Where Are MLLMs Looking When They Hallucinate? Mitigating Visual Hallucination via Gaze Steering

Yebo Wu[1*] **Han Jin**[1*] **Feng Liu**[2] **Changwang Zhang**[2] **Jun Wang**[2†] **Li Li**[1†]
[1]University of Macau [2]OPPO Research Institute
{yc37926,mc56727,llili}@um.edu.mo
{liufeng4hit,junwang.lu}@gmail.com changwangzhang@foxmail.com

Abstract

Multimodal Large Language Models (MLLMs) have achieved remarkable progress, yet visual hallucination remains a critical barrier to reliable deployment. This paper aims to answer a fundamental question: *when MLLMs hallucinate, where are they actually looking?* By tracing the inter-layer consistency of visual attention, we reveal that hallucinated reasoning trajectories bifurcate into two pathological states: *Attention Locking*, where the gaze becomes overly focal and rigidly anchored to limited visual evidence, and *Attention Collapse*, where attention becomes excessively dispersed and fails to ground reasoning in meaningful cues. In contrast, faithful reasoning maintains a dynamic equilibrium between established visual evidence and broader peripheral exploration. Building on this insight, we propose REGAZE, a training-free framework that redirects the model's gaze toward this healthy equilibrium to mitigate hallucinations. Specifically, REGAZE monitors layer-wise gaze consistency and triggers timely interventions whenever an unhealthy gaze state emerges. For the locked state, REGAZE redistributes attention from over-dominant critical tokens to peripheral regions, encouraging the model to explore richer visual details. For the collapsed state, it gathers scattered attention from non-critical tokens and injects it into critical anchors, amplifying key visual signals. Experimental results demonstrate that REGAZE enables MLLMs to process visual information more effectively and significantly mitigates hallucinations.

1 Introduction

Multimodal Large Language Models (MLLMs) [2] have significantly advanced vision-language intelligence, demonstrating remarkable success across various tasks such as visual question answering [18, 8, 43], visual grounding [33, 27], and fine-grained visual understanding [45, 47, 46]. Most MLLMs bridge visual perception and language reasoning through a modular architecture: a pre-trained vision encoder (e.g., CLIP [28]) extracts high-dimensional visual representations, which are projected into the textual embedding space and processed by a Large Language Model (LLM). Benefiting from the collaborative training and adaptation strategies of modern LLMs [31, 40, 41, 42], this paradigm effectively grounds linguistic generation in rich visual evidence.

Despite these advances, MLLMs remain vulnerable to visual hallucination, producing fluent descriptions that are inconsistent with the actual visual content [14]. These errors can manifest as fabricated objects [21, 30], incorrect attributes [34], or misidentified spatial relations [38], posing serious risks in safety-critical decision-making applications [12]. Such hallucinations indicate that the model may lose a stable connection between its language reasoning and the visual evidence.

[*]Equal contribution.
[†]Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Existing hallucination mitigation methods can be broadly grouped into three categories [39]: training-based alignment, such as reinforcement learning from human feedback (RLHF) [32] and high-quality instruction tuning [50]; post-hoc correction with external tools or iterative re-decoding [48]; and inference-time intervention, including contrastive decoding [19, 36] and attention calibration [17, 24]. Although these methods are effective in different scenarios, they largely overlook the model's internal visual reasoning process. As a result, they provide limited insight into how a response that should be grounded in visual evidence gradually drifts into hallucination. This motivates a fundamental question: *when MLLMs hallucinate, where are they actually looking?*

In this paper, we approach hallucination through the lens of internal gaze dynamics. Rather than treating hallucination as an output-level failure, we trace the inter-layer consistency of visual attention throughout the reasoning process and examine how the model's gaze evolves from perception to answer generation. Our analysis shows that hallucinated reasoning trajectories do not simply ignore the image; instead, they drift toward two pathological extremes. ① *Attention Locking*: The gaze becomes overly focal, rigidly anchored to a few visual patches and unable to incorporate emerging evidence. ② *Attention Collapse*: Visual focus becomes excessively dispersed, preventing the model from grounding its reasoning in meaningful cues. Faithful reasoning, by contrast, maintains a dynamic equilibrium between established visual evidence and broader peripheral exploration. These findings suggest a new principle for hallucination mitigation: rather than correcting hallucinated outputs after generation, the model's gaze should be recalibrated as soon as its visual dynamics begin to drift.

