Token-level Inference-Time Alignment for Vision-Language Models

Kejia Chen[1,2], Junjun Zheng[3], Jiawen Zhang[1,2], Manxi Lin[3], Xiao Pan[5], Jiacong Hu[1,2], Jian Lou[4], Zunlei Feng[1,2,†], Mingli Song[1,2]

[1]The State Key Laboratory of Blockchain and Data Security, Zhejiang University
[2]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security
[3]Taobao & Tmall Group, Alibaba Inc. [4]Sun Yat-sen University [5]Shenzhen University
{chenkejia, kevinzh, zunleifeng}@zju.edu.cn, {fangcheng.zjj, linmanxi.lmx@alibaba-inc.com}

Abstract

Vision-Language Models (VLMs) often prioritize linguistic fluency over visual fidelity, leading to hallucinations where generated text contradicts the image. Countering this bias typically requires resource-heavy fine-tuning or high-latency verification methods that provide feedback only after the full response is generated. To overcome these limitations, we present a framework for **T**oken-level **I**nference-**T**ime **A**lignment (TITA) that steers the decoding process without updating the base model parameters. By training a lightweight reward model to capture visual preferences, TITA extracts implicit guidance through log-probability ratios. This approach functions as an inference-time adaptation of Direct Preference Optimization (DPO), injecting dense feedback to correct the output distribution at every generation step. Across diverse architectures including LLaVA-1.5, Qwen3-VL, and InternVL3.5, TITA consistently improves performance on 13 benchmarks. For example, TITA boosts LLaVA-1.5-7B by +8.6% on MMVet and achieves a 74.0 MMStar score with Qwen3-VL-8B. Specifically, these gains incur negligible overhead (0.2s per query), offering a superior trade-off between alignment effectiveness and efficiency. Our code is available at: https://github.com/Thecommonirin/TITA

1 Introduction

Vision-Language Models (VLMs) have fundamentally reshaped multimodal AI, enabling capabilities ranging from visual question answering (VQA) to complex instruction following by anchoring text generation in visual input (Li et al., 2023c; Liu et al., 2024a; Wu et al., 2024a,c; Yang et al., 2025; Wang et al., 2025). Despite their widespread adoption, VLMs frequently exhibit a critical failure mode: hallucination. These models often produce fluent, coherent text that contradicts the provided visual evidence (Zhao et al., 2023; Bai et al., 2024; Huang et al., 2024; Leng et al., 2024; Zang et al., 2025). Such discrepancies not only degrade generation quality but also introduce substantial safety risks, preventing the deployment of trustworthy multimodal systems in high-stakes environments.

Fundamentally, these hallucinations stem from misalignment during the decoding process: large-scale pretraining instills strong linguistic priors that can override visual grounding, particularly when visual signals are ambiguous or fine-grained (Li et al., 2023a; Zhu et al., 2023; Hurst et al., 2024; Shen et al., 2025). In these scenarios, the model defaults to statistical correlations learned from text data rather than attending to the image, amplifying factual inconsistencies. Consequently, mitigating hallucination requires intervening in this dominance of language priors to restore balance between visual adherence and textual fluency.

Current alignment paradigms attempt to address this trade-off but struggle with limitations in training costs and granularity as illustrated in Figure 1. Training-time alignment methods leverage supervised fine-tuning or reinforcement learning with human or model-based feedback (Xiong et al., 2024; Zhou et al., 2024b; Kapuriya et al., 2024). While effective, this paradigm suffers from inherent rigidity and prohibitive scalability costs. Relying on static parameter updates means that adapting to new domains or refining preference criteria necessitates retraining the entire backbone. This process requires not only massive computational resources but also expensive, curated annotation budgets or proprietary preference labels, severely limiting the accessibility and adaptability of such methods in rapidly evolving multimodal landscapes (Zhao et al., 2024; Favero et al., 2024; Bai et al., 2025).

Inference-time methods offer an alternative by steering frozen VLMs with external reward models (Yan et al., 2024; Zhou et al., 2024c; Liu et al., 2025b). Most operate at the sequence level: they

† Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 25012–25029

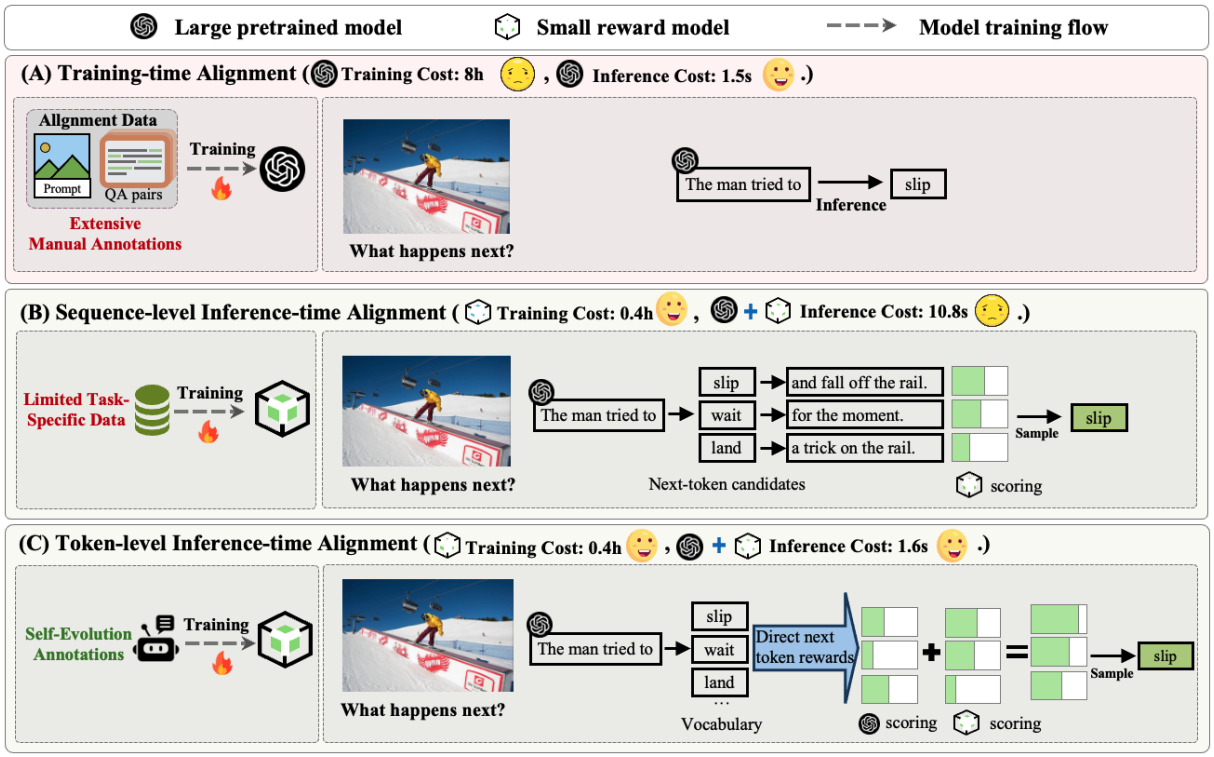

Figure 1: Efficiency Trade-offs across VLM Preference Alignment Paradigms. (A) Training-time alignment fine-tunes base model π_θ with human-labeled preferences. (B) Sequence-level inference-time alignment reranks complete responses with reward models. (C) TITA achieves dual efficiency via token-level decoding guidance.

assign rewards to entire responses, offering only delayed and coarse-grained feedback while incurring heavy overhead from sampling and reranking. This "generate-then-evaluate" mechanism suffers from two critical drawbacks: its feedback is too late to correct errors that manifested early in the decoding trajectory, and its need to sample full sequences create prohibitive computational overhead.

We argue that overcoming these bottlenecks require shifting from delayed, coarse-grained feedback to timely, fine-grained intervention. Since hallucinations often originate from deviations during the decoding trajectory, correction signals should be applied at the token level. Inspired by the duality between reward modeling and language modeling (Fu et al., 2024), we observe that preference information does not require heavy external critics or human annotation. Instead, it can be implicitly captured via the log-probability ratios between a reference model and the target model. This allows a dense, autoregressive guidance that steers the model away from hallucinations step-by-step.

Building on these insights, we introduce TITA (**T**oken-level **I**nference-**T**ime **A**lignment), a framework that transforms sparse sequence-level feedback into dense autoregressive signals. Rather than fine-tuning the base VLM, TITA trains a lightweight reward model to approximate the preferred distribution. During inference, it extracts implicit preference signals as log-probability ratios between the reward model and the target VLM, dynamically steering the decoding process. A token-mapping mechanism ensures compatibility across heterogeneous tokenizers, enabling plug-and-play alignment for off-the-shelf VLMs without modifying their parameters (Figure 1(C)).

This work establishes TITA as a general and principled paradigm for efficient and precise VLM alignment. **Methodologically**, we design a pipeline for constructing preferences via self-supervision. By leveraging augmented visual inputs, we synthesize robust reward signals at the token level, effectively eliminating the reliance on costly human annotations. **Theoretically**, we provide a rigorous proof that this autoregressive formulation approximates the dense reward distribution over sequences, formally bridging the gap between coarse verification of full sequences and granular guidance during decoding (Section A in Appendix). **Empirically**, extensive evaluations across three VLM families and 13 benchmarks demonstrate that TITA consistently reduces hallucinations while preserving base model capabilities and incurring minimal computational overhead (Section 4).

