G}\hat{A}_i\cdot\frac{\pi_\theta(o_i|x)}{\pi_{\theta_{\text{old}}}(o_i|x)}\right] - \beta\cdot\text{KL}(\pi_\theta(\cdot|x)\,||\,\pi_{\text{ref}}(\cdot|x)). \tag{2}$$

Here, π_{ref} acts as the reference policy, typically initialized from the supervised fine-tuned baseline, while β controls the strength of the divergence penalty. Critically, standard GRPO assigns uniform optimization weights to all samples, ignoring the model's inherent predictive confidence. This fundamental limitation directly motivates our entropy-regulated formulation detailed in Sec. 5.

3.2 Natural Language Inference as Reward Proxy

Since obtaining scalar rewards for open-ended visual captioning is non-trivial, we leverage Natural Language Inference, or NLI [7, 14], as a robust proxy for factual correctness. These models categorize the logical relationship between a premise and a hypothesis into three distinct classes: Entailment, Contradiction, and Neutral. In our framework, we decompose ground-truth annotations into atomic visual propositions to serve as the premise, subsequently evaluating whether the generated text, functioning as the hypothesis, entails these facts. This approach allows us to construct a fine-grained, reference-based reward signal without relying on expensive human feedback or black-box LLM scoring.

4 Investigating the Roots of Hallucination

In this section, we dismantle the reasoning-faithfulness trade-off to understand why Chain-of-Thought models exhibit stronger reasoning yet poorer factual grounding. We first confirm the paradoxical effect of these methods. Crucially, we

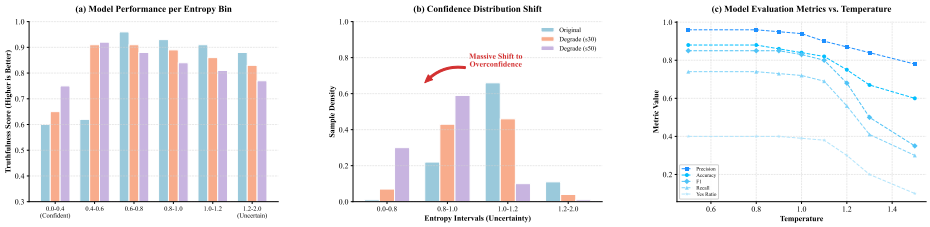

Fig. 3: Unveiling the Confidence Trap. Our analysis across three dimensions highlights the pathology of reasoning models: (a) In the evaluation of truthfulness versus entropy, models maintain faithfulness in high-entropy zones but shift toward blind overconfidence under visual degradation. (b) The confidence distribution shift reveals that as visual noise increases, output density paradoxically migrates toward lower entropy ranges. (c) Sensitivity analysis across temperatures shows consistent performance degradation in accuracy and F1 score, suggesting that this overconfidence is an intrinsic policy trait rather than a heuristic decoding artifact.

then introduce a controlled image degradation experiment revealing a counter-intuitive phenomenon of "Overconfidence under Degradation", an observation that identifies the true root of failure as pathological overconfidence rather than mere uncertainty.

4.1 The Paradoxical Effect of Chain-of-Thought

CoT is designed to enhance reasoning via slow thinking. However, our experiments on the LLaVA-CoT-100k [47] dataset reveal a critical side effect. Quantitatively, as shown in Table 1, while the instruction-tuned variant of this paradigm (Tuned-CoT) is intended to improve reasoning benchmarks, it paradoxically degrades factual consistency. This manifests as lower Accuracy and F1 scores, alongside a decreased Yes-Ratio on POPE and AMBER [42] relative to the vanilla baseline.

This statistical decline is starkly corroborated by our qualitative observations. As illustrated in Fig. 2, forcing the model to generate explicit reasoning steps can inadvertently trigger visual hallucinations that are entirely absent during direct answering. Together, these findings confirm that optimizing for complex, multi-step reasoning often comes at the cost of visual grounding, as the model learns to prioritize plausible narrative generation over strict factual adherence.

4.2 Overconfidence under Degradation

Contrary to the common assumption that hallucinations mainly arise from high uncertainty, our controlled image-degradation experiments reveal a counter-intuitive pattern: visual ambiguity can push CoT-tuned LVLMs toward lower-entropy, high-confidence errors. Specifically, we add Gaussian noise in RGB pixel space