Motivated by this principle, we propose REGAZE, a training-free framework that mitigates MLLM hallucinations by redirecting the model's gaze toward a healthy equilibrium. Specifically, REGAZE monitors layer-wise gaze consistency and maintains a sliding window over recent attention patterns to identify critical and non-critical visual tokens. Once an unhealthy gaze state emerges, REGAZE triggers a diagnosis-specific intervention. For locked states, it redistributes attention from over-dominant tokens to peripheral regions, encouraging the model to move beyond local fixation and explore richer visual details. For collapsed states, it gathers scattered attention from non-critical tokens and injects it into critical anchors, amplifying key visual signals submerged in noise. Through these interventions, REGAZE effectively reduces hallucinations while enhancing the model's ability to reason from clear visual evidence. Our main contributions are summarized as follows:

- We introduce an internal-gaze perspective to understand MLLM hallucinations. By tracing the inter-layer consistency of visual attention, we reveal that hallucinated reasoning trajectories drift toward two pathological extremes: *Attention Locking* and *Attention Collapse*.
- We propose REGAZE, a training-free framework that mitigates hallucinations by redirecting the model's gaze toward a healthy dynamic equilibrium. REGAZE monitors layer-wise gaze consistency and applies timely interventions to correct unhealthy gaze states.
- We conduct extensive experiments on mainstream hallucination and multimodal reasoning benchmarks. Results show that REGAZE significantly outperforms existing baselines, reducing sentence-level hallucination rates by up to **5.7%** over the strongest method.

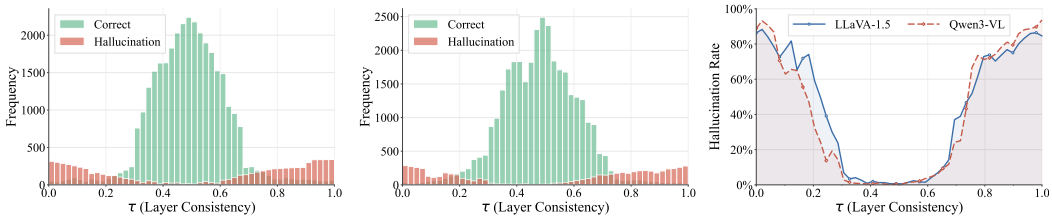
2 Gaze Dynamics: Preliminaries and Analysis

In this section, we investigate the internal gaze dynamics of MLLMs during reasoning, examining whether hallucinated responses exhibit attention patterns distinct from faithful ones.

2.1 Tracking the Gaze via Inter-Layer Rank Consistency

Notation. Consider an MLLM with L transformer layers. At layer l, the multi-head self-attention module produces an attention tensor $A^{(l)} \in \mathbb{R}^{H\times N\times N}$, where H denotes the number of heads and N denotes the sequence length. Let $\mathcal{V}$ be the set of visual token indices, and let $\mathcal{Q}$ be the set of query tokens. Since our goal is to characterize where the model allocates its gaze during reasoning, we measure the total attention received by each visual token from the query side, averaged across heads. Formally, for each visual token $j \in \mathcal{V}$, we define its received attention at layer l as:

$$s_j^{(l)} = \frac{1}{H}\sum_{h=1}^{H}\sum_{q\in\mathcal{Q}} A_{h,q,j}^{(l)}. \tag{1}$$



(a) Sample Distribution: LLaVA-1.5. (b) Sample Distribution: Qwen3-VL. (c) Hallucination Rate Distribution.

Figure 1: Global distribution of inter-layer rank consistency on POPE [22]. Panels (a) and (b) compare the distributions of faithful and hallucinated samples on LLaVA-1.5-7B and Qwen3-VL-8B, respectively, while panel (c) reports the hallucination rate across different consistency intervals.

Collecting these scores over all visual tokens yields a layer-wise attention profile $\mathbf{s}^{(l)} = \{s_j^{(l)}\}_{j \in \mathcal{V}}$. The ranking induced by $\mathbf{s}^{(l)}$ captures the model's visual priority at layer l, providing a compact representation of its internal gaze state.