2 Related Work

Hallucination in VLMs. Despite the success of VLMs in multimodal tasks, they suffer from a fundamental misalignment where strong language priors often override visual evidence during generation. This dominance of parametric knowledge leads to hallucinations, defined as generated text that is linguistically fluent but contradicts the visual input (Huang et al., 2024; Leng et al., 2024; Guo et al., 2025a). Such ungrounded outputs compromise factual accuracy and restrict model deployment in safety-critical domains like healthcare and scientific reasoning (Chen et al., 2024a; Wu et al., 2024b; Guo et al., 2025b). Consequently, current research has pivoted toward aligning VLM outputs with human preferences to ensure that generation remains faithful to the provided visual context (Zhang et al., 2025b; Sun et al., 2025).

Preference Alignment in VLMs. Recent efforts aim to align VLMs with human preferences via training-time or inference-time strategies. Training-time alignment involves supervised fine-tuning or reinforcement learning based on human-annotated (Chen et al., 2024c; Guo et al., 2025b; Shen et al., 2025) or model-generated preference data (Ren et al., 2024; Zhang et al., 2025a; Wan et al., 2025). These approaches often yield strong performance but require substantial computational resources and repeated retraining when adapting to new tasks or preferences. Inference-time strategies offer a lightweight alternative by using external reward models to guide the decoding of frozen VLMs. However, most current inference-time methods operate at the sequence level, (Gou et al., 2024; Dong et al., 2025; Sun et al., 2025), assessing response quality only after generating full candidates. This coarse granularity delays error correction and significantly increases inference latency due to the need for multiple sampling passes.

Data Augmentation in VLMs. While data augmentation is traditionally employed in computer vision to enforce representation invariance (Grill et al., 2020; He et al., 2020; Yuan et al., 2024). Rather than viewing this sensitivity merely as a lack of robustness, recent research (Zhu et al., 2024; Awais et al., 2025) repurposes it for alignment. By analyzing divergent outputs triggered by augmentations, these methods identify unstable or hallucinated content, effectively turning augmentation-induced inconsistency into a source of negative preference pairs. This transforms augmentation from a regularization technique into a weak supervision tool for preference mining.

Self-Evolution Strategies. Self-evolution reduces reliance on manual annotation by enabling models to generate their own supervision signals. While techniques such as self-consistency ranking and feedback distillation have shown promise in LLMs (Chen et al., 2024d; Patel et al., 2024; Wang et al., 2024; Ding and Zhang, 2025; Liu et al., 2025b), their application to multimodal settings remains underexplored. The primary challenge lies in establishing reliable verification criteria for visual grounding without external labels. `TITA` addresses this gap by leveraging visual perturbations to construct token-level preference signals automatically. This approach extends self-evolution to VLMs, enabling scalable and efficient alignment that enforces fine-grained consistency between visual inputs and textual outputs.

3 Methods

In response to the inherent tendency of aligned VLMs to develop shallow heuristics rather than principled reasoning, this work present a token-level preference optimization framework that fundamentally rethinks the alignment process.

3.1 Preference Dataset Construction

In preference optimization, the dataset consists of quadruplets $\mathcal{D} = \{(q_n, I_n, y_w^n, y_l^n)\}_{n=1}^N$, where q_n is the input question, I_n is the associated image, y_w is the preferred response, and y_l is the less preferred one. Preferences are modeled with the Bradley–Terry (BT) formulation:

$$p(y_w \succ y_l|q, I) = \frac{\exp(r(q, I, y_w))}{\exp(r(q, I, y_w)) + \exp(r(q, I, y_l))}, \tag{1}$$

where $r(q, I, y)$ is the reward score for response y conditioned on the input (q, I). This formulation naturally captures our intuition that the winning answer should have a higher probability of being preferred, while maintaining a meaningful comparison with the competitive loser.

To construct more informative preference pairs, we leverage the diversity of model outputs generated under multiple image augmentations. Given an input (q, I), we first obtain a baseline response

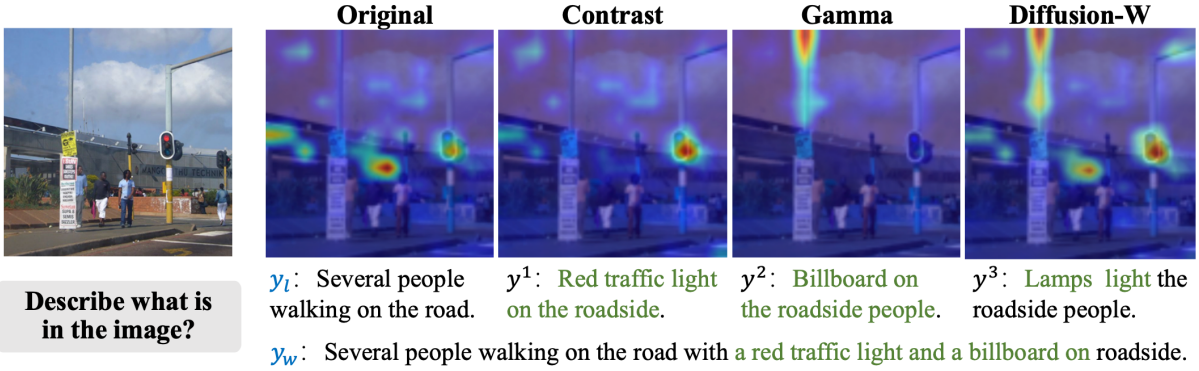

Figure 2: The winner answer y_w is generated by fusing multiple responses obtained from augmented versions of the image, capturing more comprehensive and details compared to the original generation y_l.

from the original image:

$$y_l \leftarrow \pi_\theta(\cdot|q, I), \tag{2}$$

$$\hat{y}^k \leftarrow \pi_\theta(\cdot|q, f_k(I)), \quad k \in [1, ..., K], \tag{3}$$

$$y_w \leftarrow \pi_\theta(\cdot|\hat{y}^1 \| \hat{y}^2 \| \dots \| \hat{y}^K), \tag{4}$$

where f_k denotes the k-th image augmentation method, and y_l serves as the *loser* response. The responses $\{\hat{y}^1, \hat{y}^2 \dots, \hat{y}^K\}$ are concatenated along with a fusion prompt (e.g., "Please provide a comprehensive fusion based on the following candidate answers."), and passed back into the model to generate a unified answer y_w, which serves as the *winner* response. This procedure encourages the model to aggregate diverse visual cues across augmentations, resulting in a more grounded target.

Figure 2 illustrates how different augmentations highlight distinct visual cues and lead to semantically richer descriptions. The fused output captures fine-grained elements (e.g., red traffic light, billboard) that are overlooked in the original response, validating the effectiveness of our augmentation-guided preference construction.

3.2 Token-Level Reward Model

Diverging from standard scalar reward models that evaluate complete sequences, `TITA` employs a generative autoregressive formulation. Instead of outputting a single score, a lightweight VLM (e.g., TinyLLaVA-1.5B) is optimized to assign higher probability mass to preferred tokens, thereby functioning as a dense, token-level reward signal.

Let $y = (y_1, y_2, \dots, y_t)$ denote the output token sequence, where y_t is the token at position t, and $y_{<t}$ is the prefix. Then the autoregressive reward model assigns token-level rewards by modeling the log-likelihood of each token conditioned on the input and its prefix:

$$r(q, I, y) = \sum_t \pi_r(y_t|q, I, y_{<t}), \tag{5}$$

where $\pi_r(y_t|q, I, y_{<t})$ is a learnable distribution function. Generating the next token requires only one forward pass through the target and reward models. This is significantly faster than previous methods (Zhang et al., 2025a) that require generating several candidate tokens, completing the full response for each, and then selecting the best next token. As shown in Table 1, our inference strategy operating at the level of tokens significantly reduces latency. And we demonstrate in Appendix A that this parameterization remains sufficiently expressive to guide target LLMs toward any distribution achievable by traditional reward models within the KL-regularized RL framework.

The reward model is trained by maximizing the likelihood margin between preferred and less preferred tokens, ensuring that the sequence-level rewards align with the preference data:

$$\mathcal{L}(\pi_r; \mathcal{D}_p) = -\mathbb{E}_{\mathcal{D}_p}\Big[\log \sigma\Big(\beta \sum_t \log \pi_r(y_{w,t}|q, I, y_{w,<t}) - \beta \sum_t \log \pi_r(y_{l,t}|q, I, y_{l,<t})\Big)\Big], \tag{6}$$

3.3 Inference-time Guidance

We now describe the auto-regressive inference-time alignment mechanism. In practical scenarios, fine-tuning a smaller, typically weaker language model (e.g., 1B/7B) is often feasible, while fine-tuning a larger, stronger model (e.g., 70B) may be impractical due to resource constraints. By leveraging our proposed auto-regressive reward model, which predicts next-token rewards $\log \pi_r(y_t|q, I, y_{<t})$ in a manner similar to how language models predict

next-token log probabilities, Equation 7 can be interpreted as a form of controlled decoding from multiple models:

$$\log \pi(y|q,I) = -\log Z(q,I) + \sum_t \log \pi_\theta(y_t|q,I,y_{<t}) + \lambda \cdot \sum_t \log \pi_r(y_t|q,I,y_{<[t]}), \tag{7}$$

This formulation allows TITA to apply previous decoding techniques (Dekoninck et al., 2023) to sample the next token y_t, conditioned on the query with image (q, I) and the partially generated response $y_{<t}$, by computing the next-token conditional probability as follows:

$$\pi(y_t|q,I,y_{<t}) \propto \pi_\theta(y_t|q,I,y_{<t})\big(\pi_r(y_t|q,I,y_{<t})\big)^\lambda. \tag{8}$$

A distinct advantage of TITA is the decoupling of the reward model from the target policy. TITA trains the autoregressive reward model without relying on any specific target LLM during the training phase. Unlike standard DPO, which binds alignment to a specific backbone during training, TITA optimizes the reward model independently. Specifically, in this work, a compact autoregressive reward model is employed to steer larger, more powerful target LLMs, enabling scalable weak-to-strong alignment. This design decouples reward modeling from the target generator, fostering diverse and adaptable inference-time applications without the need for target-specific retraining.

