TIIMAGE: MITIGATING MLLM HALLUCINATIONS VIA UNIFIED CONTRASTIVE DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite the success of Multimodal Large Language Models (MLLMs), hallucinations severely undermine their reliability, largely driven by modality misalignment. Existing efforts primarily address modality misalignment via architectural projector enhancements or contrastive fine-tuning in the latent space, which incur substantial training overhead while leaving the fundamental decoupling of dual-stream inputs unsolved. Inspired by picture–word integration that unifies text and vision (Yum et al., 2021), we propose a novel training-free framework, **T**ext **i**n **I**mage (**TIImage**). TIImage renders textual instructions directly into image pixels, bridging the modality gap at the input level. Furthermore, to maximally exploit text-in-image integration without corrupting visual integrity, we systematically analyze spatial instruction rendering and introduce a non-invasive bottom-padding scheme. Paired with this unified input, TIImage employs a distribution-guided dynamic contrastive decoding strategy that isolates co-occurrence biases while preventing over-penalization.Extensive experiments across multiple benchmarks demonstrate that TIImage consistently suppresses hallucinations while maintaining fine-grained visual perception.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) excel at visual comprehension and instruction following (Koh et al., 2024; Peng et al., 2025; Huo et al., 2025), yet remain prone to hallucinations that contradict visual facts (Liu et al., 2025b; Su et al., 2025), raising safety concerns in critical domains like autonomous driving and medical diagnosis (Lin et al., 2025; Ding et al., 2024). Such errors largely stem from two coupled factors: modality misalignment during integration and reliance on training-data co-occurrence priors (Jain et al., 2025; Huang et al., 2025).

To mitigate hallucinations, prior works explore fine-tuning (Jiang et al., 2024), retrieval-augmented generation (RAG) (Zhao et al., 2025), attention intervention (Huang et al., 2024a), and contrastive decoding (CD) (Li et al., 2023b). Training-free CD methods are particularly promising for their inference efficiency, typically penalizing biased priors via noise-perturbed negative inputs (Leng et al., 2024; Wang et al., 2024).

Nevertheless, current CD approaches face two key limitations: (1) static noise injection introduces uncontrollable uncertainty and risks over-penalization, and (2) operating under dual-stream inputs leaves modality misalignment unaddressed. Drawing inspiration from picture–word integration that naturally unifies text and vision (Dehaene & Cohen, 2011; Ernst & Bülthoff, 2004), we explore rendering textual instructions directly into images to form a unified visual representation, which empirically alleviates feature misalignment (Figure 1).

Yet converting text into visual pixels is non-trivial. It introduces practical trade-offs between textual legibility and visual occlusion, while different spatial placements and canvas paddings distinctly affect multimodal reasoning. More critically, our preliminary probe reveals an unexpected hurdle: generating responses purely from this visual-rendered input actually exacerbates hallucinations (Figure 3). Without explicit text tokens, the model loses linguistic conditioning and defaults heavily to learned co-occurrence priors.

This finding indicates that input-level unification alone is insufficient to curb hallucinations, highlighting the value of coupling spatial instruction rendering with debiased decoding. We thus propose **TIImage**, a training-free framework coordinating input-level alignment with adaptive contrastive

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

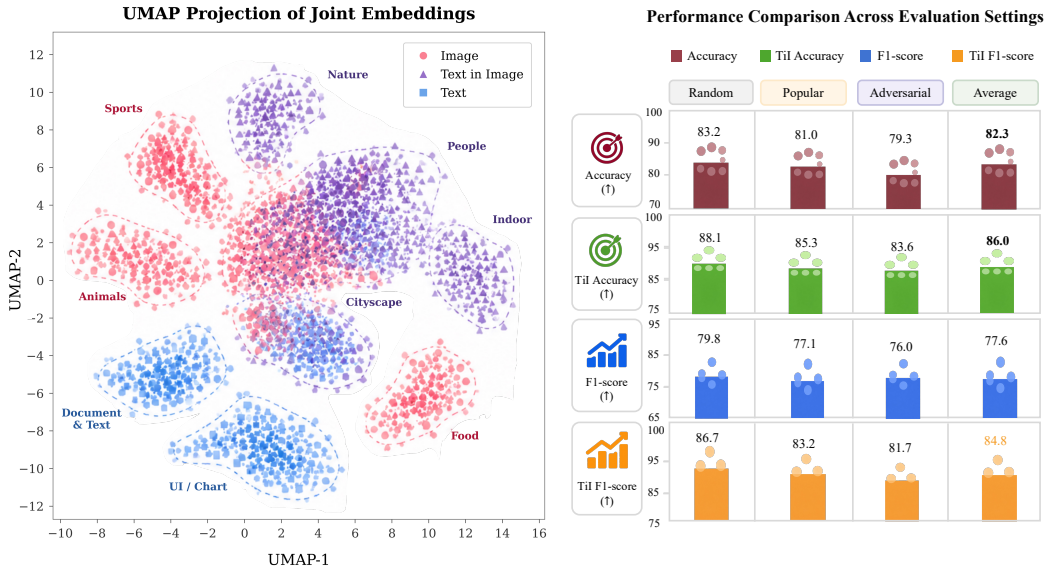


Figure 1: (Left) UMAP visualization of image, text, and text embedded image features. (Right) Performance comparison across three POPE settings, where TiI denotes the Text-in-Image.

decoding (Figure 2). Specifically, TIImage first renders the prompt T onto the image I to bridge the modality gap without visual occlusion (Stage 1). During autoregressive decoding, it decouples visual and textual representations (z_v, z_t) through respective adapters f_θ and g_θ, contrasting their logits via a balance coefficient α to penalize co-occurrence priors (Stage 2). Finally, to avoid over-penalizing valid tokens, TIImage applies an adaptive truncation threshold $\beta \cdot \max p_\theta$ before token sampling, promoting more faithful generation (Stage 3).

Quantitative evaluations demonstrate that TIImage substantially reduces hallucinations across mainstream benchmarks: it cuts instance-level hallucinations ($CHAIR_I$) by 19.8%–21.8% on the CHAIR benchmark while boosting object recall to 67.4%, and consistently yields 4.9%–5.6% accuracy gains on challenging POPE settings across various model scales. To summarize, our contributions include:

- **Modality-Prior Dilemma in Hallucinations:** We identify that conventional dual-stream inputs induce modality misalignment, while removing textual constraints or perturbing visual inputs forces MLLMs to fall back on object co-occurrence priors, providing key insights into hallucination triggers.
- **Unified Spatial-Preserving Integration:** We present a training-free input unification scheme that embeds textual queries directly into extended image canvas pixels, effectively bridging the cross-modal representation gap without visual occlusion.
- **Distribution-Aware Contrastive Decoding:** We propose a dynamic contrastive decoding strategy that isolates co-occurrence biases via an intentionally perturbed view, dynamically regulating contrast penalties through Jensen-Shannon Divergence and filtering uncertain tokens via entropy-guided truncation.
- **Superior Empirical Performance:** Extensive experiments across POPE, CHAIR, and MME demonstrate that TIImage substantially reduces object hallucinations (e.g., cutting $CHAIR_I$ by up to 21.8%) while preserving rich perceptual detail and fine-grained visual reasoning.

2 Related Work

MLLMs and Hallucinations. Despite the remarkable progress of MLLMs in multimodal generation (Koh et al., 2024; Peng et al., 2025; Huo et al., 2025), hallucinations—generating text inconsistent with visual facts—remain a critical bottleneck to their reliability (Liu et al., 2025a; Li et al., 2023b; Huang et al., 2024b; Chen et al., 2024b). Systematic analyses confirm that modality misalignment predominantly drives MLLM hallucinations (Bai et al., 2024; Jain et al., 2025; Huang

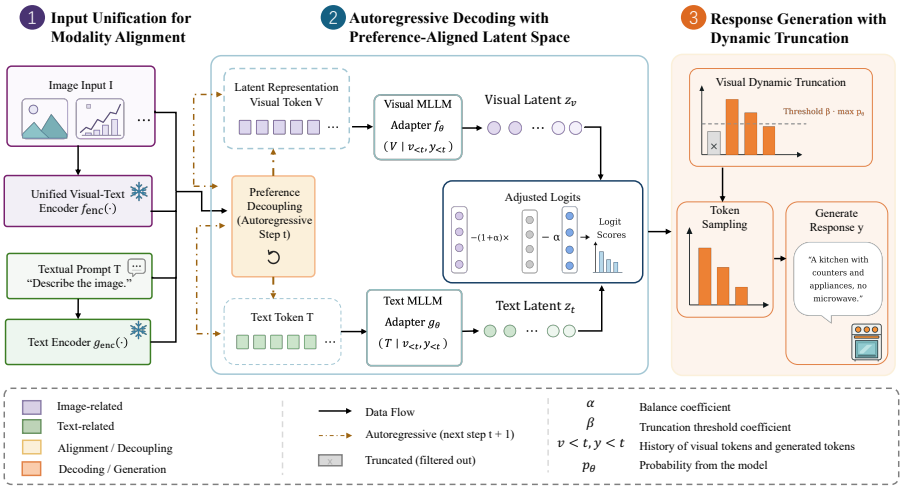


Figure 2: TIImage mitigates hallucinations via input unification, contrastive adjustment, and dynamic truncation.

et al., 2025), revealing a persistent modality gap that conventional alignment schemes struggle to bridge (Jiang et al., 2024). Formally, we define visual hallucination as generating outputs inconsistent with the input image V. An effective mitigation strategy aims to reduce erroneous assertions while maximizing factual faithfulness to V when responding to query T, demanding a delicate balance between visual grounding and linguistic fluency.