Inter-Layer Rank Consistency. To trace how visual gaze evolves during reasoning, we consider not only the attention magnitude assigned to each token, but also the stability of token-importance rankings across adjacent layers. This ranking-based view captures whether the model keeps reinforcing the same visual anchors or reorganizes its gaze as computation deepens. We quantify this behavior using the Kendall rank correlation coefficient (τ) [4]. However, selecting the top-K tokens independently at each layer can introduce instability: a token near the boundary may move from rank K to $K+1$ with nearly unchanged attention, yet disappear from the comparison set. To avoid this artifact, we adopt a *Fixed Reference Set* strategy. For adjacent layers $l-1$ and l, we first select the top-K visual tokens from the previous layer:

$$\mathcal{R}^{(l)} = \mathrm{TopK}_{j \in \mathcal{V}} \left(s_j^{(l-1)} \right). \tag{2}$$

We then compare the relative rankings of these fixed reference tokens $\mathcal{R}^{(l)}$ under $\mathbf{s}^{(l-1)}$ and $\mathbf{s}^{(l)}$, and define the inter-layer rank consistency as:

$$\tau^{(l)} = \frac{N_c - N_d}{K(K-1)/2}, \tag{3}$$

where N_c and N_d denote the numbers of concordant and discordant token pairs between the two rankings. A high τ indicates that the model continues to prioritize the same visual anchors across layers, corresponding to a rigid and stagnant gaze. In contrast, a low τ indicates that the relative order of previously important visual tokens is heavily disrupted, suggesting a disorganized gaze. Thus, τ provides a compact layer-wise signal for tracing the evolution of visual attention.

2.2 Dual Pathologies of Visual Hallucination

With inter-layer rank consistency, we revisit the central question: *when MLLMs hallucinate, where are they actually looking?* We conduct experiments on POPE [22], splitting samples into faithful and hallucinated groups according to response correctness, and compute τ across the network depth.

Global Distribution. We first summarize each sample by averaging its inter-layer consistency scores across all layers. As shown in Figure 1(a) and (b), faithful and hallucinated responses exhibit markedly different global distributions across both LLaVA-1.5-7B and Qwen3-VL-8B. Faithful samples are mostly concentrated in a moderate consistency region near the center of the τ axis, suggesting that correct reasoning maintains a dynamic balance between preserving reliable visual anchors and adapting to surrounding evidence. In contrast, hallucinated samples are much more likely to appear near the two extremes, revealing a polarization of gaze dynamics. High-consistency samples ($\tau \to 1$) correspond to *Attention Locking*, where the model repeatedly reinforces a limited set of visual patches and loses the flexibility to revise its focus. Low-consistency samples ($\tau \to 0$) correspond to *Attention Collapse*, where visual priority becomes disorganized and critical cues are diluted by noisy attention shifts. Figure 1(c) further supports this observation from a complementary perspective: the hallucination rate is consistently higher in both low- and high-consistency intervals, while the middle consistency range exhibits a substantially lower hallucination rate.

TAKEAWAY I*: Faithful reasoning tends to stay in a balanced consistency regime, whereas hallucination emerges when gaze dynamics polarize toward either excessive stability or excessive instability.*

Layer-Wise Evolution. The global distributions above reveal the overall tendency of faithful and hallucinated samples across the consistency spectrum, but they do not explain how such polarization develops during inference. To examine this process, we track τ layer by layer and divide hallucinated samples into two groups according to their global consistency: low-consistency hallucinations ($\tau < 0.5$) and high-consistency hallucinations ($\tau > 0.5$). As shown in Figure 2, the three trajectories are initially similar and begin to diverge only as the model progresses from early visual alignment to deeper multimodal reasoning.

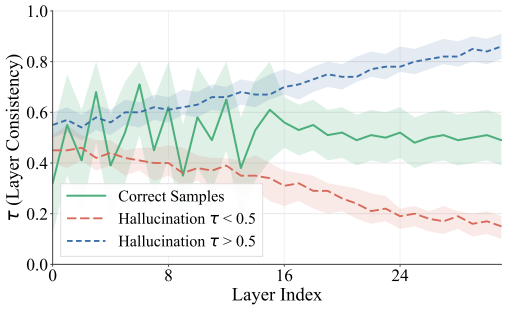


Figure 2: Layer-wise evolution of inter-layer rank consistency on LLaVA-1.5-7B using POPE.