The complete inference pipeline is summarized in Algorithm 1 (Appendix B). A unique challenge of this decoupled architecture arises when the small reward model π_r and the target VLM π_θ utilize different tokenizers. To align their probability spaces during the token generation step (Equation 8), we deploy a Top-k cross-encoding mapping strategy. Specifically, we extract the top-k tokens with the highest probabilities from π_r, decode them into text strings, and re-encode them using the target model's tokenizer to assign the corresponding probability mass. Sensitivity analysis reveals that TITA is remarkably robust to the choice of k, with performance fluctuations remaining negligible (± 0.3 on MMVet) for $k \in \{5, 10, 20, 50\}$.

Furthermore, while this cross-vocabulary mapping is stable due to the overlap of high-probability subwords, we recommend using same-family RM-LM pairings (e.g., Qwen-VL-2B guiding Qwen-VL-72B) as a practical best practice to completely bypass tokenization discrepancies.

4 Experiments

4.1 Experimental Setup

Backbones. To evaluate the generality of TITA, we conduct experiments on two distinct VLM categories. (1) Established Research Backbones: We utilize the *LLaVA-1.5* series (Liu et al., 2024a) to ensure fair comparisons with prior hallucination mitigation studies. (2) Contemporary High-Performance Models: To explore scalability to advanced architectures, we extend evaluation to *Qwen3-VL-8B-Instruct* (Yang et al., 2025), *InternVL3.5-8B* (Wang et al., 2025), and the *DeepSeek-VL2* family (Wu et al., 2024c).

For the reward model, we employ the lightweight *TinyLLaVA-1.5B* (Zhou et al., 2024a), trained on preference pairs from OCRVQA (Mishra et al., 2019) and TextVQA (Singh et al., 2019). Training takes only ∼0.4 hours on 8 A100 GPUs (see Appendix C for training details).

Baselines. To situate TITA within the alignment landscape, we compare it with three paradigms of hallucination mitigation.

(1) Training-time Alignment: Methods that internalize human preferences by fine-tuning the base VLM parameters, including *Fact-RLHF* (Sun et al., 2023), *CSR* (Zhou et al., 2024c), and *SeVa* (Zhu et al., 2024). As shown in 1, these methods incur notable computational overhead (7.5–16.4 hours when applied to the *LLaVA-1.5-7B* base model) due to parameter-efficient fine-tuning.

(2) Decoding Heuristics: Training-free methods that modify decoding logits based on priors or noise intervention. We evaluate against *VCD* (Leng et al., 2024), *M3ID* (Favero et al., 2024), and *MARINE* (Zhao et al., 2024) to contrast heuristic-based adjustments with the learned, fine-grained semantic guidance provided by TITA.

(3) Inference-time Alignment: Strategies that employ external critics or iterative self-correction to rerank or refine generated responses. We compare with *Critic-V* (Zhang et al., 2025a), *MM-Verify* (Sun et al., 2025), and *Sherlock* (Ding and Zhang, 2025). While effective, these "System 2" approaches operate at the sequence level, often necessitating multiple generation passes.

Benchmarks. Evaluation is conducted across three dimensions: (1) *Comprehensive Evaluation*: SEED (Li et al., 2023b), LLaVA-Bench (Liu et al., 2024b), MMbench (Liu et al., 2025a), MME (Yin

Table 1: Comparative study on LLaVA-1.5-7B across three alignment paradigms. TITA establishes a new Pareto frontier with minimal training cost. "**Inference Time**" indicates the average latency per query.

Model	MME^P	MME^C	SEED	$LLaVA^W$	MMVet	SQA	GQA	POPE	Optimization	Training Time	Inference Time
Base Model: LLaVA-1.5-7B	1510.7	348.2	58.6	63.4	30.5	66.8	62.0	85.9	-	-	1.5s
Paradigm 1: Training-time Alignment											
Fact-RLHF (Sun et al., 2023)	1490.6	335.0	58.1	63.7	31.4	65.8	61.3	81.5	RLHF	16.4h	1.5s
CSR (Zhou et al., 2024c)	1524.2	367.9	60.3	71.1	33.9	**70.7**	**62.3**	86.8	DPO	6.8h	1.5s
SeVa (Zhu et al., 2024)	1531.0	369.2	65.8	72.2	37.2	67.5	60.7	86.7	DPO	7.5h	1.5s
Paradigm 2: Training-free Decoding Heuristics											
VCD (Leng et al., 2024)	1450.1	354.0	61.7	66.6	32.9	65.4	61.3	86.3	Decoding	-	2.9s
M3ID (Favero et al., 2024)	1436.4	342.8	59.3	64.3	36.2	66.9	61.8	88.0	Decoding	-	2.4s
MARINE(Zhao et al., 2024)	1517.5	360.2	62.4	67.0	38.5	68.4	61.6	90.5	Decoding	-	3.8s
Paradigm 3: Inference-time Alignment											
Critic-V (Zhang et al., 2025a)	1528.4	355.0	63.4	67.8	35.7	66.5	59.4	86.5	DPO	2.9h	7.9s
MM-Verify (Sun et al., 2025)	1505.0	342.7	59.3	67.6	34.2	66.0	58.0	86.2	SFT	4.8h	5.9s
Sherlock (Ding and Zhang, 2025)	1523.0	350.6	61.4	67.5	38.3	69.6	61.7	88.7	DPO	19.0h	21.4s
TITA † (Ours)	**1538.4**	**369.5**	**66.6**	**72.5**	**39.1**	**70.7**	**62.3**	**91.7**	DPO	**0.4h**	**1.6s**

† denotes our token-level method, whereas other inference-time baselines operate at the sequence level.

et al., 2023), MMVet (Yu et al., 2023). (2) *General Visual Question Answering (VQA)*: VizWiz (Gurari et al., 2018), GQA (Hudson and Manning, 2019), ScienceQA (Lu et al., 2022), MMStar (Chen et al., 2024b). (3) *Hallucination Detection*: CHAIR (Rohrbach et al., 2018) and POPE (Li et al., 2023d). To ensure hardware-agnostic latency comparisons, all inference speed and FLOPs measurements are standardized on a single NVIDIA A100 (80GB) GPU using greedy decoding with a batch size of 4.

4.2 Main Results

Efficiency-Effectiveness Trade-off. Table 1 illustrates the performance landscape on the LLaVA-1.5 benchmark, demonstrating that TITA redefines the balance between alignment quality and computational cost. TITA surpasses training-time methods while eliminating the need for expensive parameter updates. Approaches like SeVa demand 7.5 hours of GPU-intensive training, whereas TITA requires only 0.4 hours for reward modeling yet achieves a superior MMVet score of 39.1%.

Crucially, TITA eliminates the latency bottlenecks of inference-time verification. Methods like Critic-V generate and rank full sequences, leading to an average latency of 7.9s per query. By contrast, TITA optimizes the token distribution in real time, adding a negligible $+0.1$s overhead to the base model's speed (1.6s vs 1.5s). Furthermore, as detailed in Table 3, compared to dual-stream decoding heuristics like MARINE (17.36 TFLOPs), TITA is highly compute-efficient (10.54 TFLOPs) since it only requires a single scalar scoring overhead per token rather than redundant forward passes. Independent runs confirm this efficiency is highly stable (e.g., MMVet 39.1 ± 0.28, POPE 91.7 ± 0.31).

Isolating the Guidance Mechanism. To rigorously decouple the quality of our synthesized preference data from the guidance mechanism itself, we compared TITA against standard SFT and DPO using the *exact same* (y_w, y_l) data pairs. As shown in Table 4, TITA (with a frozen base model) successfully matches the performance of DPO across multiple benchmarks. This proves that our performance gains are not merely data dividends; rather, TITA effectively extracts and applies preference signals at inference time, achieving the same alignment quality as intensive weight fine-tuning but at a fraction ($\sim 5\%$) of the computational cost.

Scalability and Weak-to-Strong Generalization. To verify scalability beyond LLaVA, we evaluate TITA on modern high-performance architectures, including Qwen3-VL-8B, InternVL3.5-8B, and DeepSeek-VL2-27B (Table 2). The results indicate that simply scaling the base model does not automatically eliminate entrenched language priors; however, TITA provides consistent gains across all capacities. On the 27B DeepSeek-VL2, it reduces object hallucination ($CHAIR_i$) from 11.7% to 4.9%—a drop of over 58%. Remarkably, these improvements are achieved using a 1.5B reward model steering targets nearly an order of magnitude larger. This weak-to-strong generalization suggests that the visual alignment principles captured by TITA are scale-invariant, offering a highly cost-efficient alignment strategy. Furthermore, early experiments indicate this token-level guidance successfully generalizes to text-only LLMs (detailed in Appendix C.2), showcasing its broad applicability beyond visual domains.

Table 2: Comparative Study of Modern VLM Architectures: TITA Consistently Improves Hallucination Robustness and General Reasoning Across Qwen, InternVL, and DeepSeek Backbones with Minimal Inference Overhead.