Training-based Methods. Extensive efforts attempt to bridge this representation gap via architectural modifications or optimization adjustments (Jiang et al., 2024; Ouali et al., 2024). For instance, HACL employs contrastive learning with hallucinated texts as negatives to align visual-text representations (Jiang et al., 2024), Honeybee designs a locality-enhanced projector to retain fine-grained visual details (Cha et al., 2024), and ShareGPT4V scales up high-quality pretraining data to enhance projector alignment (Chen et al., 2024a). However, these approaches rely on costly fine-tuning or complex architectural alterations, incurring heavy computational overhead and data annotation costs that restrict their practical deployment across diverse architectures.

Training-free Methods. To bypass retraining overhead, training-free methods modify inference dynamics. For example, Opera intervenes in self-attention maps (Huang et al., 2024a) but incurs high inference latency, whereas Marine relies on external object detectors (Zhao et al., 2025), making it sensitive to detection noise. Contrastive decoding (CD) approaches, such as ICD (Wang et al., 2024) and VCD (Leng et al., 2024), adjust logits to penalize biased priors (Li et al., 2023b; Shi et al., 2024). However, operating under standard dual-stream encoders (Wadekar et al., 2024; Liu et al., 2024), existing CD methods leave the input modality gap completely unaddressed and risk over-penalization via static perturbations. Unlike dual-stream paradigms, human picture–word integration naturally unifies textual and visual cues into a coherent signal (Dehaene & Cohen, 2011; Ernst & Bülthoff, 2004). Inspired by this, TIImage bridges modality misalignment at the input level by rendering instructions directly into image pixels. It then introduces an adaptive contrastive decoding mechanism that contrasts this unified input against a visually perturbed negative view, dynamically penalizing statistical priors without compromising visual grounding.

3 METHODOLOGY

Problem Formulation and Notations. Following (Leng et al., 2024; Zhao et al., 2025; Zou et al., 2025), we consider a multimodal large language model $\mathcal{M}_\theta^{\text{MLLM}}$ parameterized by θ, comprising a visual encoder, a vision-language projector, and an autoregressive text decoder. In the standard image-grounded text generation setting, the model takes an image V and a textual query T as inputs. The visual encoder first projects V into a sequence of continuous visual tokens $\mathbf{v} = [v_1, \ldots, v_p]$ with sequence length p. Simultaneously, the textual query T is tokenized into textual tokens $\mathbf{x} = [x_1, \ldots, x_q]$. These visual tokens $\mathbf{v}$ are concatenated with textual tokens $\mathbf{x}$ to construct the multimodal context, from which the decoder autoregressively generates the response $\mathbf{y}$:

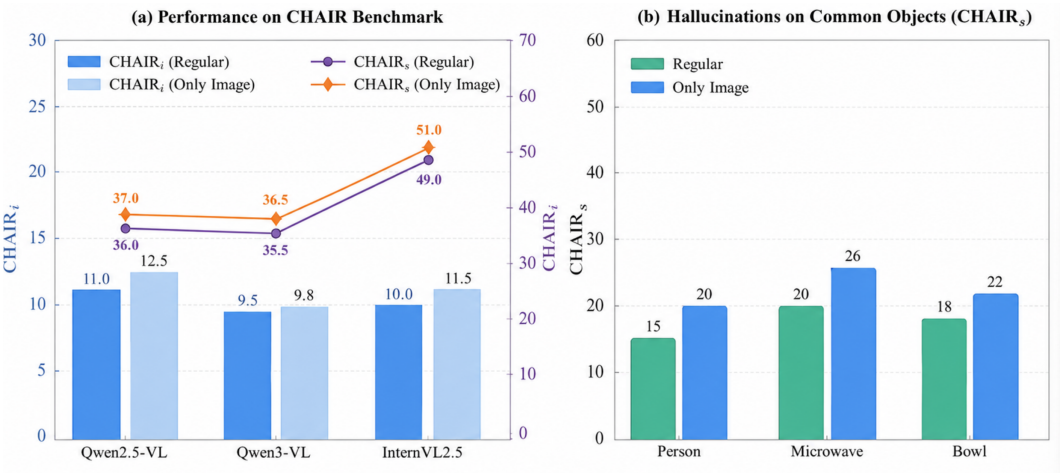


Figure 3: Evaluation of MLLM hallucinations on the CHAIR (Rohrbach et al., 2018) benchmark using only image inputs. Both CHAIR_s and CHAIR_i scores increase significantly.

$$y_t \sim p_\theta(y_t \mid \mathbf{v}, \mathbf{x}, \mathbf{y}_{<t}) = \text{softmax}(f_\theta(y_t \mid \mathbf{v}, \mathbf{x}, \mathbf{y}_{<t})), \tag{1}$$

where y_t denotes the predicted token at time step t, $\mathbf{y}_{<t}$ represents previously generated tokens, and $f_\theta(\cdot)$ denotes the logit predictions.

3.1 Overview of the TIImage Framework

Although rendering instructions into images reduces feature misalignment (Figure 1), omitting explicit text tokens causes hallucinations to surge on benchmarks like CHAIR (Figure 3), where the model defaults to learned co-occurrence priors. In instruction-driven tasks where textual constraints must be retained, intentionally blurring the visual input similarly exposes these statistical biases (Leng et al., 2024).

To jointly address modality misalignment and prior-driven hallucinations, we propose **TIImage** (Figure 2). The pipeline proceeds in three steps: (1) rendering instructions into an extended canvas to form a unified input V_{text} without visual occlusion; (2) contrasting unified logits against a blurred negative view $\tilde{V}$ to penalize statistical priors; and (3) adaptively calibrating penalties and truncation thresholds via distributional divergence and token entropy to safeguard generation fluency.

3.2 Visual Instructions for Alignment

To address the modality misalignment identified in Section 3.1, we introduce a rendering function $R(\cdot)$ to integrate both modalities into a unified visual input. To ensure text readability while preserving the integrity of original image features, we adopt a padding and rendering strategy. Specifically, we extend the pixel space at the bottom of the original image canvas with white pixels to construct a region for instruction placement. Subsequently, the text query T is rendered into this area using a standard black font. This design prevents text from overlapping with key objects in the image. Therefore, it effectively preserves the integrity of the original visual information. Formally, given the raw image V and the text query T, we construct the unified visual input V_{text} as follows:

$$V_{\text{text}} = R(V, T), \tag{2}$$

where V_{text} denotes the result of integrating textual and visual content into a single image. Consequently, we reformulate the generation of the output response y as an autoregressive process relying solely on unified visual features. This process is defined as follows:

$$y_t \sim p_\theta(y_t \mid \mathbf{v}_{\text{text}}, \mathbf{y}_{<\mathbf{t}}) = \text{softmax}(f_\theta(y_t \mid \mathbf{v}_{\text{text}}, \mathbf{y}_{<\mathbf{t}})), \tag{3}$$

where $\mathbf{v} = [v_1, \ldots, v_o]$ denotes the sequence of visual tokens encoded from V_{text}. Under this paradigm, although the textual queries are deeply unified into the visual space to mitigate modality misalignment via input-level visual instruction augmentation, the model continues to receive the separate text token sequences. The primary objective of retaining these textual instructions is to strictly constrain the model's generation behavior and ensure that the multimodal reasoning process remains solidly anchored within a unified semantic space.

3.3 TIIMAGE CONTRASTIVE DECODING STRATEGY

As established, modality alignment alone cannot fully suppress hallucinations, as models still default to statistical priors when visual evidence is compromised. To address this, we introduce the TIImage Contrastive Decoding (TCD) strategy. TCD contrasts predictions from the unified input against a visually perturbed negative view. While the unified input captures accurate semantics, the degraded view forces the model to rely on visual priors, effectively isolating its hallucination tendencies. Both branches retain identical textual constraints to ensure strict adherence to the task format.