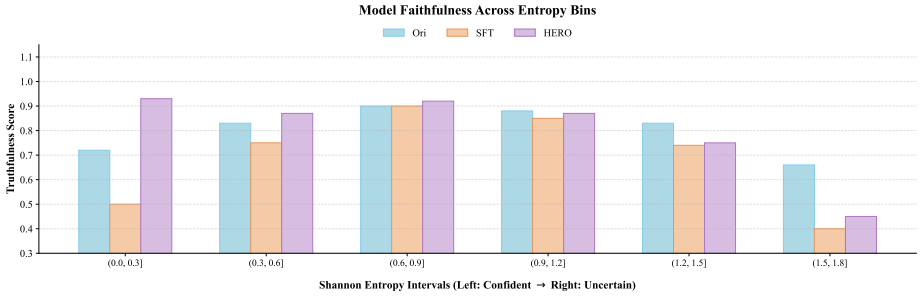

Fig. 4: Truthfulness across entropy bins. Standard CoT-SFT exhibits a sharp drop in the ultra-low-entropy region (Entropy ≤ 0.3), indicating that many of its most confident predictions are factually incorrect. In contrast, HERO substantially improves truthfulness in this critical region while maintaining stable performance across other entropy ranges.

as

$$I_{\text{noise}} = \text{clip}(I + \epsilon, 0, 255), \quad \epsilon \sim \mathcal{N}(0, \sigma^2), \quad \sigma \in \{0, 30, 50\},$$

where $\sigma = 0$ denotes the clean image and larger values correspond to stronger visual degradation.

Intuitively, degrading the visual input should reduce the model's confidence and increase predictive entropy. However, Fig. 3(a) shows that truthfulness drops markedly in low-entropy regions as visual quality decreases. Meanwhile, Fig. 3(b) shows that degradation also shifts more samples into these low-entropy bins. In other words, the model does not merely become uncertain under corrupted visual evidence; instead, it increasingly falls back on linguistic priors and produces confident but visually unsupported predictions. This behavior forms the Confidence Trap: the most harmful hallucinations can appear as high-confidence outputs, thereby weakening the reliability of uncertainty-based detection or mitigation strategies.

To examine whether this failure can be resolved by inference-time adjustment, we further evaluate the sensitivity of faithfulness metrics to sampling temperature. As shown in Fig. 3(c), increasing temperature consistently degrades accuracy and F1 score. This trend suggests that the hallucinations are not merely artifacts of greedy decoding, but are embedded in the learned policy distribution. Therefore, simple decoding heuristics such as temperature scaling are insufficient to restore visual faithfulness. These findings motivate direct policy realignment, where low-entropy erroneous outputs are explicitly penalized during reinforcement learning.

4.3 The Confidence Trap: Why SFT Fails

To further quantify this risk, we analyze the relationship between predictive entropy and truthfulness scores across various tuning stages. As presented in the

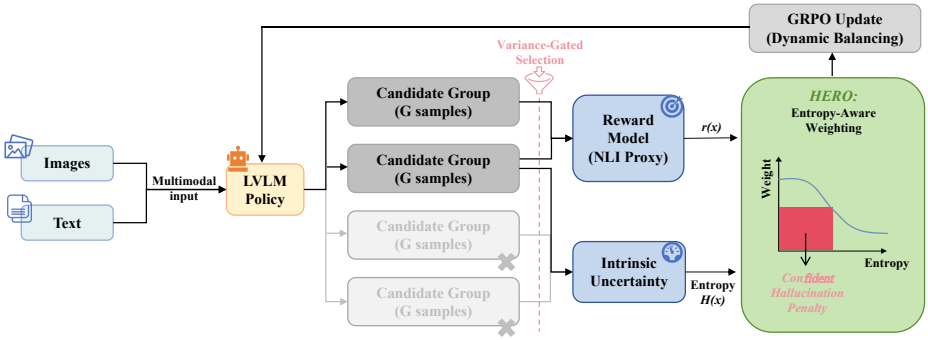

Fig. 5: Overview of the HERO Framework. Given multimodal inputs, the LVLM policy generates candidate groups. We employ Variance-Gated Selection to filter out uninformative samples, retaining active hard negatives. These candidates then undergo a dual evaluation: a Reward Model assesses factual correctness $r(x)$, while the model's Intrinsic Uncertainty yields the predictive entropy $H(x)$. Crucially, the HERO module utilizes an Entropy-Aware Weighting mechanism to modulate the optimization; as illustrated in the central plot, it assigns a severe penalty (red zone) specifically targeting confident hallucinations (low entropy, low reward). The final policy is refined via dynamically balanced GRPO updates.

detailed bin analysis in Fig. 4, we observe that standard supervised fine-tuning on Chain-of-Thought data severely exacerbates the issue. While SFT pushes a significant portion of samples into the high-confidence zone where entropy is below 0.3, the corresponding truthfulness score in this regime drops precipitously compared to our RL-tuned model. This suggests that SFT forces the model to mimic the structural style of reasoning chains, leading to a collapse in entropy even when the underlying reasoning is flawed.