Backbone	Method	MMVet	MMBench	MMStar	VizWiz	POPE	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	Inference Time
LLaVA-1.5-13B	Base Model	35.4	67.7	41.6	53.6	85.9	48.3	14.1	2.7s
	+ CSR (Zhou et al., 2024c)	37.8	**68.8**	42.4	**56.8**	87.3	28.0	7.3	2.7s
	+ SeVa (Zhu et al., 2024)	41.0	68.7	44.2	54.7	87.4	23.6	**6.5**	2.7s
	+ Critic-V (Zhang et al., 2025a)	39.2	66.7	43.0	52.5	80.1	26.0	7.4	14.3s
	+ MM-Verify (Sun et al., 2025)	40.4	67.0	43.7	53.0	88.7	24.5	8.1	10.5s
	+ Sherlock (Ding and Zhang, 2025)	41.4	67.7	44.6	55.0	90.3	23.8	7.2	37.8s
	+ **TITA (Ours)**	**42.3**	68.2	**45.1**	55.2	**92.6**	**23.5**	6.6	2.8s
Qwen3-VL-8B	Base Model	85.5	84.7	70.9	39.0	91.5	50.6	23.5	1.4s
	+ CSR (Zhou et al., 2024c)	86.3	85.0	71.3	40.3	92.7	30.1	15.6	1.4s
	+ SeVa (Zhu et al., 2024)	88.4	86.8	72.6	44.2	94.8	25.3	11.7	1.4s
	+ Critic-V (Zhang et al., 2025a)	86.3	85.7	71.6	43.1	94.3	28.0	16.4	8.5s
	+ MM-Verify (Sun et al., 2025)	86.5	85.9	71.8	43.0	94.8	22.6	13.5	7.5s
	+ Sherlock (Ding and Zhang, 2025)	88.0	87.2	73.2	44.7	95.1	21.8	12.8	20.4s
	+ **TITA (Ours)**	**89.1**	**88.3**	**74.0**	**44.9**	**97.5**	**20.3**	**12.2**	1.6s
InternVL3.5-8B	Base Model	83.1	79.5	69.3	54.3	88.7	53.5	27.7	1.4s
	+ CSR (Zhou et al., 2024c)	84.2	80.1	70.4	54.5	90.2	36.8	18.7	1.4s
	+ SeVa (Zhu et al., 2024)	86.8	82.8	72.3	55.1	94.3	33.6	10.5	1.4s
	+ Critic-V (Zhang et al., 2025a)	84.5	80.7	70.5	54.6	93.5	34.4	12.9	8.5s
	+ MM-Verify (Sun et al., 2025)	84.1	80.4	70.2	55.0	92.0	35.7	16.9	7.6s
	+ Sherlock (Ding and Zhang, 2025)	**88.2**	**84.0**	**73.8**	55.2	96.0	28.5	9.4	21.5s
	+ **TITA (Ours)**	87.7	83.7	73.4	**55.3**	**96.3**	**22.3**	**8.5**	1.6s
DeepSeek-VL2-27B	Base Model	52.8	71.3	49.0	47.4	88.8	41.3	16.7	3.9s
	+ CSR (Zhou et al., 2024c)	54.8	72.4	50.3	48.6	90.4	26.6	12.9	3.9s
	+ SeVa (Zhu et al., 2024)	56.3	73.0	51.6	50.5	92.6	22.7	10.6	3.9s
	+ Critic-V (Zhang et al., 2025a)	56.0	72.8	51.3	50.0	94.1	16.7	8.3	23.5s
	+ MM-Verify (Sun et al., 2025)	55.8	72.9	51.4	50.7	93.6	17.5	9.2	17.0s
	+ Sherlock (Ding and Zhang, 2025)	56.8	74.1	51.6	**51.0**	93.7	14.5	7.0	54.2s
	+ **TITA (Ours)**	**57.3**	**73.9**	**52.0**	50.4	**94.7**	**12.5**	**4.9**	4.2s

Table 3: Computational overhead comparison on MMVet (LLaVA-1.5-7B). TITA achieves state-of-the-art accuracy while requiring significantly fewer FLOPs than dual-stream decoding (MARINE).

Method	MMVet	FLOPs ($\times 10^{12}$)
Base Model	30.5	8.68
+ CSR (Training-time)	33.9	8.68
+ MARINE (Inference-time)	38.5	17.36
+ **TITA (Ours)**	**39.1**	**10.54**

Comparison with Heuristic Decoding. Finally, we differentiate TITA from training-free heuristics like VCD (Leng et al., 2024), M3ID (Favero et al., 2024), and MARINE (Zhao et al., 2024) (Table 5). These methods attempt to mitigate hallucinations by reweighting the base model's logits using fixed contrastive formulas. VCD contrasts with perturbed images, M3ID contrasts with text-only inputs, and MARINE contrasts with caption-conditioned outputs. These approaches rely on hand-crafted perturbations without learned visual grounding signals. While TITA incorporates a learned multimodal reward model $\pi_{reward}(y||q, I)$ that provides explicit preference supervision. By leveraging this learned signal, TITA guides the model toward visually faithful tokens more effectively than methods relying only on noise or priors.

Table 4: Decoupling Guidance from Data: TITA (frozen base) matches or exceeds the performance of SFT/DPO (fine-tuned base) using the **exact same** preference data, but requires only $\sim 5\%$ of the training cost.

Method	Training Cost	MMVet	MMBench	POPE
Backbone: LLaVA-1.5-13B				
Base Model	-	35.4	67.7	85.9
+ SFT	3.4h	39.6	68.0	89.0
+ DPO	11.0h	42.5	68.1	92.9
+ **TITA (Ours)**	**0.6h**	**42.3**	**68.2**	**92.6**
Backbone: Qwen3-VL-8B				
Base Model	-	85.5	84.7	91.5
+ SFT	3.9h	86.7	86.5	93.0
+ DPO	7.5h	89.4	88.3	97.5
+ **TITA (Ours)**	**0.4h**	**89.1**	**88.3**	**97.5**

4.3 Ablations and Analysis

Impact of scale factor λ. We investigate the sensitivity of the hyperparameter λ, which governs the trade-off between the base model's original priors and the reward model's preference guidance. As illustrated in Figure 3, increasing λ from 0 to 0.6 steadily improves MMVet, culminating in a peak score of 39.1%. This trend is mirrored in POPE, which rises from 85.9% to 91.7% at the same threshold. This consistent peak across reasoning and grounding tasks confirms that the optimal window $\lambda \in [0.5, 0.7]$ is robust and transferable,

Table 5: Unlike heuristic methods that rely on noise perturbations or linguistic priors, TITA leverages a learned reward model for explicit visual grounding.

Model	Inference logits	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$
Backbone: LLaVA-1.5-7B	$\log \pi_\theta(y\|q, I)$	48.8	14.9
+ VCD	$(1+\lambda)\log \pi_\theta(y\|q, I) - \lambda \log \pi_\theta(y\|q, \hat{I})$	28.1	11.0
+ M3ID	$(1-\lambda)\log \pi_\theta(y\|q, I) + \lambda \log \pi_\theta(y\|q)$	27.1	6.4
+ MARINE	$(1-\lambda)\log \pi_\theta(y\|q, c, I) + \lambda \log \pi_\theta(y\|q, I)$	17.8	7.2
+ **TITA (Ours)**	$(1-\lambda)\log \pi_{\text{reward}}(y\|q, I) + \lambda \log \pi_\theta(y\|q, I)$	**16.3**	**5.6**

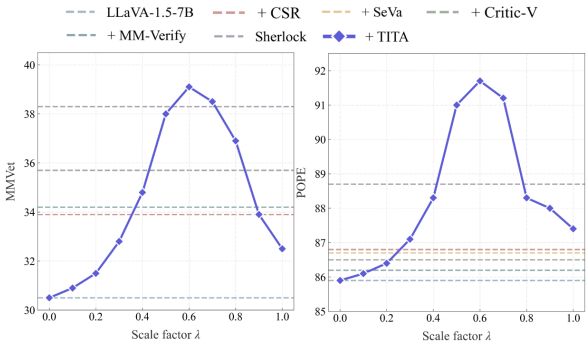

Figure 3: Ablation studies on reward integration factor λ in Eq. 8. TITA achieves optimal performance at $\lambda \approx 0.6$ across both reasoning and hallucination tasks.

rather than task-specific.

Effectiveness of Multi-view Preference Fusion. We evaluate the efficacy of our multi-view preference fusion strategy by contrasting it with standard single-view image augmentations. The left panel of Figure 4 shows that using a single augmentation (e.g., *RandFlip*, *Contrast*) leads to modest gains over the baseline. For example, *Contrast* and *Diffusion-W* provide slight boosts on MMVet, but they struggle to offer robust gains across hallucination benchmarks due to limited semantic variation. Conversely, the right panel of Figure 4 highlights the superiority of our fusion approach, which aggregates consensus from multiple perturbed views to construct a high-quality target. A clear monotonic trend is observed: as the number of fused views increases from 1 to 6, the quality of the constructed preference pairs improves significantly, driving performance up to 39.1% on MM-Vet and 91.7% on

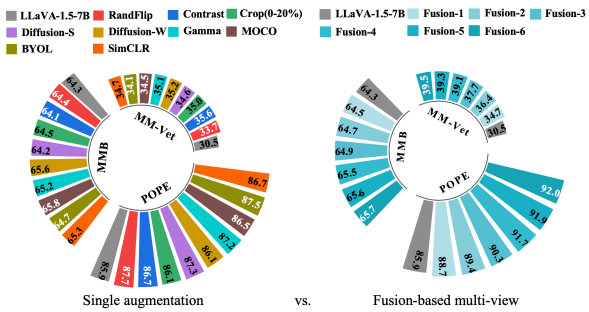

Figure 4: Ablation study on preference pair construction: Single augmentation vs. Multi-view fusion.