Specifically, given a text query T and an original image V, we empirically construct the degraded image $\tilde{V}$ by applying a Gaussian blur with a radius of 7. The model generates distinct output distributions conditioned on the unified image V_{text} and the degraded image $\tilde{V}$. The contrastive distribution p_{TCD} is then derived directly from their discrepancy:

$$p_{\text{TCD}}(y \mid V_{\text{text}}, \tilde{V}) = \text{softmax}\Big[(1+\alpha) f_\theta(y \mid V_{\text{text}}) - \alpha f_\theta(y \mid \tilde{V})\Big], \tag{4}$$

Here, α controls the strength of the contrastive penalty. Essentially, TCD serves as an explicit calibration mechanism. By contrasting with the distribution from the visually perturbed image, which is highly prone to statistical bias, TCD effectively penalizes tokens that are generated based merely on co-occurrence frequencies, thereby robustly suppressing hallucinations.

3.4 DYNAMIC PROBABILITY TRUNCATION

The contrastive distribution defined in Equation 4 carries a potential challenge. It may impose an excessive penalty on the generation distribution conditioned on the image alone. However, this penalty assumption is not always correct. Despite the influence of statistical bias, the model still maintains basic linguistic standards and reflects reasonable commonsense reasoning (Leng et al., 2024). Applying penalties without distinction could incorrectly reduce the probability of valid tokens. To mitigate this issue, we adopt the approach from (Li et al., 2023a). We introduce a dynamic probability truncation mechanism based on the output confidence under the unified visual input:

$$\begin{aligned} \mathcal{V}_{\text{top}}(\mathbf{y}_{<t}) = \Big\{ y_t \in \mathcal{V} : p_\theta(y_t \mid \mathbf{v}_{\text{text}}, \mathbf{y}_{<t}) \geq \beta \max_w p_\theta(w \mid \mathbf{v}_{\text{text}}, \mathbf{y}_{<t}) \Big\}, \\ p_{\text{TCD}}(y_t \mid \mathbf{v}_{\text{text}}, \tilde{\mathbf{v}}) = 0, \quad \text{if } y_t \notin \mathcal{V}_{\text{top}}(\mathbf{y}_{<t}), \end{aligned} \tag{5}$$

where $\mathcal{V}$ denotes the output vocabulary of MLLMs, and the maximum operation is taken over all possible tokens $w \in V$. We employ $\beta \in [0, 1]$ to determine the threshold for truncating the distribution. The complete formulation, which incorporates both TCD and the adaptive probability truncation, is defined as:

$$y_t \sim \text{softmax}\Big[(1+\alpha) f_\theta(y_t \mathbf{v}_{\text{text}}, \mathbf{y}_{<t}) - \alpha f_\theta(y_t \mid \tilde{\mathbf{v}}, \mathbf{y}_{<t})\Big], \quad \text{s.t. } y_t \in \mathcal{V}_{\text{top}}(\mathbf{y}_{<t}) \tag{6}$$

This formulation establishes the core mechanism of TIImage. It leverages TCD to mitigate statistical biases stemming from unconstrained visual priors, while dynamic truncation restricts selection to high-confidence candidate tokens. The complete algorithmic workflow is detailed in Appendix A. Together, these mechanisms promote more factually grounded generation.

3.5 DECODING VIA DISTRIBUTIONAL DYNAMICS

While the contrastive strategy effectively suppresses hallucinations, relying on static hyperparameters (α and β) across all decoding steps introduces a new challenge: it tends to over-penalize the

model during robust reasoning and under-filter during highly uncertain steps. To address this, we propose a dynamic modulation mechanism driven by Jensen-Shannon Divergence (JSD) and Information Entropy.

JSD-guided Gating Penalty. We utilize JSD to quantify the distributional discrepancy between the unified view and the negative view. A high divergence indicates that the model is heavily influenced by statistical bias, necessitating a contrastive penalty. Let $p_{\text{unif}} = p_\theta(\cdot \mid \mathbf{v}_{\text{text}}, \mathbf{y}_{<t})$ and $p_{\text{img}} = p_\theta(\cdot \mid \mathbf{v}, \mathbf{y}_{<t})$. Rather than applying a continuous penalty, we introduce a strict gating mechanism with an adaptive coefficient α_t:

$$\begin{aligned} M_t &= \frac{1}{2}(p_{\text{unif}} + p_{\text{img}}), \\ D_{\text{JSD}} &= \frac{1}{2}D_{\text{KL}}(p_{\text{unif}} \parallel M_t) + \frac{1}{2}D_{\text{KL}}(p_{\text{img}} \parallel M_t), \\ \alpha_t &= \begin{cases} \alpha, & \text{if } D_{\text{JSD}} > \tau, \\ 0, & \text{otherwise}, \end{cases} \end{aligned} \tag{7}$$

where D_{KL} denotes the Kullback-Leibler divergence, and τ is a predefined safety threshold (empirically set to 0.05). When visual reasoning is coherent and the divergence remains below τ, α_t is strictly zeroed out, entirely neutralizing the penalty to preserve original generation fluency. Conversely, when the divergence exceeds the threshold, the penalty is fully activated to suppress hallucinated priors.

Entropy-Aware Dynamic Truncation. Furthermore, to dynamically filter long-tail noise without over constraining the model, we calibrate the truncation threshold β based on the Shannon Information Entropy $\mathcal{H}_t$ of the unified distribution p_{unif}. High entropy reflects severe uncertainty, where hallucination tokens frequently emerge. Instead of a linear scaling, we utilize an inverse exponential mapping bounded by $[\beta_{\text{min}}, \beta_{\text{max}}]$ to formulate the adaptive threshold β_t:

$$\begin{aligned} \mathcal{H}_t &= -\sum_{w \in \mathcal{V}} p_{\text{unif}}(w) \log p_{\text{unif}}(w), \\ \beta_t &= \beta_{\text{max}} - (\beta_{\text{max}} - \beta_{\text{min}}) \exp(-\gamma \mathcal{H}_t), \end{aligned} \tag{8}$$

where γ governs the entropy decay rate (further details regarding the configuration and empirical calibration of γ are provided in Appendices A and D). Under this formulation, when the model hesitates (high $\mathcal{H}_t$), the exponential term decays to zero, pushing β_t towards β_{max} for rigorous truncation. Conversely, during confident predictions (low $\mathcal{H}_t$), β_t relaxes towards β_{min}.

To prevent catastrophic truncation during extreme uncertainty, we introduce an absolute Top-k (k=5) fallback mechanism, ensuring the valid candidate set $\mathcal{V}_{\text{top}}$ always retains the k most probable tokens.

Table 1: Performance on the POPE benchmark. Best results in green. We report accuracy (Acc.) and F1-score (F1) across three settings (*Random*, *Popular*, *Adversarial*) and their *Average*.