This analysis identifies "Confident Hallucination", a phenomenon characterized by low entropy paired with high error rates, as the most pernicious failure mode. A naive approach of penalizing high entropy proves insufficient because the most dangerous errors are mathematically confident. This observation directly motivates our HERO framework, which specifically targets these low-entropy errors through a dynamic entropy-aware penalty mechanism.

5 Methodology

5.1 Overview of HERO Framework

The HERO framework (Fig. 5) is designed to mitigate the Confidence Trap by penalizing overconfident hallucinations. It operates on a refined subset of prompts $\mathcal{B}$, obtained by variance gating (Sec. 5.2) to retain samples with informative reward variation. The policy is optimized with a dynamic objective that

balances policy improvement and regularization:

$$\begin{aligned}\mathcal{J}_{\mathrm{HERO}}(\theta) = \mathbb{E}_{x\sim\mathcal{B}}\left[p_{\mathrm{offense}}(x)\cdot w(H(x))\cdot \mathcal{J}_{\mathrm{GRPO}}(x)\right] \\ -\beta\cdot\mathbb{E}_{x\sim\mathcal{B}}\left[p_{\mathrm{defense}}(x)\cdot \mathcal{L}_{\mathrm{KL}}(x)\right].\end{aligned} \tag{3}$$

Here, $\mathcal{J}_{\mathrm{GRPO}}$ encourages reward maximization, while the entropy-aware weight $w(H(x))$ (Sec. 5.3) amplifies updates on confident erroneous outputs. The balancing probabilities p_{offense} and p_{defense} (Sec. 5.4) further stabilize training by adaptively shifting between policy improvement and conservative regularization.

5.2 Variance-Gated Sampling

Since post-training LVLMs already possess strong foundational capabilities, standard online RL can spend substantial computation on prompts whose sampled responses provide little learning signal. For a prompt x, we generate G responses and compute the reward variance

$$\sigma_R^2(x) = \mathrm{Var}(r_1,\ldots,r_G). \tag{4}$$

Prompts with extremely low reward variance are treated as plateaued groups and are filtered out before the policy update. This selection is best understood as an efficiency-oriented hard-example mining strategy rather than the primary mechanism for detecting confident hallucinations. It focuses the online updates on prompts where the policy produces diverse outcomes and therefore exposes recoverable decision boundaries. The Confidence Trap itself is addressed by the entropy-aware weighting in Sec. 5.3, which acts on the retained responses and assigns larger weights to low-entropy erroneous outputs. Although variance gating introduces a deliberate sampling bias, our experiments show that it reduces unnecessary backward updates and improves practical training efficiency without degrading final performance.

5.3 Entropy-Aware Weighting: Breaking the Trap

To specifically penalize confident hallucinations, we introduce an entropy-aware weight $w(H(x))$ that dynamically modulates the optimization objective based on the predictive uncertainty of the model. Letting $H(x)$ represent the average token-level Shannon entropy of the response, we define this weight as follows:

$$w(H(x)) = 1 + \sigma\left(-k\cdot(H(x)-H_0)\right). \tag{5}$$

Here, H_0 serves as a robust confidence baseline derived from the empirical mean entropy of the reference model. The hyperparameters k controls the steepness of the transition, while σ denotes the sigmoid function.

This weighting mechanism operates inversely to entropy, precisely regulating the learning process across different confidence regimes. In scenarios where the model is overly confident and its entropy falls significantly below the baseline, the sigmoid function approaches one, driving the weight toward its upper bound.

Consequently, if such a highly confident output yields a low reward due to factual hallucination, the optimization penalty is severely amplified. This strict regularization forces the policy to actively unlearn its blind confidence. Conversely, when the model exhibits high entropy and demonstrates legitimate uncertainty, the weight naturally decays to one. This graceful degradation ensures that the framework safely reverts to standard optimization, preventing appropriate hesitations from being unfairly penalized.

5.4 Dynamic Offensive and Defensive Balancing

We implement a dynamic arbitration mechanism to mitigate reward hacking and catastrophic forgetting, effectively shifting between maximizing reward (*Offense*) and anchoring to the reference policy (*Defense*). This balance is controlled by the group's mean reward $\bar{r}(x) = \frac{1}{G}\sum_{i=1}^{G} r_i$:

$$p_{\text{offense}}(x) = \text{sigmoid}\left(-(\bar{r}(x) - \theta_{\text{base}})\right), \tag{6}$$

$$p_{\text{defense}}(x) = 1 - p_{\text{offense}}(x). \tag{7}$$

Here, θ_{base} is a fixed hyperparameter set to the average reward of the SFT model. The mechanism creates a soft transition between two distinct behaviors: When the model underperforms relative to the baseline ($\bar{r}(x) < \theta_{\text{base}}$), it enters an Offensive Mode ($p_{\text{offense}} \to 1$), aggressively exploring the policy space to find better solutions. Conversely, when performance is satisfactory ($\bar{r}(x) > \theta_{\text{base}}$), it shifts to a Defensive Mode ($p_{\text{defense}} \to 1$), prioritizing KL regularization to consolidate gains and prevent distribution drift.