• **Shallow-Layer Homogeneity.** In the early layers, all groups fluctuate within a similar consistency range. This suggests that hallucination is not simply caused by an immediately abnormal visual focus; at this stage, faithful and hallucinated responses still share broadly similar visual alignment dynamics.

• **Middle-Layer Transition.** As visual evidence is integrated with the language context, the trajectories begin to diverge. Faithful samples remain near the middle of the consistency axis, indicating that the model preserves useful visual anchors while still allowing its gaze to adapt. In contrast, hallucinated samples start drifting in two directions: one branch becomes increasingly unstable with decreasing consistency, while the other becomes increasingly rigid with steadily rising consistency.

• **Deep-Layer Polarization.** In deeper layers, the separation becomes more pronounced. Low-consistency hallucinations continue to drift downward, indicating *Attention Collapse*, where the model repeatedly reorganizes its visual priority and fails to preserve stable evidence. High-consistency hallucinations instead move upward, indicating *Attention Locking*, where the model over-reinforces a fixed set of visual anchors and becomes insensitive to alternative cues. Faithful samples instead contract toward the moderate region, thereby avoiding rigid fixation and uncontrolled gaze shifting.

TAKEAWAY II*: Visual hallucination emerges through a progressive polarization of gaze dynamics in mid-to-deep layers. Layer-wise consistency therefore serves as an effective online signal for detecting when the model begins to drift away from the balanced reasoning regime.*

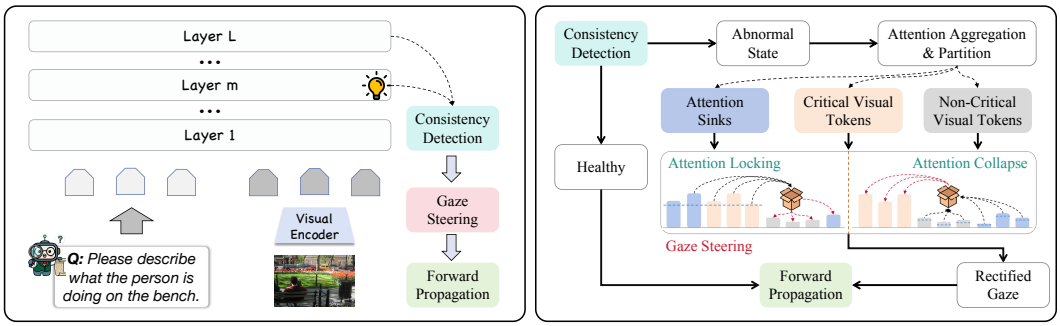
3 Our Method: REGAZE

Guided by the analysis in Section 2, we propose REGAZE, a training-free framework that mitigates MLLM hallucinations through adaptive gaze steering. As shown in Figure 3, REGAZE continuously monitors gaze dynamics during inference and, once an abnormal gaze state is detected, applies pathology-aware calibration to steer the model toward a balanced and healthy gaze regime.

3.1 Gaze Monitoring and Token Partitioning

Detection Onset. As observed in Section 2.2, faithful and hallucinated trajectories remain indistinguishable in shallow layers. Their divergence becomes pronounced only in the mid-to-deep layers, where visual evidence is progressively integrated into multimodal reasoning. Therefore, REGAZE begins monitoring from a designated layer l_s and leaves earlier layers untouched. For each layer $l \geq l_s$, REGAZE computes the inter-layer rank consistency τ; deviations from the healthy range indicate that the current trajectory is drifting toward either Attention Locking or Attention Collapse.

Sliding-Window Gaze Aggregation. Once an abnormal gaze state is detected, REGAZE must determine which visual tokens should be suppressed and which should receive additional attention. Indiscriminately intervening on all visual tokens may corrupt useful evidence or amplify irrelevant background regions. We therefore partition visual tokens into two groups: *critical visual tokens*, which serve as the main visual anchors during reasoning, and *non-critical visual tokens*, which correspond to peripheral regions that may still contain overlooked cues. However, token categorization based on



(a) Overall Framework.

(b) Workflow of Gaze Steering.