Model	Method	Random		Popular		Adversarial		Average	
		Acc. ↑	F1 ↑	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑
Qwen2.5-VL-7B	Regular	83.20 ↑0.0	80.09 ↑0.0	82.30 ↑0.0	79.11 ↑0.0	81.86 ↑0.0	78.73 ↑0.0	82.45 ↑0.0	79.31 ↑0.0
	Greedy	84.56 ↑1.4	81.83 ↑1.7	84.33 ↑2.0	81.61 ↑2.5	83.56 ↑1.7	80.79 ↑2.1	84.15 ↑1.7	81.41 ↑2.1
	ICD ((Wang et al., 2024))	85.00 ↑1.8	82.57 ↑2.5	84.76 ↑2.5	82.60 ↑3.5	82.46 ↑0.6	81.72 ↑3.0	84.07 ↑1.6	82.30 ↑3.0
	VCD ((Leng et al., 2024))	85.60 ↑2.4	83.38 ↑3.3	84.80 ↑2.5	82.52 ↑3.4	83.53 ↑1.7	81.30 ↑2.6	84.64 ↑2.2	82.40 ↑3.1
	Marine ((Zhao et al., 2025))	87.60 ↑4.4	86.39 ↑6.3	84.53 ↑2.2	84.50 ↑5.4	84.10 ↑2.2	84.35 ↑5.6	85.41 ↑3.0	85.08 ↑5.8
	TIImage (ours)	89.50 ↑6.3	88.47 ↑8.4	87.90 ↑5.6	86.94 ↑7.8	86.80 ↑4.9	85.92 ↑7.2	88.07 ↑5.6	87.11 ↑7.8
Qwen3-VL-8B	Regular	86.70 ↑0.0	86.42 ↑0.0	85.86 ↑0.0	85.66 ↑0.0	85.03 ↑0.0	85.02 ↑0.0	85.86 ↑0.0	85.70 ↑0.0
	Greedy	87.56 ↑1.9	87.10 ↑1.7	86.60 ↑0.7	86.30 ↑0.6	85.86 ↑1.8	85.66 ↑1.6	86.67 ↑1.5	86.35 ↑1.3
	ICD ((Wang et al., 2024))	88.40 ↑1.7	87.90 ↑1.5	87.50 ↑1.6	87.60 ↑1.9	85.96 ↑0.9	85.74 ↑0.7	87.29 ↑1.4	87.08 ↑1.4
	VCD ((Leng et al., 2024))	87.73 ↑1.0	86.45 ↑0.0	86.63 ↑0.8	86.25 ↑0.6	85.23 ↑0.2	85.65 ↑0.6	86.53 ↑0.7	86.12 ↑0.4
	Marine ((Zhao et al., 2025))	89.53 ↑2.8	88.41 ↑2.0	86.46 ↑0.6	86.44 ↑0.8	85.03 ↑0.0	85.14 ↑0.1	87.01 ↑1.1	86.66 ↑1.0
	TIImage (ours)	91.57 ↑4.9	91.04 ↑4.6	88.23 ↑2.4	87.96 ↑2.3	85.90 ↑0.9	85.76 ↑0.7	88.57 ↑2.7	88.26 ↑2.6
InternVL2.5-8B	Regular	82.10 ↑0.0	82.57 ↑0.0	81.16 ↑0.0	81.55 ↑0.0	79.70 ↑0.0	80.79 ↑0.0	80.99 ↑0.0	81.64 ↑0.0
	Greedy	83.43 ↑1.3	83.75 ↑1.2	82.10 ↑0.9	82.19 ↑0.6	80.56 ↑0.9	81.72 ↑0.9	82.03 ↑1.0	82.55 ↑0.9
	ICD ((Wang et al., 2024))	84.56 ↑2.5	83.67 ↑1.1	84.23 ↑3.1	84.29 ↑2.7	80.33 ↑0.6	80.94 ↑0.1	83.04 ↑2.1	82.97 ↑1.3
	VCD ((Leng et al., 2024))	84.33 ↑2.2	85.63 ↑3.1	83.56 ↑2.4	84.67 ↑3.1	81.56 ↑1.9	82.68 ↑1.9	83.15 ↑2.2	84.33 ↑2.7
	Marine ((Zhao et al., 2025))	84.23 ↑2.1	84.30 ↑1.7	85.63 ↑4.5	85.16 ↑3.6	83.46 ↑3.8	84.92 ↑4.1	84.44 ↑3.5	84.79 ↑3.2
	TIImage (ours)	90.13 ↑8.0	89.32 ↑6.8	88.60 ↑7.4	87.86 ↑6.3	87.33 ↑7.6	86.69 ↑5.9	88.69 ↑7.7	87.96 ↑6.3

4 Experimental Results

Tables 1 and 2 summarize the results on POPE and CHAIR.

Results on POPE Benchmark. As shown in Table 1, TIImage consistently outperforms existing methods across all backbones and splits. On Qwen2.5-VL-7B, it yields a 5.6% accuracy and 7.8% F1 improvement, reaching an average accuracy of 88.07%, surpassing the larger Qwen3-VL-8B baseline (85.86%). Under the adversarial setting, TIImage effectively suppresses co-occurrence priors: it elevates Qwen2.5-VL-7B's accuracy to 86.80% (+4.9% over the 81.86% baseline), outperforming VCD and Marine. Steady improvements across the Qwen and InternVL series across scales validate its strong generalization.

Table 2: Performance on the CHAIR benchmark. Best results in green. We report CHAIR_S (C_S), CHAIR_I (C_I), and recall (Rec.) across three MLLMs (*Qwen2.5-VL*, *Qwen3-VL*, *InternVL2.5*).

Method	Qwen2.5-VL			Qwen3-VL			InternVL2.5			Average		
	$C_S \downarrow$	$C_I \downarrow$	Rec. ↑	$C_S \downarrow$	$C_I \downarrow$	Rec. ↑	$C_S \downarrow$	$C_I \downarrow$	Rec. ↑	$C_S \downarrow$	$C_I \downarrow$	Rec. ↑
Regular	35.8 ↓0.0	11.0 ↓0.0	51.9 ↑0.0	47.0 ↓0.0	10.1 ↓0.0	64.5 ↑0.0	35.2 ↓0.0	9.7 ↓0.0	50.9 ↑0.0	39.3 ↓0.0	10.3 ↓0.0	55.8 ↑0.0
Greedy	34.2 ↓1.6	10.7 ↓0.3	53.6 ↑1.7	41.4 ↓5.6	8.7 ↓1.4	65.4 ↑0.9	34.0 ↓1.2	7.7 ↓2.0	51.1 ↑0.2	36.5 ↓2.8	9.0 ↓1.2	56.7 ↑0.9
ICD	40.4 ↑4.6	11.8 ↑0.8	51.8 ↓0.1	44.6 ↓2.4	10.4 ↑0.3	62.7 ↓1.8	38.2 ↑3.0	9.5 ↓0.2	46.7 ↓4.2	41.1 ↑1.7	10.6 ↑0.3	53.7 ↓2.0
VCD	33.9 ↓1.9	9.7 ↓1.3	53.0 ↑1.1	42.8 ↓4.2	9.4 ↓0.7	63.7 ↓0.8	34.8 ↓0.4	8.4 ↓1.3	47.4 ↓3.5	37.2 ↓2.2	9.2 ↓1.1	54.7 ↓1.1
Marine	33.8 ↓2.0	9.5 ↓1.5	51.1 ↓0.8	43.0 ↓4.0	8.0 ↓2.1	66.1 ↑1.6	31.8 ↓3.4	8.5 ↓1.2	53.8 ↑2.9	36.2 ↓3.1	8.7 ↓1.6	57.0 ↑1.2
TIImage (ours)	33.6 ↓2.2	8.6 ↓2.4	67.3 ↑15.4	43.6 ↓3.4	8.1 ↓2.0	68.6 ↑4.1	32.4 ↓2.8	7.7 ↓2.0	66.2 ↑15.3	36.5 ↓2.8	8.1 ↓2.2	67.4 ↑11.6

Table 3: Ablation study on the position of visual text.

Method	Accuracy ↑	F1-score ↑	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	Recall ↑
Qwen2.5-VL	81.86 ↑0.0	78.73 ↑0.0	35.8 ↓0.0	11.0 ↓0.0	51.9 ↑0.0
Top	82.16 ↑0.3	80.07 ↑1.3	38.2 ↑2.4	10.1 ↓0.9	63.1 ↑11.2
Left	83.66 ↑1.8	82.63 ↑3.9	43.0 ↑7.2	9.5 ↓1.5	63.9 ↑12.0
Right	83.90 ↑2.0	82.63 ↑3.9	37.2 ↑1.4	8.8 ↓2.2	63.8 ↑11.9
Bottom	84.53 ↑2.7	83.55 ↑4.8	33.8 ↓2.0	9.5 ↓1.5	51.1 ↓0.8

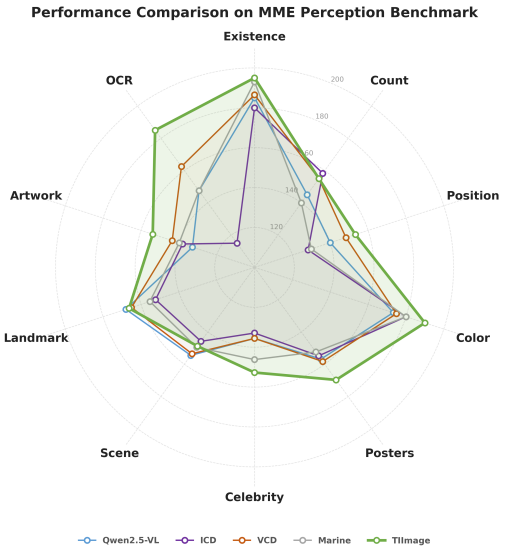


Figure 4: MME benchmark evaluation. Results across 10 diverse sub-tasks demonstrate the comprehensive perception capabilities of TIImage.