Table 6: GPT-4o pairwise evaluation between fusion-based winners (y_w) and original responses (y_l). The overwhelming preference for y_w validates the high quality of our self-supervised preference construction.

Dataset	GPT-4o Preference (%)		
	y_w Win rate ↑	y_l Win rate ↓	Tie rate
TextVQA	**97.30**	0.44	2.26
OCRVQA	**85.12**	2.95	11.93

POPE. This confirms that fusing diverse visual perspectives effectively filters out noise, thereby distilling a more reliable and grounded supervision signal for the reward model.

Quantitative Validation of Winners y_w. To further verify the quality of TITA 's automatically constructed data, we employ GPT-4o-2025-03-26 as an impartial judge to compare the fusion-based winners y_w against original responses y_l. Evaluation sets are constructed from TextVQA and OCRVQA. As shown in Table 6, y_w substantially outperforms y_l across both datasets: 97.3% on TextVQA and 85.1% on OCRVQA. Importantly, we conduct a rigorous qualitative analysis to rule out potential *verbosity bias*, ensuring that the evaluation does not merely favor longer outputs. Our findings confirm that the superiority of y_w stems from **increased factuality** rather than redundant linguistic padding. While y_w responses are occasionally more comprehensive, this reflects **denser visual grounding**. Specifically, the multi-view fusion mechanism enables y_w to capture fine-grained visual evidence (e.g., precise spatial coordinates and localized signage texts) that single-view generations frequently miss. This rich, grounded detail directly correlates with higher correctness and reduced hallucination, thus proving the efficacy of our data synthesis. Detailed evaluation protocols, scoring criteria, and comprehensive sample pairs are provided in Appendix C.1 and Table 7.

4.4 Qualitative Alignment Analysis

To elucidate the mechanism driving the performance gains of TITA, we analyze attention dynamics during token generation. Qualitatively, visualizations in Appendix C.4 (Figure 6) reveal that the baseline model often exhibits diffuse attention, failing to anchor on relevant visual regions, leading to hallucinations. This observation is substantiated by a layer-wise diagnosis (Figure 7), which identifies a critical "visual accumulation" phase in the

middle layers (5–18). By reinforcing visual evidence accumulation within this specific window, TITA prevents subsequent semantic refinement layers from generating text based solely on language probability, confirming that our token-level rewards effectively steer the model to prioritize visual fidelity over parametric knowledge.

5 Conclusion

We presented TITA , a framework that shifts VLM alignment from costly parameter retraining or delayed sequence reranking to precise, token-level intervention. By reframing preference optimization as a decoding-time guidance problem, TITA transforms sparse sequence-level rewards into dense autoregressive signals. This approach effectively counteracts the dominance of linguistic priors by steering the generation trajectory using log-probability ratios between a lightweight reward model and a frozen target model. To ensure broad interoperability, we introduce a dynamic cross-tokenizer mapping mechanism, allowing TITA to function as a plug-and-play module across disparate architectures. Extensive empirical evidence across diverse VLM families—including LLaVA, Qwen3-VL, InternVL3.5, and DeepSeek-VL2—confirms that our method consistently suppresses hallucinations and enhances multimodal reasoning. By achieving these gains with minimal computational overhead, TITA establishes a scalable and efficient paradigm for deploying reliable, visually grounded vision-language models.

Ethics Statement

In this work, we propose an inference-time alignment framework for vision-language models that significantly improves factual consistency without requiring expensive human annotations. By enhancing visual grounding, our method has the potential to increase the reliability and accessibility of multimodal systems in critical domains such as education, assistive technology, and scientific analysis. However, as a dual-use technology, this improved steerability could theoretically be exploited to generate more convincing misinformation or to circumvent safety guardrails. Furthermore, because our approach relies on self-supervised preference construction, the resulting model may inherit or inadvertently reinforce systemic biases present in the foundational pre-training data. We advocate for careful deployment, transparency in alignment mechanisms, and continued research into fairness-aware multimodal reward modeling to mitigate these risks.

Large Language Model Usage

During the preparation of this manuscript, we employed large language models exclusively for editorial purposes, including language polishing, grammar correction, and structural refinement. We strictly confined their usage to improving the readability and English presentation of the text. These models did not contribute to the conception of the core ideas, theoretical formulations, experimental design, data analysis, or the interpretation of the results. We take full responsibility for the entirety of the content within this paper.

Limitations

The performance of TITA is inherently bounded by the discriminative capacity of the reward model, while achieving the optimal trade-off between visual adherence and linguistic diversity necessitates precise calibration of the guidance scale.

Acknowledgments

This work is supported by Taobao&Tmall Group of Alibaba through Alibaba Research Intern Program and AlibabaGroup Innovative Research Program.

References

Muhammad Awais, Muzammal Naseer, Salman Khan, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, and Fahad Shahbaz Khan. 2025. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and 1 others. 2024a. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024*

Conference on Empirical Methods in Natural Language Processing, pages 7346–7370.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024b. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2024c. Dress: Instructing large vision-language models to align and interact with humans via natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14239–14250.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. 2024d. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*.

Jasper Dekoninck, Marc Fischer, Luca Beurer-Kellner, and Martin Vechev. 2023. Controlled text generation via language model arithmetic. *arXiv preprint arXiv:2311.14479*.

Yi Ding and Ruqi Zhang. 2025. Sherlock: Self-correcting reasoning in vision-language models. *arXiv preprint arXiv:2505.22651*.

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

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312.

Deqing Fu, Tong Xiao, Rui Wang, Wang Zhu, Pengchuan Zhang, Guan Pang, Robin Jia, and Lawrence Chen. 2024. Tldr: Token-level detective reward model for large vision language models. *arXiv preprint arXiv:2410.04734*.

Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. 2024. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation. In *European Conference on Computer Vision*, pages 388–404. Springer.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, and 1 others. 2020. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, and 1 others. 2025a. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*.

Jiawei Guo, Tianyu Zheng, Yizhi Li, Yuelin Bai, Bo Li, Yubo Wang, King Zhu, Graham Neubig, Wenhu Chen, and Xiang Yue. 2025b. Mammoth-vl: Eliciting multimodal reasoning with instruction tuning at scale. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13869–13920.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. 2024. Visual hallucinations of multimodal large language models. *arXiv preprint arXiv:2402.14683*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2025. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25004–25014.

Janak Kapuriya, Chhavi Kirtani, Apoorv Singh, Jay Saraf, Naman Lal, Jatin Kumar, Adarsh Raj Shivam, Astha Verma, Avinash Anand, and Rajiv Ratn Shah. 2024. Mm-phyrlhf: Reinforcement learning framework for multimodal physics question-answering. *arXiv preprint arXiv:2404.12926*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. 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*, pages 13872–13882.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023a. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023b. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023c. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023d. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2025a. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pages 216–233. Springer.

Zhenhua Liu, Lijun Li, Ruizhe Chen, Yuxian Jiang, Tong Zhu, Zhaochen Su, Wenliang Chen, and Jing Shao. 2025b. Iterative value function optimization for guided decoding. *arXiv preprint arXiv:2503.02368*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE.

Ajay Patel, Markus Hofmarcher, Claudiu Leoveanu-Condrei, Marius-Constantin Dinu, Chris Callison-Burch, and Sepp Hochreiter. 2024. Large language models can self-improve at web agent tasks. *arXiv preprint arXiv:2405.20309*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. 2024. Pixellm: Pixel reasoning with large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26374–26383.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, and 1 others. 2025. Vlm-r1: A stable and generalizable r1-style large vision-language model. *arXiv preprint arXiv:2504.07615*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Linzhuang Sun, Hao Liang, Jingxuan Wei, Bihui Yu, Tianpeng Li, Fan Yang, Zenan Zhou, and Wentao Zhang. 2025. Mm-verify: Enhancing multimodal reasoning with chain-of-thought verification. *arXiv preprint arXiv:2502.13383*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Zhongwei Wan, Zhihao Dou, Che Liu, Yu Zhang, Dongfei Cui, Qinjian Zhao, Hui Shen, Jing Xiong, Yi Xin, Yifan Jiang, and 1 others. 2025. Srpo: Enhancing multimodal llm reasoning via reflection-aware reinforcement learning. *arXiv preprint arXiv:2506.01713*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin

Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Zhaoyang Wang, Weilei He, Zhiyuan Liang, Xuchao Zhang, Chetan Bansal, Ying Wei, Weitong Zhang, and Huaxiu Yao. 2024. Cream: Consistency regularized self-rewarding language models. *arXiv preprint arXiv:2410.12735*.

Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, and 1 others. 2024a. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *Advances in Neural Information Processing Systems*, 37:69925–69975.

Jinge Wu, Yunsoo Kim, and Honghan Wu. 2024b. Hallucination benchmark in medical visual question answering. *arXiv preprint arXiv:2401.05827*.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, and 1 others. 2024c. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. 2024. Llava-critic: Learning to evaluate multimodal models. *arXiv preprint arXiv:2410.02712*.

Siming Yan, Min Bai, Weifeng Chen, Xiong Zhou, Qixing Huang, and Li Erran Li. 2024. Vigor: Improving visual grounding of large vision language models with fine-grained reward modeling. In *European Conference on Computer Vision*, pages 37–53. Springer.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.

Yuhang Zang, Xiaoyi Dong, Pan Zhang, Yuhang Cao, Ziyu Liu, Shengyuan Ding, Shenxi Wu, Yubo Ma, Haodong Duan, Wenwei Zhang, and 1 others. 2025. Internlm-xcomposer2. 5-reward: A simple yet effective multi-modal reward model. *arXiv preprint arXiv:2501.12368*.