Figure 3: Illustration of the REGAZE framework. (a) REGAZE tracks inter-layer visual consistency during MLLM reasoning to identify abnormal gaze dynamics. (b) When an unhealthy trajectory is detected, REGAZE applies pathology-aware gaze calibration to mitigate either Attention Locking or Attention Collapse, steering the model back toward a balanced visual grounding state.

a single layer can be unreliable, since attention may transiently shift to irrelevant textures, sink-like tokens, or short-lived reasoning artifacts. To obtain a more stable estimate of token importance, REGAZE aggregates attention patterns over a recent sliding window of layers. Let W denote the window size. For the current layer l, the cached layer set is defined as:

$$\mathcal{W}^{(l)} = \{\, r \mid \max(1, l - W + 1) \le r \le l \,\}. \tag{4}$$

The smoothed attention is obtained by averaging attention patterns over the cached layer set:

$$\bar{A}^{(l)} = \frac{1}{|\mathcal{W}^{(l)}|} \sum_{r \in \mathcal{W}^{(l)}} A^{(r)}. \tag{5}$$

Here, $\bar{A}^{(l)} \in \mathbb{R}^{H \times N \times N}$ suppresses transient layer-wise fluctuations while preserving persistently influential visual tokens, thereby providing a stable basis for subsequent token partitioning.

Attention Sink Filtering. Before splitting visual tokens, we first identify *attention sinks*, which absorb attention from most queries regardless of semantics. These tokens typically correspond to punctuation, separators, or other non-visual attractors, and serve as reservoirs for redundant attention. For each token j, we compute the fraction of tokens that effectively attend to it as:

$$c_j^{(l)} = \frac{1}{HN} \sum_{h=1}^{H} \sum_{q=1}^{N} \mathbb{I}\left[\bar{A}_{h,q,j}^{(l)} > \epsilon\right], \tag{6}$$

where ϵ is the threshold for determining whether token j is effectively attended to by a given token. If $c_j^{(l)}$ exceeds the sink threshold ρ, token j is marked as an attention sink. Let $\mathcal{S}^{(l)}$ denote the resulting sink set. These sink tokens are excluded from the critical/non-critical visual-token partition, allowing REGAZE to distinguish meaningful visual evidence from systematic non-visual attractors.

Visual Token Partitioning. After excluding attention sinks, REGAZE partitions the remaining visual tokens based on their accumulated importance. Let $\mathcal{V}^{(l)} = \mathcal{V} \setminus \mathcal{S}^{(l)}$ denote the set of valid visual tokens. For each token $j \in \mathcal{V}^{(l)}$, we compute its smoothed attention intensity as:

$$s_j^{(l)} = \frac{1}{H} \sum_{h=1}^{H} \sum_{q \in \mathcal{Q}} \bar{A}_{h,q,j}^{(l)}. \tag{7}$$

We compute the average visual intensity as $\mu_V^{(l)} = \frac{1}{|\mathcal{V}^{(l)}|} \sum_{j \in \mathcal{V}^{(l)}} s_j^{(l)}$ and use it to define a scale-adaptive threshold $T^{(l)} = \gamma \mu_V^{(l)}$, which separates visual tokens into critical and non-critical sets:

$$\mathcal{K}^{(l)} = \{\, j \in \mathcal{V}^{(l)} \mid s_j^{(l)} \ge T^{(l)} \,\}, \qquad \mathcal{U}^{(l)} = \{\, i \in \mathcal{V}^{(l)} \mid s_i^{(l)} < T^{(l)} \,\}. \tag{8}$$

Here, $\mathcal{K}^{(l)}$ denotes critical visual tokens that serve as persistent evidence anchors, while $\mathcal{U}^{(l)}$ denotes non-critical tokens forming the peripheral visual context. Stabilized by the sliding window, this adaptive partition provides the token-level basis for subsequent gaze calibration.

3.2 Pathology-Aware Gaze Steering

At each monitored layer, REGAZE uses the inter-layer rank consistency τ to decide whether intervention is needed. We define a healthy interval $[\tau_{\text{low}}, \tau_{\text{high}}]$: if τ falls within this range, the model is considered to maintain a balanced gaze and the attention map is left unchanged. Values above τ_{high} indicate Attention Locking, while values below τ_{low} indicate Attention Collapse.

For readability, we omit the layer superscript and write $\mathcal{K}, \mathcal{U}, \mathcal{S}$, and T for $\mathcal{K}^{(l)}, \mathcal{U}^{(l)}, \mathcal{S}^{(l)}$, and $T^{(l)}$, respectively. Let a_j denote the attention assigned to token j within one head, satisfying $\sum_j a_j = 1$. The goal is to redistribute attention, thereby restoring gaze patterns to a healthy state.