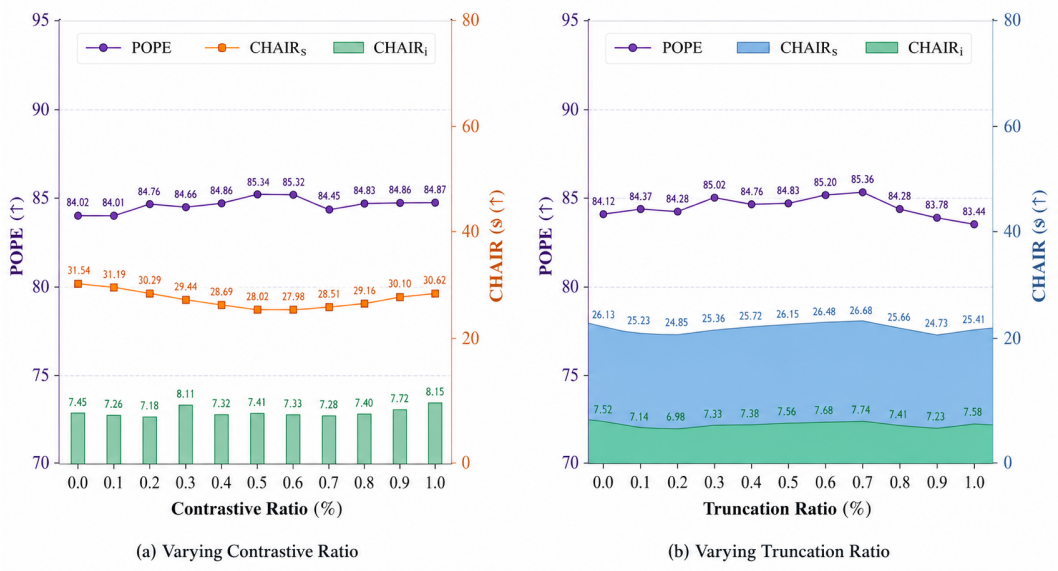
Results on the CHAIR Benchmark. Table 2 shows that TIImage balances descriptive richness with hallucination mitigation across three MLLMs. TIImage achieves the highest average Recall of 67.4 (+11.6% over baseline), capturing fine-grained visual details rather than relying on conservative outputs to avoid errors. Despite this higher informativeness, TIImage strongly curbs hallucinations, yielding the lowest instance-level CHAIR_I of 8.1. While Marine scores marginally lower on sentence-level CHAIR_S, it suffers a substantially lower Recall. Thus, TIImage effectively suppresses hallucinations without sacrificing perceptual granularity.

Results on the MME Benchmark. We further evaluate TIImage on the MME benchmark to test object comprehension and fine-grained perception. As shown in Figure 4, TIImage consistently improves on hallucination-sensitive subsets, outperforming the Qwen2.5-VL baseline by 12.5 points on average and showing clear advantages in spatial and attribute reasoning. Full results appear in Appendix E (8).

4.1 Ablation Study

To validate TIImage's components, we conduct in-depth ablation studies using Qwen2.5-VL-7B.

Impact of Contrastive and Truncation Ratios. As shown in Figure 5, performance remains stable across contrastive ratios from 0.0 to 1.0. However, the optimal truncation ratio β falls within $\beta \leq 0.6$. Exceeding this threshold degrades performance, indicating that overly aggressive truncation may filter out critical semantic information.



(a) Varying Contrastive Ratio (b) Varying Truncation Ratio

Figure 5: (Left) Results under different contrastive ratios α. (Right) Results under different truncation ratios β on POPE and CHAIR.

Impact of Visual Text Position. Table 3 ablates visual text placement without contrastive decoding. Bottom placement achieves the highest Accuracy and F1-score on POPE while effectively suppressing hallucinations (yielding the lowest CHAIR$_S$ and competitive CHAIR$_I$). While top and left placements improve Recall, they exacerbate hallucinations. Thus, we adopt bottom placement as the default.

Impact of TIImage Decoding. Table 4 details the ablation of the TCD strategy. TIImage$_{\text{w/o TCD}}$ improves Accuracy and reduces CHAIR$_S$, affirming the benefits of visual alignment, though a slight Recall drop suggests overly conservative generation. Integrating the full TCD strategy effectively mitigates this issue, increasing Recall to 67.3 while further lowering CHAIR$_I$ (8.6) and CHAIR$_S$ (33.6). This indicates TCD enhances perceptual richness without merely truncating outputs to avoid errors.

Inference Latency. Evaluated on Qwen2.5-VL-7B (Table 5), TIImage incurs a modest latency overhead over vanilla decoding ($\times$1.91), yet remains more efficient than contrastive methods like VCD and ICD across per-token latency, throughput, and total time cost.

Table 4: Ablation study of the decoding strategy.

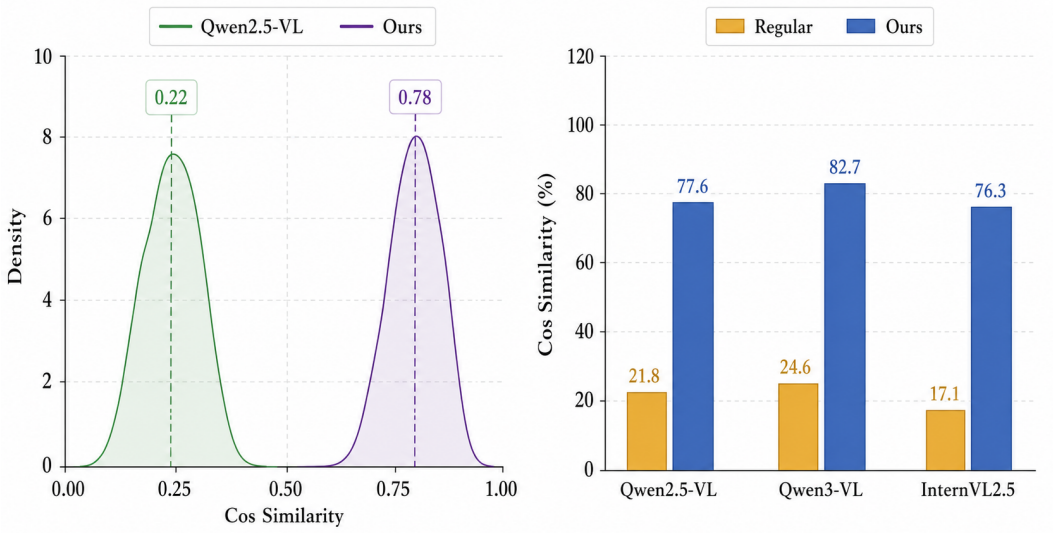
Method	Accuracy ↑	F1-score ↑	CHAIR$_S$ ↓	CHAIR$_I$ ↓	Recall ↑
Qwen2.5-VL	81.86 ↑0.0	78.73 ↑0.0	35.8 ↓0.0	11.0 ↓0.0	51.9 ↑0.0
TIImage$_{\text{w/o TCD}}$	84.53 ↑2.7	83.55 ↑4.8	33.8 ↓2.0	9.5 ↓1.5	51.1 ↓0.8
TIImage	85.43 ↑3.6	84.96 ↑6.2	33.6 ↓2.2	8.6 ↓2.4	67.3 ↑15.4

TIImage reduces modality misalignment. Figure 6 shows embedding text into visual signals boosts text-vision cosine similarity from ∼0.2 to ∼0.8, effectively bridging the modality gap. Consistent gains across Qwen2.5-VL, Qwen3-VL, and InternVL2.5 prove TIImage is a generalized, architecture-agnostic alignment method.

Attribution. To further investigate the intrinsic mechanisms of our method, Table 6 presents the ablation study. *Rendered Random Text* fills the white bottom region with random text; *Blank Padding Only* retains the white bottom region without rendering any text; *TIImage (No CD)* renders text onto the image without classifier-free guidance; *TIImage (Static CD)* fixes $\alpha = 0.6$ and $\beta = 0.1$; and

Table 5: Performance comparison of SOTA methods and TIImage in latency, throughput, and time cost for 80 tokens. The best and suboptimal results **among contrastive decoding methods** are highlighted in green and blue, respectively.

Method	Latency ↓ (ms/token)	Throughput ↑ (token/ms)	Time Cost ↓ for 80 Tokens (ms)
Qwen2.5-VL	33.1 (×1.00)	0.030 (×1.00)	2651.7 (×1.00)
Greedy	32.7 (×0.99)	0.031 (×1.01)	2616.6 (×0.99)
VCD	65.6 (×1.98)	0.015 (×0.50)	5248.0 (×1.98)
Marine	68.8 (×2.08)	0.015 (×0.48)	5505.9 (×2.08)
ICD	65.3 (×1.97)	0.015 (×0.51)	5221.1 (×1.97)
TIImage	63.1 (×1.91)	0.016 (×0.53)	5049.2 (×1.90)



(a) Cos Similarity Distribution (b) Cos Similarity Comparison

Figure 6: Visualization of Modality Alignment.

TIImage represents our full framework. These results strongly demonstrate that TIImage is a holistic methodology rather than a simple stacking of features. Simple layout changes (Blank Padding), unrelated OCR cues (Random Text), or simply injecting instructions without decoding intervention (Pure Text-in-Image) actually disrupt the model's original attention and lead to performance drops. Furthermore, applying a rigid contrastive penalty (Static RCD) leads to severe over-penalization. The significant performance gain of Full TIImage proves the irreplaceable synergy between input-level visual instruction augmentation and our dynamic JSD/entropy-aware decoding.