5.5 Multi-Component Reward Model

We employ a composite reward $R(C)$ that combines Format Compliance (S_{format}) and Accuracy (S_{acc}) to provide a holistic supervision signal:

$$R(C) = S_{\text{obj}} + S_{\text{attr}} + \lambda S_{\text{format}} \tag{8}$$

where S_{acc} is hierarchically decomposed into an object existence score S_{obj} and an attribute score S_{attr}. To compute S_{acc}, we utilize an NLI model f_{NLI} to evaluate the generated caption C. For each ground-truth object o in the object set $\mathcal{O}$, we define an existence indicator $E_o = \mathbb{I}(f_{\text{NLI}}(C, s_o^{\text{obj}}) = \text{Entailment})$ based on its atomic statement s_o^{obj}. The object existence score is simply $S_{\text{obj}} = \frac{1}{|\mathcal{O}|}\sum_{o\in\mathcal{O}} E_o$.

Subsequently, for objects successfully entailed by the caption ($E_o = 1$), we evaluate their corresponding attribute set $\mathcal{A}_o$ using attribute-level statements $s_{o,a}^{\text{attr}}$. The overall attribute score S_{attr} is aggregated strictly across the existing objects:

$$S_{\text{attr}} = \frac{\sum_{o\in\mathcal{O}} \frac{E_o}{|\mathcal{A}_o|} \sum_{a\in\mathcal{A}_o} \mathbb{I}(f_{\text{NLI}}(C, s_{o,a}^{\text{attr}}) = \text{Entailment})}{\sum_{o\in\mathcal{O}} E_o + \epsilon} \tag{9}$$

where ϵ is a negligibly small constant to prevent division by zero. This hierarchical approach resolves ambiguity and ensures a rigorous reward construction.

Table 2: Main results on hallucination benchmarks. HERO improves THRONE/POPE and achieves higher AMBER coverage.

Method	THRONE [17]		POPE [24]			AMBER-G [42]				AMBER-D	
	Acc	$F_{0.5}$	Acc	F_1	$Yes_{\%}$	$CHAIR_{\downarrow}$	$Cover_{\uparrow}$	$Hal_{\downarrow}$	$Cog_{\downarrow}$	Acc	F_1
3B Multimodal LLMs											
Qwen2.5-3B (Base)	78.6	85.7	85.0	83.8	42.8	7.5	58.5	**35.1**	**2.2**	69.8	80.5
+ CoT-SFT [47]	78.3	85.6	84.1	82.8	42.4	8.1	55.6	47.5	3.4	74.6	80.4
+ VCD [19] (Inference)	78.8	86.1	84.8	83.2	42.0	7.8	58.5	46.2	3.2	74.8	79.8
+ HA-DPO [57] (Offline)	78.5	85.5	84.2	82.0	42.5	7.5	57.5	35.5	2.5	73.0	79.5
+ GRPO (Online) [38]	82.2	88.2	85.5	83.8	42.8	7.3	61.5	36.2	2.6	**75.0**	80.2
+ HERO (Ours)	**82.8**	**88.6**	**86.5**	**85.4**	43.4	**7.2**	**62.3**	36.2	2.9	74.4	**80.5**
7B Multimodal LLMs											
Qwen2.5-7B (Base)	77.9	84.8	86.5	84.6	37.5	6.6	60.8	**34.2**	**1.9**	70.3	85.1
+ CoT-SFT [47]	77.3	84.9	85.9	84.2	42.2	6.2	54.1	39.2	2.3	80.1	85.0
+ VCD [19](Inference)	78.5	85.5	87.0	85.0	42.5	6.2	56.5	38.5	2.3	79.8	84.8
+ HA-DPO [57](Offline)	79.5	86.0	87.2	86.0	43.8	6.2	55.5	37.2	2.3	79.5	84.6
+ GRPO [38](Online)	81.9	88.5	89.2	90.0	45.2	**6.0**	59.0	36.8	2.2	80.8	85.3
+ HERO (Ours)	**82.9**	**89.0**	**90.8**	**90.4**	46.0	6.3	**61.2**	37.0	2.1	**81.0**	**85.6**

6 Experiments

6.1 Experimental Settings

Datasets. We utilize a dual-dataset strategy. LLaVA-CoT-100k is employed for Supervised Fine-Tuning (SFT) to instill fundamental reasoning capabilities. For RL alignment, we adopt Visual Attributes in the Wild (VAW), leveraging its fine-grained attribute annotations and real-world complexity to construct a rigorous reward signal.