Di Zhang, Jingdi Lei, Junxian Li, Xunzhi Wang, Yujie Liu, Zonglin Yang, Jiatong Li, Weida Wang, Suorong Yang, Jianbo Wu, and 1 others. 2025a. Critic-v: Vlm critics help catch vlm errors in multimodal reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9050–9061.

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

Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. 2024. Mitigating object hallucination in large vision-language models via image-grounded guidance. *arXiv preprint arXiv:2402.08680*.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. 2024a. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*.

Xiongtao Zhou, Jie He, Yuhua Ke, Guangyao Zhu, Víctor Gutiérrez-Basulto, and Jeff Z Pan. 2024b. An empirical study on parameter-efficient fine-tuning for multimodal large language models. *arXiv preprint arXiv:2406.05130*.

Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. 2024c. Calibrated self-rewarding vision language models. *arXiv preprint arXiv:2405.14622*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Ke Zhu, Liang Zhao, Zheng Ge, and Xiangyu Zhang. 2024. Self-supervised visual preference alignment. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 291–300.

Appendix

To ensure reproducibility, this supplementary material provides comprehensive technical details. Appendix A theoretically justifies using log-probability rewards for token-level guidance. Appendix B formalizes our inference pipeline and Top-k Tokenizer Mapping. Appendix C details the experimental setup, encompassing benchmark descriptions (Appendix C.1), granular results (Table 7), text-only LLM generalization (Appendix C.2), and training strategies (Appendix C.3). Finally, Appendix C.4 analyzes Visual Attention Dynamics to elucidate the mechanisms behind our performance gains.

A Theoretical Justification for Log-Probability Reward in VLMs

In this section, we provide a theoretical foundation for parameterizing the reward function as a log-probability distribution, $\log \pi_r(y \mid q, I)$. We demonstrate that under the Bradley-Terry (BT) model, this parameterization is not merely an approximation but a theoretically complete representation. Specifically, we prove that for any valid reward function r, there exists a corresponding autoregressive distribution π_r that induces an identical preference ranking over responses. This ensures that optimizing against $\log \pi_r$ is mathematically equivalent to optimizing against the original reward r.

Theorem I. Let $\mathcal{R}$ denote the class of reward functions consistent with the Plackett-Luce model over multimodal input (q, I). Then, for every $r \in \mathcal{R}$, there exists a probability distribution $\pi_r(y \mid q, I)$ such that the log-probability reward $\log \pi_r(y \mid q, I)$ belongs to the same preference equivalence class as r. Moreover, this parameterization is unique within each equivalence class.

This result implies that using the autoregressive likelihood $\log \pi_r(y \mid q, I)$ as a surrogate reward function in VLMs is not merely an approximation but a complete and expressive formulation under the Plackett-Luce framework. Despite the complexity of multimodal grounding where visual evidence and linguistic instructions jointly influence the response, the log-probability form preserves the full range of expressible preferences encoded by reward functions in $\mathcal{R}$. To formalize this claim, we first define equivalence classes of reward functions based on the preference distributions they induce.

Lemma. (Adapted from (Rafailov et al., 2024)) Under the Plackett-Luce or Bradley-Terry model, two reward functions $r_1(q, I, y)$ and $r_2(q, I, y)$ are equivalent if they induce the same pairwise preference probabilities over responses:

$$P(y \succ y' \mid q, I) = \frac{\exp(r(q, I, y))}{\exp(r(q, I, y)) + \exp(r(q, I, y'))},$$

Furthermore, any pair of equivalent reward functions leads to the same optimal policy in constrained reinforcement learning settings.

Proof. Let $r(q, I, y) \in \mathcal{R}$ be an arbitrary reward function. Define its normalized variant via the softmax transformation:

$$\hat{r}(q, I, y) := \log \frac{\exp(r(q, I, y))}{\sum_z \exp(r(q, I, z))} = r(q, I, y) - \log \sum_z \exp(r(q, I, z)),$$

The corresponding conditional distribution is:

$$\pi_r(y \mid q, I) = \frac{\exp(r(q, I, y))}{\sum_z \exp(r(q, I, z))},$$

and hence $\log \pi_r(y \mid q, I) = \hat{r}(q, I, y)$.

We now show that $\hat{r}(q, I, y)$ and $r(q, I, y)$ belong to the same preference equivalence class. Observe that the transformation introduces only a constant shift:

$$r(q, I, y) - \hat{r}(q, I, y) = \log \sum_z \exp(r(q, I, z)),$$

which is independent of y. Therefore, the pairwise preference between any two outputs remains unchanged:

$$\frac{\exp(r(q,I,y))}{\exp(r(q,I,y))+\exp(r(q,I,y'))}=\frac{\exp(\hat{r}(q,I,y))}{\exp(\hat{r}(q,I,y))+\exp(\hat{r}(q,I,y'))}.$$

Since the preference structure is preserved, the same ranking over outputs is induced, and thus the same optimal policy is obtained when optimizing under such preferences. This confirms that $\log \pi_r(y \mid q, I)$ is a faithful representative of the equivalence class defined by $r(q, I, y)$. □

Theorem II. All reward equivalence classes can be represented with the parameterization $\log \pi_r(y|q, I)$ for some probability distribution $\pi_r(y|q, I)$.

Proof Sketch. Take any reward function $r(q, I, y)$. Consider the following reward function

$$\hat{r}(q,I,y) := \log \frac{\exp r(q,I,y)}{\sum_z \exp r(q,I,z)}.$$

First, $\hat{r}(q, I, y)$ is consistent with the parameterization $\log \pi_r(y|q, I)$ with $\pi_r(y|q, I) = \frac{\exp r(q,I,y)}{\sum_z \exp r(q,I,z)}$. Second, since $r(q, I, y) - \hat{r}(q, I, y) = \log \sum_z \exp r(q, I, z)$ does not depend of y, $\hat{r}(q, I, y)$ and $r(q, I, y)$ are equivalent. Therefore, $\hat{r}(q, I, y)$ is a member of the equivalence class of $r(q, I, y)$ with the desired form, and we do not lose any generality in our reward model from the proposed parameterization. □

B Algorithms

Algorithm 1 formalizes the TITA pipeline, integrating multi-view preference construction, reward optimization, and cross-tokenizer token-level guidance.

Algorithm 1 Token-level Inference-time Alignment

Require: Dataset with query prompts and images: $\mathcal{D} = \{(q_n, I_n)\}_{n=1}^{N}$; target model π_θ; target model tokenizer $\mathcal{T}_\theta$; reward model π_r; reward model tokenizer $\mathcal{T}_r$; alignment hyper-parameter β; inference query prompt and image: (q^*, I^*); number of output tokens T; scaling factor λ; Image augmentation methods $\{f_k(\cdot)\}_{k=1}^{K}$, $\mathbb{P}$ is the softmax-derived token probability distribution.

$\mathcal{D}_p \leftarrow \{\}$ // Construct preference dataset $\mathcal{D}_p$ for reward model training.
for $n = 1, \ldots, N$ **do**
  for each augmentation methods $f_k(\cdot)$ **do**
    $I_n^k \leftarrow f_k(I_n)$ // Augment images.
    $\hat{y}_n^k \sim \pi_\theta(\cdot|q_n, I_n^k)$ // Generate candidate response from augmented input.
  end for
  $y_l^n \sim \pi_\theta(\cdot|q_n, I_n)$ // Loser response generated by the pretrained model.
  $y_w^n \sim$ Fusion$(\hat{y}_n^1, \hat{y}_n^2, \ldots, \hat{y}_n^K)$ // Winner response generated from fusion candidate answers.
  $\mathcal{D}_p \leftarrow \mathcal{D}_p \cup (q_n, I_n, y_w^n, y_l^n)$ // Adding the triplet to the preference dataset.
end for
// Training the auto-regressive reward model π_r.

$$\min_{\pi_r} -\mathbb{E}_{(q,I,y_w,y_l)\sim\mathcal{D}_p}\Big[\log\sigma\Big(\beta\sum_t \log\pi_r(y_{w,t}|q,I,y_{w,<t}) - \beta\sum_t \log\pi_r(y_{l,t}|q,I,y_{l,<t})\Big)\Big]$$

// Token-level reward guidance during inference stage.
for $t = 0, \ldots, T-1$ **do**
  if $\mathcal{T}_r \neq \mathcal{T}_{\text{target}}$ **then**
    $\mathbb{P}[\mathcal{T}_r(\mathcal{V})] \leftarrow \pi_r(y_t|q^*, I^*, y_{<t})$
    // Logits mapping with top-k tokens.
    $\mathcal{V}^{(k)} \leftarrow$ top-k tokens with highest likelihood
    $\mathbb{P}[\mathcal{T}_\theta(\mathcal{V}^{(k)})] \leftarrow \mathbb{P}[\mathcal{T}_r(\mathcal{V}^{(k)})]$
    $\pi_{\text{decode}}(y_t|q^*, I^*, y_{<t}) \leftarrow \pi_\theta(y_t|q^*, I^*, y_{<t})\big(\mathbb{P}[\mathcal{T}_\theta(\mathcal{V}^{(k)})]\big)^\lambda$
  else
    $\pi_{\text{decode}}(y_t|q^*, I^*, y_{<t}) \leftarrow \pi_\theta(y_t|q^*, I^*, y_{<t})\big(\mathbb{P}[\mathcal{T}_r(\mathcal{V})]\big)^\lambda$
  end if
  // Next predict token sampling:
  $y_t \leftarrow$ top-1 token from logits $\pi_{\text{decode}}(y_t|q^*, I^*, y_{<t})$
  $y_{<t+1} \leftarrow y_{<t} \,||\, y_t$
end for
Ensure: Generated response $y_{<t}$

C Experimental Details

C.1 Evaluation Benchmarks

LLaVA-Bench (In the wild) (Liu et al., 2024b): A challenging benchmark of 60 diverse tasks designed to evaluate models in naturalistic settings. It specifically tests visual instruction-following and question-answering capabilities in real-world scenarios, offering insights into practical applicability.