Table 6: Attribution on the MME benchmark using the InternVL2.5-8B.

Strategy	**MME Score**
Rendered Random Text	1524.31
Blank Padding Only	1515.63
TIImage (No CD)	1463.17
TIImage (Static CD)	1397.14
Full TIImage (Dynamic CD)	**1693.14**

5 Conclusion

We propose TIImage, a novel decoding paradigm that mitigates MLLM hallucinations via modality alignment. By embedding text constraints into images, TIImage offers a training-free, task-agnostic framework that avoids expensive fine-tuning while significantly reducing computational overhead. Extensive experiments show TIImage achieves superior hallucination mitigation and enhances overall perception. While the current static bottom-text layout may restrict multimodal fusion in complex images, future work will explore adaptive text positioning based on visual spatial distribution to further optimize attention alignment.

AI USE STATEMENT

In this work, we used generative AI tools for code completion, unit test generation, and language editing. We have not used generative AI tools for formulating research hypotheses or drafting core scientific arguments. All generated code snippets were independently executed, unit-tested, and verified for correctness by the authors. We take full responsibility for the final content and technical artifacts presented in this work.

ETHICS STATEMENT

This work focuses on foundational MLLMs algorithms and is evaluated entirely on standard, publicly available benchmark datasets. To the best of our knowledge, this work does not involve human participants, proprietary personal data, or direct applications with potential negative societal impacts. All scientific sources and baseline works have been properly cited.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, all algorithmic procedures, mathematical formulations, and experimental configurations are thoroughly documented. Specifically, the core formulation of TIImage is detailed in Section 4, and the complete step-by-step inference pipeline is summarized in Algorithm 1 (Appendix A.3). Comprehensive implementation details, including the font and canvas padding specifications for Text-in-Image rendering, the entropy-based theoretical derivation for parameter γ, evaluation protocols, and exact decoding hyperparameters across all baselines, are provided in Appendices A, C, and D. Furthermore, all experiments are conducted on publicly available benchmarks (POPE, CHAIR, MME, and LLaVA-Bench) using open-source MLLM backbones (Qwen2.5-VL, Qwen3-VL, and InternVL2.5) with fixed random seeds. The complete source code and configuration files are included in the supplementary material and will be open-sourced upon publication.

REFERENCES

Z. Bai, P. Wang, T. Xiao, T. He, Z. Han, Z. Zhang, and M. Z. Shou. Hallucination of multimodal large language models: A survey. *CoRR*, abs/2404.18930, 2024.

J. Cha, W. Kang, J. Mun, and B. Roh. Honeybee: Locality-enhanced projector for multimodal LLM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13817–13827. IEEE, 2024.

L. Chen, J. Li, X. Dong, P. Zhang, C. He, J. Wang, F. Zhao, and D. Lin. ShareGPT4V: Improving large multimodal models with better captions. In *European Conference on Computer Vision (ECCV)*, volume 15075, pp. 370–387. Springer, 2024a.

X. Chen, C. Wang, Y. Xue, N. Zhang, X. Yang, Q. Li, Y. Shen, L. Liang, J. Gu, and H. Chen. Unified hallucination detection for multimodal large language models. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 3235–3252, 2024b.

W. Dai, Z. Liu, Z. Ji, D. Su, and P. Fung. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2128–2140, 2023.

S. Dehaene and L. Cohen. The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, 15(6):254–262, 2011.

X. Ding, J. Han, H. Xu, X. Liang, W. Zhang, and X. Li. Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13668–13677, 2024.

M. O. Ernst and H. H. Bülthoff. Merging the senses into a robust percept. *Trends in Cognitive Sciences*, 8(4):162–169, 2004.

C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, Z. Qiu, W. Lin, J. Yang, X. Zheng, K. Li, X. Sun, and R. Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394, 2023.

L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55, 2025.

Q. Huang, X. Dong, P. Zhang, B. Wang, C. He, J. Wang, D. Lin, W. Zhang, and N. Yu. 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, 2024a.

W. Huang, H. Liu, M. Guo, and N. Gong. Visual hallucinations of multi-modal large language models. In *Findings of the Association for Computational Linguistics (ACL)*, pp. 9614–9631, 2024b.

F. Huo, W. Xu, Z. Zhang, H. Wang, Z. Chen, and P. Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2025.

A. Jain, M. Vatsa, and R. Singh. Words over pixels? rethinking vision in multimodal large language models. In *Proceedings of the International Joint Conferences on Artificial Intelligence (IJCAI)*, pp. 10481–10489, 2025.

C. Jiang, H. Xu, M. Dong, J. Chen, W. Ye, M. Yan, Q. Ye, J. Zhang, F. Huang, and S. Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27026–27036. IEEE, 2024.

J. Y. Koh, D. Fried, and R. R. Salakhutdinov. Generating images with multimodal language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2024.

S. Leng, H. Zhang, G. Chen, X. Li, S. Lu, C. Miao, and L. Bing. 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 (CVPR)*, pp. 13872–13882. IEEE, 2024.

X. L. Li, A. Holtzman, D. Fried, P. Liang, J. Eisner, T. Hashimoto, L. Zettlemoyer, and M. Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 12286–12312, 2023a.

Y. Li, Y. Du, K. Zhou, J. Wang, X. Zhao, and J.-R. Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023b.

Q. Lin, Y. Zhu, X. Mei, L. Huang, J. Ma, K. He, Z. Peng, E. Cambria, and M. Feng. Has multimodal learning delivered universal intelligence in healthcare? a comprehensive survey. *Inf. Fusion*, 116: 102795, 2025.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755. Springer, 2014.

H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024.

J. Liu, Y. Fu, R. Xie, R. Xie, X. Sun, F. Lian, Z. Kang, and X. Li. PhD: A ChatGPT-prompted visual hallucination evaluation dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19857–19866. IEEE, 2025a.

S. Liu, H. Ye, and J. Zou. Reducing hallucinations in large vision-language models via latent space steering. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2025b.

Y. Ouali, A. Bulat, B. Martínez, and G. Tzimiropoulos. CLIP-DPO: Vision-language models as a source of preference for fixing hallucinations in LVLMs. In *European Conference on Computer Vision (ECCV)*, volume 15134, pp. 395–413. Springer, 2024.

B. Peng, J. Lu, G. Zhang, and Z. Fang. An information-theoretical framework for understanding out-of-distribution detection with pretrained vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko. Object hallucination in image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4035–4045, 2018.

W. Shi, X. Han, M. Lewis, Y. Tsvetkov, L. Zettlemoyer, and W.-t. Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 783–791, 2024.

J. Su, J. Chen, H. Li, Y. Chen, L. Qing, and Z. Zhang. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 12964–12974, 2025.

S. N. Wadekar, A. Chaurasia, A. Chadha, and E. Culurciello. The evolution of multimodal model architectures. *CoRR*, abs/2405.17927, 2024.

X. Wang, J. Pan, L. Ding, and C. Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics (ACL)*, pp. 15840–15853, 2024.

Y. N. Yum, N. Cohn, and W. K.-W. Lau. Effects of picture-word integration on reading visual narratives in l1 and l2. *Learning and Instruction*, 71:101397, 2021.

L. Zhao, Y. Deng, W. Zhang, and Q. Gu. Mitigating object hallucination in large vision-language models via image-grounded guidance. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2025.

X. Zou, Y. Wang, Y. Yan, Y. Lyu, K. Zheng, S. Huang, J. Chen, P. Jiang, J. Liu, C. Tang, and X. Hu. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. In *Proceedings of the International Conference on Machine Learning (ICML)*. OpenReview.net, 2025.

A Experimental Details for Modality Alignment

We conduct all of the experiments using 2 A40 GPU with 48GB GPU memory. Each single experiment can be run on a single A40 GPU.

A.1 Text-in-Image Rendering Details

We implement the Text-in-Image input on the Qwen2.5-VL-7B model to achieve modal unification by rendering text instructions directly into the image pixel space. For the rendering configuration, we utilize a 26pt black Arial font to ensure readability across varying resolutions. To prevent occluding key visual areas, we vertically extend the image bottom by approximately 10% with pure white padding, where the text is subsequently rendered. Consequently, assuming the original image resolution is $H \times W$, the unified image resolution after rendering becomes $1.1H \times W$. This strategy maximally preserves the original visual content while ensuring text clarity.

A.2 UMAP Visualization Settings

We conduct feature visualization using the adversarial setting of the POPE benchmark (Li et al., 2023b) based on MS-COCO (Lin et al., 2014). As the most challenging configuration available, it effectively evaluates modality alignment under strong interference. We randomly sample 3,000 instances for UMAP analysis, extracting embeddings from the final layer's hidden states of Qwen2.5-VL-7B. In the standard setting with separate inputs, the text representation is derived from the final text token's hidden state, while the image representation is calculated as the average of all image tokens' hidden states.