Baselines. To rigorously validate HERO, we primarily utilize the Qwen2.5-VL family (3B/7B). To compute the reward of each response, we utilize a DeBERTa-v3 NLI model. We compare against a comprehensive suite of methods categorized as follows: (1) Foundations: The pre-trained Base model and its CoT-tuned counterpart (CoT-SFT). (2) Inference-Time Interventions: Uncertainty-based decoding strategies, specifically VCD [19]. (3) Offline Alignment: Standard preference optimization methods HA-DPO [57]. (4) Vanilla GRPO: Standard GRPO devoid of our entropy-regulation and dynamic balancing mechanisms, serving as a direct ablation baseline. Furthermore, we adopt greedy decoding (temperature = 0) for all inference to ensure deterministic outputs.

6.2 Quantitative Analysis

Superiority over Inference-Time Methods. As shown in Table 2, HERO outperforms VCD by a significant margin. VCD relies on the contrastive logit signal between the original image and a perturbed visual input to suppress visually unsupported tokens. However, under the low-entropy Confidence Trap, both

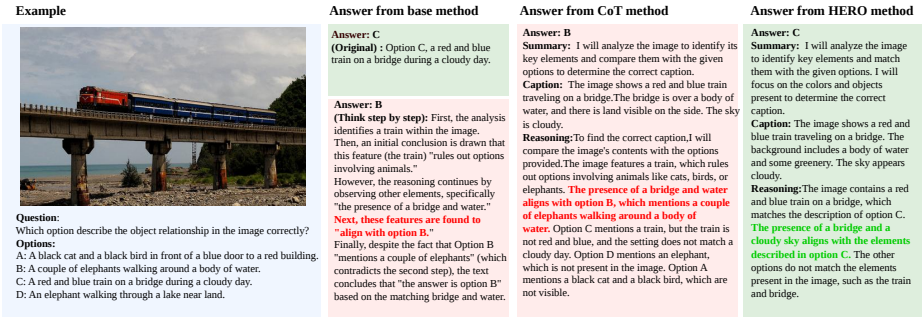

Fig. 6: Qualitative Comparison: Dismantling the Language Prior Trap. Left: The input image displays a train on a bridge. Middle: Baseline and CoT models succumb to language priors, where the context of "water and bridge" triggers a spurious linguistic association with "elephants" (Option B) that overrides visual evidence. Right: HERO suppresses this hallucination by penalizing high-confidence errors, effectively enforcing visual verification of the "red and blue train" to correctly select Option C.

branches can assign high confidence to the same hallucinated token, making the contrastive signal weak or noisy. In contrast, HERO corrects such confident hallucinations during training by explicitly penalizing low-entropy erroneous outputs, thereby addressing failure cases that decoding-time interventions tend to miss.

Superiority over Offline Alignment. HERO also surpasses offline methods like HA-DPO [57]. While offline methods are limited by the static nature of their preference datasets, HERO's on-policy exploration allows it to dynamically discover and correct the model's specific "blind spots" (confident errors) during training, resulting in superior data efficiency and final performance.

AMBER Coverage–Hallucination Trade-off. On AMBER, HERO improves coverage but does not dominate all hallucination-rate metrics. For the 7B model, it slightly increases CHAIR and Hal/sample compared with GRPO, reflecting a coverage–precision trade-off under detailed CoT-style generation. Thus, AMBER results should be interpreted jointly through both coverage and hallucination metrics.

6.3 Dismantling the Confidence Trap

Targeting Confident Hallucinations. To strictly validate whether HERO effectively resolves the Confidence Trap, we analyze the relationship between Truthfulness Score and predictive entropy zones, as shown in Fig. 4. The results reveal a stark contrast in confidence calibration. Specifically, in the high-confidence zone (Entropy ≤ 0.3), the SFT baseline suffers a severe collapse, with truthfulness dropping to approximately 50.0%. This indicates a pathological misalignment where nearly half of the model's most confident assertions are factually incorrect. In contrast, HERO successfully rectifies this pathology, recovering the truthfulness score to over 92.0% within this critical low-entropy

regime. This represents a relative improvement of 84%, empirically demonstrating that our entropy-regulated optimization forces the model to align its high confidence strictly with factual correctness rather than linguistic priors.