MM-Vet (Yu et al., 2023): A comprehensive evaluation suite comprising 218 diverse samples that assess six core visual-language capabilities. This benchmark uniquely integrates mathematical reasoning, logical inference, and visual knowledge understanding, providing a rigorous test of broad multi-modal comprehension.

MM-Bench (Liu et al., 2025a): A large-scale multi-modal benchmark with 4.7K samples, focusing on visual knowledge and reasoning capabilities. This dataset provides a balanced assessment of both factual knowledge and analytical reasoning in multi-modal contexts.

POPE (Li et al., 2023d): A specialized benchmark containing 8,440 samples designed to evaluate model hallucination. It specifically tests models' ability to provide accurate Yes/No responses about object presence in images, serving as a critical measure of visual grounding reliability.

MME (Yin et al., 2023): A benchmark with 14 tasks assessing perception and cognition in LVLMs, challenging interpretative and analytical skills.

SEED (Li et al., 2023b): A benchmark designed to evaluate the generative comprehension capabilities of large vision-language models (LVLMs). It includes an extensive dataset of 19K multiple-choice questions with precise human annotations, spanning 12 distinct evaluation dimensions that cover both spatial and temporal understanding across image and video modalities.

ScienceQA (Lu et al., 2022): A multimodal benchmark crafted to evaluate and diagnose the multi-hop reasoning abilities and interpretability of AI systems within the science domain. It features an extensive dataset of approximately 21k multiple-choice questions, spanning a broad spectrum of scientific topics and supplemented with detailed answer annotations, associated lectures, and explanations.

GQA (Hudson and Manning, 2019): A dataset specifically engineered for advanced real-world visual reasoning, utilizing scene graph-based structures to generate 22 million diverse, semantically-programmed questions. It incorporates novel evaluation metrics focusing on consistency, grounding, and plausibility, thereby establishing a rigorous standard for vision-language task assessment.

VizWiz (Gurari et al., 2018): A visual question answering (VQA) dataset derived from naturalistic settings, featuring over 31,000 visual questions. It is distinguished by its goal-oriented approach, with images captured by blind individuals and accompanied by their spoken queries, along with crowd-sourced answers.

MMStar (Chen et al., 2024b): A benchmark of 1,500 test samples designed to address issues of low vision–language alignment and potential training-data leakage. It is carefully curated and spans 6 core capability areas and 18 fine-grained evaluation axes.

CHAIR (Rohrbach et al., 2018): A well-established benchmark for evaluating object hallucination in image captioning tasks, with two variants: CHAIR_i and CHAIR_s, which assess hallucination at the instance and sentence levels, respectively. we randomly sampled 500 images from the COCO (Lin et al., 2014) validation set and evaluated object hallucination using the CHAIR metric. Note that a lower CHAIR score indicates fewer hallucinations, which implies better alignment between the captions and the actual content of the images.

$$\text{CHAIR}_i = \frac{\text{Number of hallucinated objects}}{\text{Number of all mentioned objects}},$$

$$\text{CHAIR}_s = \frac{\text{Number of captions with hallucinated objects}}{\text{Number of all captions}}.$$

C.2 Additional Detail Results

Granular Performance Breakdown. Table 7 provides a granular breakdown across three representative benchmarks. Specifically, MMVet assesses seven capabilities (including reasoning, OCR, spatial, and math); MMBench evaluates multilingual knowledge via English and Chinese subsets; and POPE tests hallucination robustness under random, popular, and adversarial settings. Across these diverse dimensions, `TITA` consistently improves upon the baselines.

Cross-Domain Generalization to Text-only LLMs. To validate the broader applicability of `TITA`, we extend our token-level alignment to text-only Large Language Models (LLMs) via a weak-to-strong paradigm. By employing a 1B reward

Table 7: Detailed performance breakdown on MMVet, MMBench, and POPE benchmarks.

Model	MMVet							MMBench		POPE			
	All	rec	ocr	know	gen	spat	math	en	cn	All	rand	pop	adv
Backbone: LLaVA-1.5-7B													
Base	30.5	35.7	21.9	17.7	19.7	24.7	7.7	64.3	58.3	85.9	89.5	86.7	81.7
+ Fact-RLHF(Sun et al., 2023)	31.4	36.5	22.7	18.1	20.9	32.3	7.7	63.4	56.8	81.5	86.5	83.9	83.0
+ CSR(Zhou et al., 2024c)	33.9	37.2	23.3	21.9	24.5	27.7	7.7	65.5	**59.4**	86.8	89.4	87.4	83.6
+ SeVa(Zhu et al., 2024)	37.2	40.2	29.9	21.8	23.9	34.3	7.7	**65.6**	59.2	86.7	89.4	87.1	83.6
+ Critic-V(Zhang et al., 2025a)	35.7	37.6	28.1	21.0	22.5	28.5	7.7	64.0	58.5	86.5	88.1	86.4	83.5
+ TITA (Ours)	**39.1**	**44.8**	**31.2**	**30.7**	**34.5**	**36.0**	7.7	65.5	59.2	**91.7**	**92.6**	**93.0**	**90.2**
Backbone: LLaVA-1.5-13B													
Base	35.4	38.9	32.2	23.3	24.8	29.7	24.8	67.7	63.6	85.9	89.6	86.5	82.0
+ Fact-RLHF (Sun et al., 2023)	32.6	41.2	28.9	22.8	23.7	34.1	25.2	64.7	58.0	86.7	89.4	87.5	82.5
+ CSR(Zhou et al., 2024c)	37.8	41.0	32.5	24.6	30.1	32.8	24.8	**68.8**	64.5	87.3	89.4	88.1	82.2
+ SeVa(Zhu et al., 2024)	41.0	**45.4**	32.8	32.4	36.7	37.0	**25.4**	68.7	**64.8**	87.4	90.5	89.0	82.7
+ Critic-V(Zhang et al., 2025a)	39.2	39.5	30.0	25.7	29.2	34.7	24.6	66.7	62.0	80.1	90.3	88.2	82.6
+ TITA (Ours)	**42.3**	44.8	**36.2**	**33.1**	**38.5**	**39.0**	24.8	68.2	64.2	**92.6**	**93.2**	**93.7**	**91.0**

model (Llama-3.2-1B-Instruct) to guide larger target models (Llama-3.1-8B and 70B), TITA consistently enhances reasoning and conversational abilities (Table 8). Notably, steering the 70B model with the 1B reward yields a +4.4% gain on MMLU and +0.16 on MT-Bench. This confirms that our explicit token-level guidance captures generalized alignment principles extending well beyond the vision-language domain.

Table 8: Cross-Domain Robustness: TITA is applied to text-only LLMs via weak-to-strong alignment, utilizing a fine-tuned Llama-3.2-1B-Instruct as the reward model to steer larger Llama-3.1 models.

Model	MMLU↑	MT-Bench↑
Llama-3.1-8B-Instruct	69.4	7.64
+ **TITA** *(1B RM)*	**74.8**	**8.51**
Llama-3.1-70B-Instruct	82.0	8.98
+ **TITA** *(1B RM)*	**86.4**	**9.14**

C.3 Experimental Setup

Image augmentation strategies To assess the impact of augmentation strategies, we analyzed 12 widely used techniques (Chen et al., 2020; Grill et al., 2020; He et al., 2020) (Figure 5). We found that overly aggressive methods (e.g., strong diffusion noise) hindered feature learning, while overly simple ones (e.g., random flipping) offered limited gains. Accordingly, we adopted a balanced combination of three effective augmentations with the

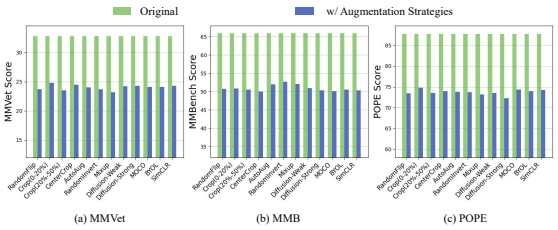

Figure 5: Comparison of 12 data augmentation strategies applied to LLaVA-1.5, including various geometric and color transformations as well as contrast learning enhancement methods. By analyzing these methods, the goal is to find the combination that best improves the performance of VLMs.

original images.

By applying these techniques to the original images, we produce multiple distinct responses which are then synthesized into a comprehensive final output. This approach enhances model robustness by introducing controlled variations in visual input while maintaining semantic consistency. The augmentation strategies include: (i) Diffusion-W (Weak): Introduces gaussian noise with 200 diffusion steps, offering a more moderate level of visual distortion. (ii) Contrast: Enhances image contrast by a factor of 2, accentuating visual boundaries and feature differences. (iii) Gamma: Performs gamma correction at a value of 0.8, lightening dark regions in the image. (Note that gamma values above 1 make shadows darker, while values below 1 make dark regions lighter).