In the Text-in-Image setting, text is rendered directly into the pixel space rather than being processed as independent text sequences. To address this, we isolate the image tokens corresponding to the newly appended text region. The average hidden state of these specific tokens serves as the text embedding, whereas the overall image embedding remains the average of all image tokens.

For the UMAP dimensionality reduction, we configure the hyperprameters as follows: `n_components`=3, `random_state`=42, `perplexity`=30, and `n_iter`=1000. This specific configuration optimally balances optimization stability with the preservation of local data structures.

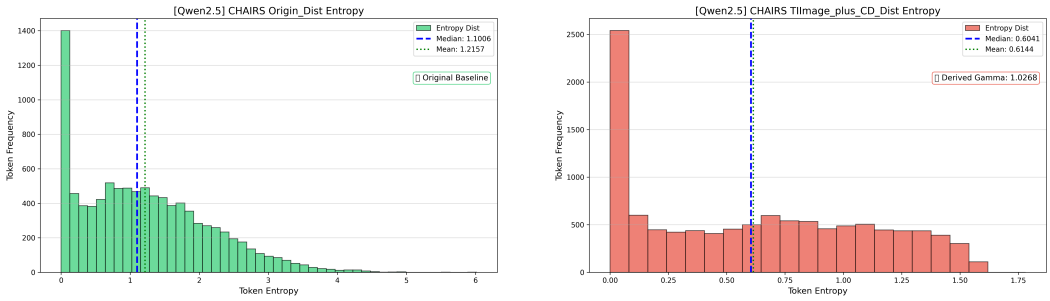
A.3 Algorithm TIImage

Algorithm 1 TIImage (Reading as Seeing)

Require: MLLM $\mathcal{M}_\theta^{\text{MLLM}}$, textual query T, and image V.
Require: Parameters α, β, and unified visual input V_{text}.
1: Initialize empty output sequence $\mathbf{y} = [\,]$.
2: **for** $t = 0, 1, \ldots, N$ **do**
3: Construct degraded image input $\mathbf{x}_{\text{img}}^{(t)} = [\tilde{\mathbf{v}}, \mathbf{y}_{<t}]$.
4: Generate image output logits: $f_\theta(\mathbf{x}_{\text{img}}^{(t)})$.
5: Construct unified visual input $\mathbf{x}_{\text{unif}}^{(t)} = [\mathbf{v}_{\text{text}}, \mathbf{y}_{<t}]$.
6: Generate unified visual output logits: $f_\theta(\mathbf{x}_{\text{unif}}^{(t)})$.
7: Update logits $f_\theta(\mathbf{x}_{\text{unif}}^{(t)})$ and $f_\theta(\mathbf{x}_{\text{img}}^{(t)})$ by Eq. 5.
8: Sample token y_t from logit space calculated by Eq. 6.
9: Update sequence $\mathbf{y} \leftarrow [\mathbf{y}, y_t]$.
10: **end for**
11: **Output:** $\mathbf{y}$.

A.4 POPE Evaluation Protocol

We evaluate object hallucination on the POPE benchmark, derived from the MS-COCO dataset (Lin et al., 2014). This benchmark comprises three settings with varying difficulty and bias: Random,



(a) Original Entropy Distribution

(b) TIImage + CD Entropy Distribution

Figure 7: Comparison of token entropy distributions. (a) shows the original distribution without intervention, while (b) demonstrates the distribution after applying TIImage Contrastive Decoding.

Popular, and Adversarial. We report Accuracy and F1-score as the primary metrics. To ensure reproducibility, we employ a consistent sampling strategy across all experiments. Unless otherwise specified, decoding parameters are set to a temperature of 0.6, top-p of 0.9, and a maximum generation length of 64, with the random seed fixed to 42.

A.5 Gamma Selection Method

The hyperparameter γ in our TIImage Contrastive Decoding (TCD) strategy controls the decay rate of the dynamic truncation threshold. To systematically determine an optimal γ, we propose an entropy-based anchor method.

First, we sample a small subset of the evaluation dataset to generate the pure contrastive probability distribution (prior to truncation). We compute the Shannon entropy H for every generated token and determine the median entropy, H_{median}, across this subset.

To establish a mathematically grounded baseline, we define γ such that the exponential decay term in the dynamic β calculation reaches its half-life exactly at the median entropy:

$$\exp(-\gamma \cdot H_{\text{median}}) = 0.5$$

Solving this equation yields the optimal recommended value:

$$\gamma = \frac{\ln(2)}{H_{\text{median}}}$$

This derivation ensures that for tokens exhibiting average uncertainty (i.e., median entropy), the dynamic truncation bound is positioned precisely at the midpoint of the predefined $[\beta_{\min}, \beta_{\max}]$ interval. This provides a mathematically rigorous and dataset-adaptive parameter configuration. As illustrated in Figure 7, we use Qwen2.5-VL-7B on the CHAIR benchmark as a representative example to analyze the token entropy distribution before and after applying TIImage. The original distribution (left) exhibits a pronounced long tail of high-entropy tokens (extending beyond 4.0), indicating substantial generation uncertainty. Following the intervention of TIImage Contrastive Decoding with our mathematically derived γ (≈ 1.03), the overall entropy is significantly reduced, with the median dropping from 1.10 to 0.60. More crucially, the high-entropy long tail is effectively suppressed and strictly bounded below 1.75. This visualization confirms that our dynamically calculated γ precisely truncates highly uncertain candidate tokens, thereby mitigating hallucination risks at the decoding level.

B Experimental Details for Statistical Bias

We evaluate performance using the CHAIR metric on 500 randomly sampled MS-COCO images under two configurations. In the standard setting, we utilize the prompt "Describe this image in detail" to guide generation. Conversely, the image-only setting omits all text instructions, forcing the model to rely solely on visual features for autoregressive prediction, thereby exposing its vulnerability to statistical priors. To capture the model's intrinsic distribution biases and ensure a fair

comparison, identical decoding strategies are applied across both settings. Specifically, we employ a sampling strategy with temperature set to 1.0, top-p at 1.0, and a maximum generation length of 512 tokens, while fixing the random seed to 42 to guarantee reproducibility.

C Datasets and Evaluation Metrics

In alignment with established evaluations from previous studies (Zhao et al., 2025; Leng et al., 2024; Dai et al., 2023), we assess our proposed TIImage using the following metrics:

CHAIR (Rohrbach et al., 2018) is designed to assess the degree to which generated captions are faithful to the visual content of the input image. It comprises two complementary variants, namely CHAIR_S, which evaluates hallucinations at the sentence level, and CHAIR_I, which measures object-level errors by computing the proportion of incorrectly mentioned objects relative to all referenced objects in a caption. In our experiments, we adopt the MS-COCO (Lin et al., 2014) dataset val2014 split, which provides annotations for 80 object categories. From this split, we randomly sample 500 images for evaluation and prompt the MLLM with "Please describe this image in detail." to generate captions. The CHAIR metric reports both CHAIR_S and CHAIR_I scores, which correspond to per-sentence and per-instance evaluations, respectively, and are formally defined as follows:

$$\text{CHAIR}_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all objects}\}|}$$

$$\text{CHAIR}_S = \frac{|\{\text{hallucinated captions}\}|}{|\{\text{all captions}\}|}$$

POPE (Li et al., 2023b) is a visual question answering-based benchmark designed to quantify object hallucinations in MLLMs. Under this protocol, the model is queried with questions of the form "Is [object] in this image?" and is required to provide a binary response. To explicitly constrain the output space, we further append the instruction "Please answer yes or no." to each query. To construct the queried object set, we adopt three distinct sampling strategies, namely random, popular, and adversarial, which are designed to reflect different levels of object co-occurrence bias. We report results on all three settings to comprehensively evaluate model robustness under diverse hallucination scenarios.

MME (Fu et al., 2023) is a comprehensive benchmark for assessing MLLMs, which organizes evaluation tasks into two primary groups, namely perception and cognition. The perception group focuses on fine-grained visual understanding and covers tasks including existence, counting, spatial location, coarse color recognition, poster recognition, celebrity identification, scene classification, landmark recognition, artwork identification, and OCR. The cognition group evaluates higher-level reasoning abilities. In this work, we evaluate our method on the full perception subset of MME to comprehensively assess both hallucination mitigation and general visual perception capabilities.