Qualitative Analysis: Escaping the Language Prior Trap. To visualize how HERO dismantles these hallucinations, we present a concrete case study in Fig. 6. The Baseline and CoT models often fall into a "Language Prior Trap." For instance, given an image of a bridge over water, the model's internal parametric priors strongly associate these elements with "elephants" (likely due to spurious training data bias). Consequently, despite the unambiguous visual presence of a train, the CoT model overrides the visual evidence and hallucinates an elephant with high confidence. In contrast, HERO successfully suppresses this linguistic hallucination. By explicitly penalizing high-confidence errors during training, the model learns to prioritize visual grounding over probabilistic language associations, correctly identifying the "red and blue train." This confirms that HERO does not merely optimize metrics but fundamentally alters the reasoning mechanism to be more visually faithful.

Table 3: Ablation results. We validate our proposed optimization components and demonstrate the superiority of variance gating over random sampling.

Configuration	THRONE		MMBench
	Acc (%)	F_1 (%)	Acc (%)
Ablation on Optimization Components			
Vanilla GRPO	80.6	77.7	65.5
+ Variance-Gated Sampling	80.6	77.8	65.4
+ Entropy-Aware Weighting	82.0	78.7	65.7
+ Dynamic Balancing	82.8	79.3	65.9
Ablation on Sampling Strategies			
Full Data	80.6	77.7	65.5
Random Sampling	79.2	76.5	64.9
Variance-Gated Sampling	80.6	77.8	65.4

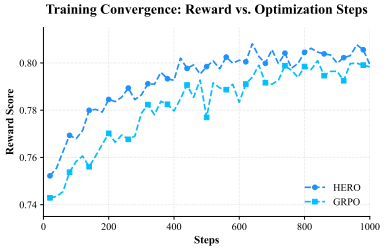

Fig. 7: Training Efficiency. Compared to vanilla GRPO, HERO converges faster and performs better by leveraging variance gating.

6.4 Ablation Study

We verify the contribution of each component in HERO through a comprehensive ablation analysis, as summarized in Table 3, where we report the accuracy on the lowest-entropy 30% of samples. First, we evaluate the impact of entropy-aware weighting. Incorporating this mechanism significantly boosts hallucination metrics, yielding a 2.2% improvement on THRONE compared to the Vanilla GRPO baseline. This performance gain confirms that treating all errors equally is suboptimal, whereas specifically penalizing confident errors produces more effective gradients for hallucination mitigation.

Furthermore, to validate the importance of our hard negative mining strategy, we compare the variance-gated selection against a random sampling approach.

As shown in Table 3, while utilizing random sampling with 60% of the data degrades performance to an accuracy of 79.2%, our variance gating achieves a superior accuracy of 80.6% under the identical budget constraint. This stark contrast demonstrates that non-informative samples offer negligible contribution to the gradient; by filtering them out, the framework reduces training costs while maintaining highly competitive accuracy.

In addition to faster reward convergence, HERO reduces practical training cost under the same GPU allocation: wall-clock time decreases from approximately 22h for vanilla GRPO to 16h for HERO, corresponding to a reduction from 274 to 185 GPU hours. This gain mainly comes from variance-gated selection, which avoids unnecessary backward updates on low-information groups.

6.5 Impact on General Capabilities

While reinforcement learning fine-tuning effectively mitigates hallucinations, it often introduces an "alignment tax" by disrupting pre-existing feature representations, thereby degrading general multimodal understanding. To verify if HERO circumvents this risk, we evaluate our HERO-3b on MMBench [33], MMMU [53] and MME [9] benchmarks.

Table 4: General capability evaluation results on general benchmarks.

Benchmark	Accuracy (%)		
	Base	CoT-SFT	**HERO**
MMBench	63.0	65.6	**65.9**
MMMU	45.22	**46.56**	45.67
MME	66.85	73.04	**73.42**

As shown in Table 4, HERO preserves general multimodal capabilities after hallucination-focused RL. Compared with CoT-SFT, HERO slightly improves MMBench and MME, while remaining close on MMMU. These results suggest that HERO does not introduce a clear alignment tax.

7 Conclusion

In this work, we identify Overconfidence under Degradation as a key factor behind the reasoning-faithfulness trade-off in Chain-of-Thought LVLMs. Our analysis shows that, when visual evidence becomes ambiguous, models can fall into a Confidence Trap, producing low-entropy hallucinations driven by linguistic priors rather than grounded visual evidence. To address this failure mode, we propose HERO, an entropy-regulated on-policy RL framework that emphasizes low-entropy erroneous outputs and improves training efficiency through

variance-gated sample selection. Experiments show that HERO improves hallucination mitigation on THRONE and POPE, enhances truthfulness in low-entropy regions, and improves AMBER coverage under detailed CoT-style generations while exposing a coverage–hallucination trade-off on AMBER rate metrics. Overall, HERO demonstrates that explicitly correcting confident hallucinations can improve multimodal faithfulness without introducing a clear degradation in general reasoning capabilities.