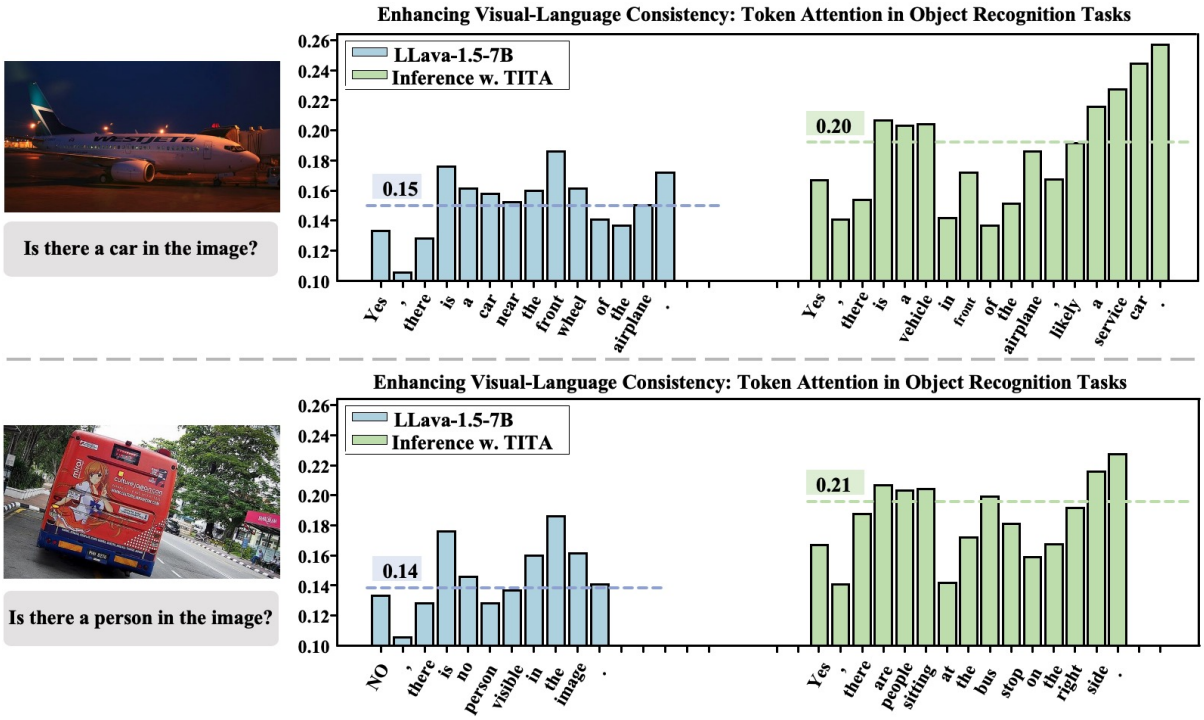

Figure 6: **Visualization of attention shifts during object generation.** The baseline LLaVA-1.5-7B often exhibits low or diffuse attention on relevant visual tokens, leading to hallucinations. TITA effectively steers the model to allocate higher attention weights to visual evidence, thereby ensuring the generated text is visually grounded.

C.4 Deep Dive into Visual Attention Dynamics

To understand why hallucinations emerge in VLMs and why TITA's decoding guidance is effective, we analyze how the model processes visual information during object-token generation. Prior work suggests (Li et al., 2023a; Zhu et al., 2023; Hurst et al., 2024; Shen et al., 2025) that VLMs rely heavily on linguistic priors, often before visual evidence is fully incorporated. Inspired by the attention diagnostic framework (Jiang et al., 2025), we conduct an independent layer-wise evaluation to contrast the dynamics of real tokens versus hallucinated tokens. We examine (a) the visual attention ratios (VAR) across layers and heads, and (b) the logit contribution of attention sublayers to the final object prediction.

This section provides a deeper investigation into the internal mechanisms of hallucination and the corrective effect of TITA. We focus on two aspects: (1) specific attention patterns in generated responses, and (2) statistical trends in attention distribution across model layers.

Response-Token Attention Visualization. Figure 6 compares the attention weights of generated tokens over image features for both the baseline LLaVA-1.5-7B and the TITA-guided inference. In the baseline case (top row), the model fails to attend to specific visual regions when generating object-related tokens (e.g., "car"), leading to hallucinations where the text describes objects absent from the image. Conversely, TITA (bottom row) produces sharper attention maps that tightly align with the corresponding visual objects. This qualitative evidence suggests that the token-level reward model explicitly penalizes ungrounded generation, forcing the decoding process to respect visual boundaries.

Layer-wise Analysis of Visual Grounding. To understand why hallucinations emerge in VLMs and why TITA's decoding guidance is effective, we analyze how visual information is processed during object-token generation, using Qwen3-VL as our representative base model. Prior work suggests (Li et al., 2023a; Zhu et al., 2023; Hurst et al., 2024; Shen et al., 2025) that VLMs rely heavily on linguistic priors, often before visual evidence is fully incorporated. Inspired by the attention diagnostic framework (Jiang et al., 2025), we conduct an independent layer-wise evaluation on Qwen3-VL to contrast the dynamics of real tokens versus hallucinated tokens. We examine (a) the visual attention ratios (VAR) across layers and heads, and (b) the logit contribution of attention sublayers to the final object prediction.

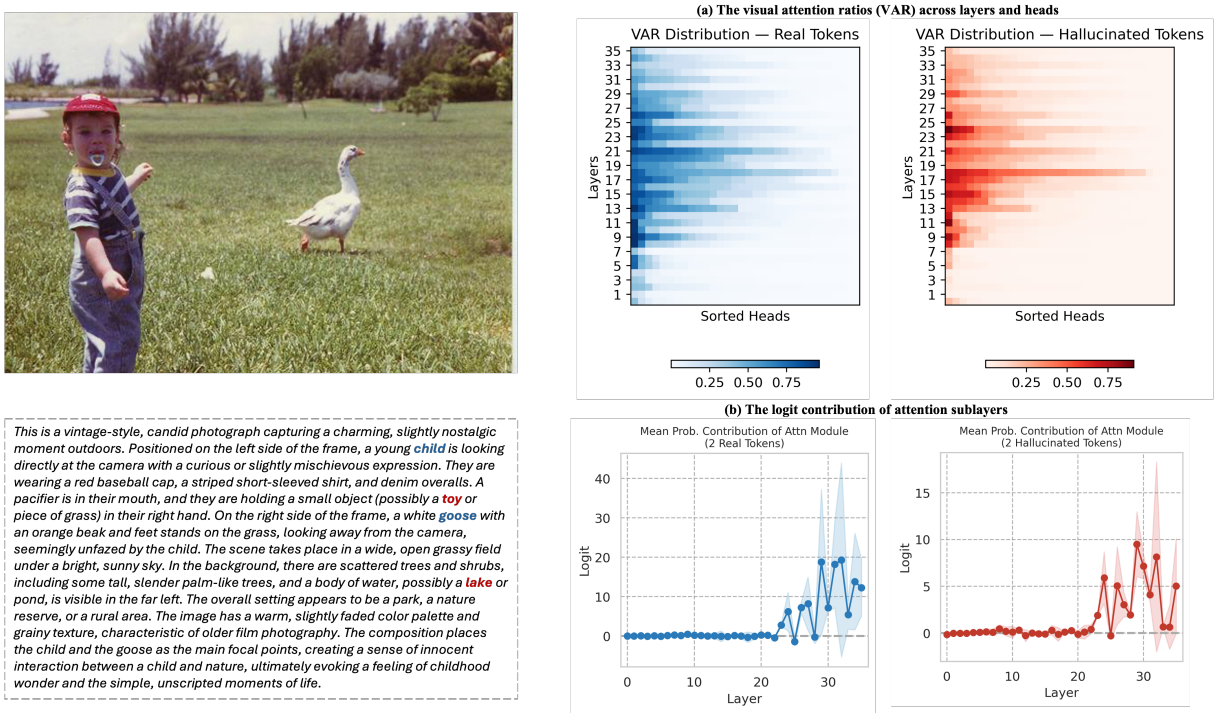

Figure 7: **Visual Attention Dynamics: Real vs. Hallucinated Tokens.** (Left) An example image and generated text, highlighting accurate groundings (child, goose) and hallucinations (toy, lake). (Right) Layer-head distribution of visual attention ratios (VAR) and mean logit contributions for real tokens (blue) versus hallucinated tokens (red). While hallucinated tokens exhibit visual attention in middle layers, their visual representations fail to translate into strong predictive signals in the upper layers, allowing linguistic priors to dominate.

As illustrated in Figure 7, contrasting the behavior of real tokens (blue) with hallucinated tokens (red) reveals a distinct two-stage processing pattern and the fundamental mechanistic cause of hallucinations. For accurately grounded tokens, the middle layers (e.g., layers 11–27) assign high and broadly sustained attention to image tokens, establishing a robust *visual evidence accumulation* stage. Following this, the upper layers (e.g., layers 24–35) exhibit a sharp, high-magnitude rise in logit contribution (peaking near 20), acting as a *semantic refinement* stage where these visual representations strongly dictate the final token prediction.

However, the dynamics of hallucinated tokens reveal a critical breakdown in this pipeline. As seen in the red distributions, while hallucinated tokens do trigger concentrated visual attention in a narrower band of middle layers (layers 15–23), this accumulation fails to be effectively propagated. Most tellingly, during the semantic refinement stage, the logit contribution from the visual attention sublayers is significantly weaker and highly unstable (peaking at merely half the magnitude of real tokens). Because the visual module fails to provide a dominant and decisive predictive signal, the generation process is inevitably hijacked by the model's internal parametric knowledge (linguistic priors), resulting in outputs driven by textual bias rather than grounded visual facts.

TITA addresses this exact vulnerability. By serving as a dense, token-level guide, it intervenes to reinforce and sustain the visual signal across the generation trajectory. By ensuring that the refinement stage receives a robust, high-magnitude visual conditioning that overpowers ungrounded textual bias, TITA fundamentally curtails the propensity for hallucination.