D Baselines and Implementation Details

Unless otherwise specified, TIImage evaluations adopt a standard configuration using greedy decoding (`do_sample = False`, `temperature = 1.0`), with the contrastive ratio $\alpha = 0.6$, dynamic truncation bounds $\beta \in [0.1, 0.6]$, and a recommended generic decay factor of $\gamma = 1.15$ for plug-and-play zero-shot inference. For optimal performance, this parameter can also be adaptively calibrated prior to inference based on the target dataset's entropy distribution, as detailed in Appendix A.5. For VCD, we set `do_sample = True`, `temperature = 1.0`, `noise_step = 999`, and the plausibility constraint hyperparameter λ to 0.1, while α, which controls the degree of contrastive emphasis, was set to 1, following the default settings from the original code and literature. For ICD, we enable sampling with `do_sample = True` and set the decoding hyperparameters to `temperature = 1.0`, $\alpha = 1$, and $\beta = 0.1$. In addition, we construct the prompt "You are a confused image caption model" and prepend it to the query prompt. For Marine, we set `do_sample = False`, `temperature = 0.6`, top-$p = 0.9$, and the guidance strength to 0.7, following the default settings from the original code and literature. The guidance prompt templates used in our experiments are provided in Table 7. For all baseline methods and TIImage, we use task-specific maximum generation lengths to ensure fair and consistent evaluation. Specifically, we set

Table 7: Details of the guidance prompts that we used in the experiments.

Template Type	Prompt Template
CHAIR	This image contains <OBJECT_GROUNDING>. Based on this, <QUERY>
	The image contains the following objects: <OBJECT_GROUNDING>. Given these detected objects, <QUERY>
	This image shows the following objects: <OBJECT_GROUNDING>. Using this information, <QUERY>
	The objects found in this image are: <OBJECT_GROUNDING>. Considering this list of objects, <QUERY>
POPE task	This image contains only the following objects: <OBJECT_GROUNDING>. Do not assume anything beyond these objects. Based solely on this list, <QUERY>
	The detected objects in the image are: <OBJECT_GROUNDING>. Answer based only on these objects. <QUERY>
	This image shows the following objects: <OBJECT_GROUNDING>. You must answer using only the objects in this list. Given these detected objects, <QUERY>
	The objects found in this image are limited to: <OBJECT_GROUNDING>. You should rely strictly on this list of objects and make no other guesses. Based on this, <QUERY>
MME	List of detected objects in the image: <OBJECT_GROUNDING_A> <OBJECT_GROUNDING_B> Based on the detected objects above, <QUERY>
	The most prominent objects detected are: <OBJECT_GROUNDING_A> <OBJECT_GROUNDING_B> Given these findings, <QUERY>
	The following objects were detected in the image: <OBJECT_GROUNDING_A> <OBJECT_GROUNDING_B> With this information, <QUERY>
	Here is a list of all objects detected in the image: <OBJECT_GROUNDING_A> <OBJECT_GROUNDING_B> Do not infer or hallucinate any additional objects. Using only the detected objects, <QUERY>

the maximum number of generated tokens to 512 for CHAIR, and to 64 for both POPE and MME. All baseline parameter settings adhere to the default configurations specified in their respective papers.

E Performance on MME Benchmark

To comprehensively evaluate the general capabilities of TIImage, we conducted a detailed assessment on the full perception subset of the MME benchmark (Fu et al., 2023). As shown in Table 8, TIImage achieves the best total perception scores across different architectures, including Qwen2.5-VL-7B, Qwen3-VL-8B, and InternVL2.5-8B. Specifically, the score of Qwen2.5-VL-7B improves significantly from 1538.5 in the Regular setting to 1669.5, substantially surpassing state-of-the-art methods such as VCD. Similarly, InternVL2.5-8B exhibits a large gain, increasing from 1517.0 to 1693.1, while Qwen3-VL-8B achieves an impressive peak score of 1802.4. These results indicate that TIImage enhances overall visual understanding through modality alignment and debiased decoding.

In addition, TIImage performs well on core metrics related to object hallucination, including Existence, Count, Position, and Color. On the challenging Count task, InternVL2.5-8B achieves a remarkable score of 200.0, substantially outperforming both the Regular baseline and VCD. Qwen2.5-VL-7B also shows clear advantages in Position and Color recognition, indicating a more precise cap-

Table 8: Performance evaluation on MME Perception benchmark. The best results are highlighted with a green background. We report the scores on 10 sub-tasks and the total Perception score.

Model	Method	Exist ↑	Count ↑	Position ↑	Color ↑	Poster ↑	Celeb ↑	Scene ↑	Landmark ↑	Artwork ↑	OCR ↑	Perception ↑
Qwen2.5-VL-7B	Regular	185.0	145.0	140.0	173.3	156.8	135.6	154.5	168.0	132.8	147.5	1538.5
	Greedy	190.0	165.0	150.0	180.0	161.9	143.8	154.0	169.5	139.8	137.5	1591.5
	ICD (Wang et al., 2024)	180.0	158.3	128.3	180.0	154.8	132.9	145.8	152.3	138.0	115.0	1485.4
	VCD (Leng et al., 2024)	186.5	155.0	148.3	175.0	158.3	135.6	153.5	165.0	143.5	162.5	1583.2
	Marine (Zhao et al., 2025)	193.0	140.0	130.0	180.0	152.4	146.2	149.5	155.3	139.8	147.5	1533.6
	TIImage (ours)	195.0	155.0	153.3	190.0	169.7	152.7	148.8	166.3	153.8	185.0	1669.5
Qwen3-VL-8B	Regular	190.0	173.3	153.3	195.0	182.0	178.8	151.0	177.3	154.5	185.0	1740.2
	Greedy	192.0	168.3	158.3	195.0	181.6	178.8	151.5	178.0	156.8	177.5	1737.9
	ICD (Wang et al., 2024)	190.0	168.3	145.0	185.0	175.5	163.2	143.5	176.8	158.3	185.0	1690.6
	VCD (Leng et al., 2024)	195.0	170.0	151.7	200.0	181.0	180.6	154.8	167.8	154.3	177.5	1732.4
	Marine (Zhao et al., 2025)	200.0	173.3	145.0	195.0	182.0	178.8	151.5	177.3	156.8	177.5	1737.1
	TIImage (ours)	193.3	186.7	180.0	193.3	182.3	171.8	140.0	187.0	178.0	190.0	1802.4
InternVL2.5-8B	Regular	195.0	155.0	126.7	158.3	159.2	130.3	147.0	159.8	153.3	132.5	1517.0
	Greedy	195.0	155.0	131.7	170.0	161.9	132.9	151.5	161.8	159.5	155.0	1574.3
	ICD (Wang et al., 2024)	195.0	141.7	120.0	156.7	166.3	133.8	147.3	149.5	150.5	160.0	1520.7
	VCD (Leng et al., 2024)	200.0	160.0	148.3	175.0	169.0	147.1	157.5	172.8	157.0	162.5	1649.2
	Marine (Zhao et al., 2025)	200.0	160.0	126.7	158.3	161.9	135.6	154.5	168.0	154.5	160.0	1579.5
	TIImage (ours)	186.7	200.0	173.3	193.3	155.1	124.7	136.0	176.0	168.0	180.0	1693.1

ture of fine-grained semantic details with unified visual signal inputs. Importantly, these gains do not come at the cost of general visual capabilities. TIImage maintains or improves performance on tasks such as OCR, Landmark, and Artwork; for example, Qwen2.5-VL-7B attains an OCR score of 185.0. Overall, these results demonstrate that TIImage effectively enhances visual perception while mitigating hallucinations.

F Performance on long-text benchmark

To evaluate TIImage in long-text generation and complex visual understanding, we conduct a systematic evaluation on the LLaVA-Bench (Liu et al., 2023; 2024). LLaVA-Bench comprises diverse visual question answering and complex instruction tasks, providing a comprehensive assessment of image understanding and generation stability.

Table 9: LLaVA-Bench performance (accuracy %).

Method	Qwen2.5-VL	Qwen3-VL	Intern2.5-VL
Regular	41.67 ↑0.0	43.33 ↑0.0	28.33 ↑0.0
Greedy	42.85 ↑1.2	45.00 ↑1.7	30.00 ↑1.7
VCD	31.67 ↓10.0	46.67 ↑3.3	31.67 ↑3.3
ICD	33.33 ↓8.3	41.67 ↓1.7	25.00 ↓3.3
TIImage (ours)	46.67 ↑5.0	46.67 ↑3.3	36.67 ↑8.3

All experiments employ decoding and implementation parameters consistent with Appendix D. The maximum generation length is set to 512 tokens to accommodate long-text responses and ensure a fair comparison. Following (Zou et al., 2025), we use `gpt-5.1-2025-11-13` as the judge model to automatically evaluate the generated responses.

Table 9 reports the accuracy of different methods on this benchmark. TIImage achieves the highest accuracies across Qwen2.5-VL, Qwen3-VL, and InternVL2.5, reaching 46.67%, 46.67%, and 36.67%, respectively. These results indicate that TIImage stably preserves question-relevant visual information in long-context generation scenarios. Furthermore, compared to existing contrastive decoding methods, TIImage demonstrates improved robustness under extended generation lengths.