8 Limitations

While HERO mitigates confident hallucinations, several limitations remain. First, the entropy-aware weight relies on a static empirical baseline H_0, whereas predictive uncertainty may vary across tasks and input conditions. Adaptive or input-conditioned baselines may further improve calibration. Second, under extreme domain shifts, some prompts may still produce uniformly confident responses with low reward variance, limiting the effectiveness of variance-gated selection. Finally, our NLI-based reward proxy provides an efficient factuality signal, but atomic text propositions may not fully capture dense spatial relations or fine-grained visual dependencies.

Acknowledgements

This work was supported by New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123100), NSFC (92470202), Guangdong NSF Project (No. 2023B1515040025), Guangdong Key Research and Development Program (No. 2024B0101040004, No. 2025B0909020002).

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., Shou, M.Z.: Hallucination of multimodal large language models: A survey. arXiv preprint arXiv:2404.18930 (2024)
4. Chen, Q., Qin, L., Zhang, J., Chen, Z., Xu, X., Che, W.: M3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. arXiv preprint arXiv:2405.16473 (2024)
5. Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al.: How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. arXiv preprint arXiv:2404.16821 (2024)

6. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24185–24198 (2024)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)
8. Dong, Y., Liu, Z., Sun, H.L., Yang, J., Hu, W., Rao, Y., Liu, Z.: Insight-v: Exploring long-chain visual reasoning with multimodal large language models. arXiv preprint arXiv:2411.14432 (2024)
9. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025)
10. Fu, S., Yang, Q., Li, Y.M., Wei, X., Xie, X., Zheng, W.S.: Love-r1: Advancing long video understanding with an adaptive zoom-in mechanism via multi-step reasoning. arXiv preprint arXiv:2509.24786 (2025)
11. Fu, Y., Xie, R., Sun, X., Kang, Z., Li, X.: Mitigating hallucination in multimodal large language model via hallucination-targeted direct preference optimization. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 16563–16577 (2025)
12. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., et al.: Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14375–14385 (2024)
13. He, J., Zhu, K., Guo, H., Fang, J., Hua, Z., Jia, Y., Tang, M., Chua, T.S., Wang, J.: Cracking the code of hallucination in lvlms with vision-aware head divergence. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3488–3501 (2025)
14. He, P., Gao, J., Chen, W.: Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. arXiv preprint arXiv:2111.09543 (2021)
15. Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., Yu, N.: Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13418–13427 (2024)
16. Huang, W., Jia, B., Zhai, Z., Cao, S., Ye, Z., Zhao, F., Hu, Y., Lin, S.: Vision-r1: Incentivizing reasoning capability in multimodal large language models. arXiv preprint arXiv:2503.06749 (2025)
17. Kaul, P., Li, Z., Yang, H., Dukler, Y., Swaminathan, A., Taylor, C., Soatto, S.: Throne: An object-based hallucination benchmark for the free-form generations of large vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27228–27238 (2024)
18. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. Advances in neural information processing systems **35**, 22199–22213 (2022)

19. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13872–13882 (2024)
20. Li, B., Zhang, K., Zhang, H., Guo, D., Zhang, R., Li, F., Zhang, Y., Liu, Z., Li, C.: Llava-next: Stronger llms supercharge multimodal capabilities in the wild (May 2024), https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/, accessed: 29 June 2026
21. Li, F., Zhang, R., Zhang, H., Zhang, Y., Li, B., Li, W., Ma, Z., Li, C.: Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895 (2024)
22. Li, J., Zhang, J., Jie, Z., Ma, L., Li, G.: Mitigating hallucination for large vision language model by inter-modality correlation calibration decoding. arXiv preprint arXiv:2501.01926 (2025)
23. Li, P., Wang, G., Gu, W., Li, X., Liu, X., Zhang, Y.: An attack detection mechanism in smart contracts based on deep learning and feature fusion. Information Sciences p. 122645 (2025)
24. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355 (2023)
25. Li, Y.M., Yang, Q., Lei, N., Fu, S., Zeng, L.A., Hu, J.F., Wei, X., Zheng, W.S.: Irg-motionllm: Interleaving motion generation, assessment and refinement for text-to-motion generation. arXiv preprint arXiv:2512.10730 (2025)
26. Lin, K.Y., Wang, H., Ren, W., Han, K.: Panoptic captioning: An equivalence bridge for image and text. In: The Thirty-Ninth Annual Conference on Neural Information Processing Systems (2025)
27. Liu, C., Xu, Z., Wei, Q., Wu, J., Zou, J., Wang, X.E., Zhou, Y., Liu, S.: More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. arXiv preprint arXiv:2505.21523 (2025)
28. Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., Peng, W.: A survey on hallucination in large vision-language models. arXiv preprint arXiv:2402.00253 (2024)
29. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
30. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023)
31. Liu, S., Ye, H., Xing, L., Zou, J.: In-context vectors: Making in context learning more effective and controllable through latent space steering. arXiv preprint arXiv:2311.06668 (2023)
32. Liu, S., Ye, H., Zou, J.: Reducing hallucinations in large vision-language models via latent space steering. In: The Thirteenth International Conference on Learning Representations (2024)
33. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)
34. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255 (2023)
35. Luo, R., Zheng, Z., Wang, Y., Yu, Y., Ni, X., Lin, Z., Zeng, J., Yang, Y.: Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. arXiv preprint arXiv:2501.04686 (2025)

36. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems **36**, 53728–53741 (2023)
37. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
38. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
39. Shi, W., Hu, Z., Bin, Y., Liu, J., Yang, Y., Ng, S.K., Bing, L., Lee, R.K.W.: Math-llava: Bootstrapping mathematical reasoning for multimodal large language models (2024), https://arxiv.org/abs/2406.17294
40. Turpin, M., Michael, J., Perez, E., Bowman, S.: Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. Advances in Neural Information Processing Systems **36**, 74952–74965 (2023)
41. Wang, C., Chen, X., Zhang, N., Tian, B., Xu, H., Deng, S., Chen, H.: Mllm can see? dynamic correction decoding for hallucination mitigation. arXiv preprint arXiv:2410.11779 (2024)
42. Wang, J., Wang, Y., Xu, G., Zhang, J., Gu, Y., Jia, H., Wang, J., Xu, H., Yan, M., Zhang, J., et al.: Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. arXiv preprint arXiv:2311.07397 (2023)
43. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models (2023), https://arxiv.org/abs/2201.11903
44. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302 (2024)
45. Xiang, K., Liu, Z., Jiang, Z., Nie, Y., Huang, R., Fan, H., Li, H., Huang, W., Zeng, Y., Han, J., Hong, L., Xu, H., Liang, X.: Atomthink: A slow thinking framework for multimodal mathematical reasoning (2024), https://arxiv.org/abs/2411.11930
46. Xie, Y., Li, G., Xu, X., Kan, M.Y.: V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. arXiv preprint arXiv:2411.02712 (2024)
47. Xu, G., Jin, P., Li, H., Song, Y., Sun, L., Yuan, L.: Llava-cot: Let vision language models reason step-by-step (2024), https://arxiv.org/abs/2411.10440
48. Xu, H., Ye, Q., Yan, M., Shi, Y., Ye, J., Xu, Y., Li, C., Bi, B., Qian, Q., Wang, W., et al.: mplug-2: A modularized multi-modal foundation model across text, image and video. In: International Conference on Machine Learning. pp. 38728–38748. PMLR (2023)
49. Yang, Y., He, X., Pan, H., Jiang, X., Deng, Y., Yang, X., Lu, H., Yin, D., Rao, F., Zhu, M., Zhang, B., Chen, W.: R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. arXiv preprint arXiv:2503.10615 (2025)
50. Yang, Z., Luo, X., Han, D., Xu, Y., Li, D.: Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10610–10620 (2025)
51. Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X., Chen, E.: Woodpecker: Hallucination correction for multimodal large language models. Science China Information Sciences **67**(12), 220105 (2024)
52. Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.T., Sun, M., et al.: Rlhf-v: Towards trustworthy mllms via behavior alignment

from fine-grained correctional human feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13807–13816 (2024)
53. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9556–9567 (2024)
54. Zhang, X., Lu, Y., Wang, W., Yan, A., Yan, J., Qin, L., Wang, H., Yan, X., Wang, W.Y., Petzold, L.R.: Gpt-4v (ision) as a generalist evaluator for vision-language tasks. arXiv preprint arXiv:2311.01361 (2023)
55. Zhang, Y., Wang, G., Li, P., Gu, W., Chen, H.: Frace: Front-running attack classification on ethereum using ensemble learning. The Computer Journal p. bxaf027 (2025)
56. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923 (2023)
57. Zhao, Z., Wang, B., Ouyang, L., Dong, X., Wang, J., He, C.: Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization (2023)
58. Zhao, Z., Wang, B., Ouyang, L., Dong, X., Wang, J., He, C.: Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. arXiv preprint arXiv:2311.16839 (2023)
59. Zheng, H., Xu, T., Sun, H., Pu, S., Chen, R., Sun, L.: Thinking before looking: Improving multimodal llm reasoning via mitigating visual hallucination. arXiv preprint arXiv:2411.12591 (2024)