Attention Locking: Diffusing Over-Focal Mass. When $\tau > \tau_{\text{high}}$, the model enters an over-concentrated gaze state, where a small set of tokens persistently absorbs dominant attention and broader visual evidence is under-explored. To relax this rigid focus, REGAZE applies peak smoothing over $\mathcal{G} = \mathcal{K} \cup \mathcal{U} \cup \mathcal{S}$. Specifically, we reuse the adaptive threshold T from token partitioning as the clipping threshold, and collect the excess attention from tokens in $\mathcal{G}$ whose attention exceeds T:

$$\mathcal{P}_{\text{lock}} = \{\, j \in \mathcal{G} \mid a_j \geq T \,\}, \qquad \Delta_{\text{lock}} = \sum_{j \in \mathcal{P}_{\text{lock}}} (a_j - T). \tag{9}$$

The attention values of these clipped tokens are then set to the threshold: $a'_j = T,\ \forall j \in \mathcal{P}_{\text{lock}}$. The reclaimed attention is then redistributed to the under-explored visual details and global context in $\mathcal{G}$, $\mathcal{B}_{\text{lock}} = \{\, i \in \mathcal{G} \mid a_i < T \,\}$, in proportion to their current attention:

$$a'_i = a_i + \Delta_{\text{lock}} \cdot \frac{a_i}{\sum_{m \in \mathcal{B}_{\text{lock}}} a_m}, \quad \forall i \in \mathcal{B}_{\text{lock}}. \tag{10}$$

This diffusion step smooths the over-focal distribution, allows low-attention visual tokens and non-dominant sinks to receive compensation, and forces the model to re-explore broader visual evidence.

Attention Collapse: Concentrating Scattered Mass. When $\tau < \tau_{\text{low}}$, the model enters an overly dispersed gaze state, where critical visual cues are weakened by attention scattered across non-critical tokens and attention sinks. To restore stable evidence anchors, REGAZE performs the opposite correction by extracting a fraction β of attention mass from redundant sources:

$$\mathcal{P}_{\text{collapse}} = \mathcal{U} \cup \mathcal{S}, \qquad \Delta_{\text{collapse}} = \beta \sum_{i \in \mathcal{P}_{\text{collapse}}} a_i. \tag{11}$$

The attention values of these source tokens are reduced accordingly: $a'_i = (1-\beta)a_i,\ \forall i \in \mathcal{P}_{\text{collapse}}$. The collected attention is then injected into critical visual anchors according to their current saliency:

$$a'_j = a_j + \Delta_{\text{collapse}} \cdot \frac{a_j}{\sum_{n \in \mathcal{K}} a_n}, \quad \forall j \in \mathcal{K}. \tag{12}$$

This concentration step strengthens essential visual signals that are otherwise submerged in noisy attention and re-establishes them as reliable anchors for subsequent reasoning.

Forward Propagation. After gaze steering, the rectified attention map is fed into the remaining computation, thereby calibrating the subsequent reasoning process. Since REGAZE requires no parameter updates, it provides a lightweight inference-time mechanism to mitigate hallucinations.

4 Experiments

4.1 Experimental Setup

Baselines. We compare REGAZE with the following methods. DoLa [7] performs contrastive decoding between layers to enhance factual knowledge. VCD [19] contrasts predictions from clean and visually distorted inputs. OPERA [13] penalizes over-trusted tokens through an attention-based decoding constraint. SID [16] leverages discrepancies in internal states for self-correction. EAZY [3] suppresses hallucination by masking image tokens identified as hallucinatory, while DaID [39] introduces dual-anchor introspective decoding for hallucination mitigation.

Evaluation Models. We validate REGAZE on a diverse set of MLLMs from different families and parameter scales, including LLaVA-1.5 [23] at 7B and 13B scales, LLaVA-NeXT (7B), MiniGPT-4 [51], Qwen3-VL [2] at 4B and 8B scales, and InternVL3-8B [6].

Table 1: Performance comparison of REGAZE against state-of-the-art methods on visual hallucination benchmarks. The best results are highlighted in **bold**, and the second-best results are underlined.

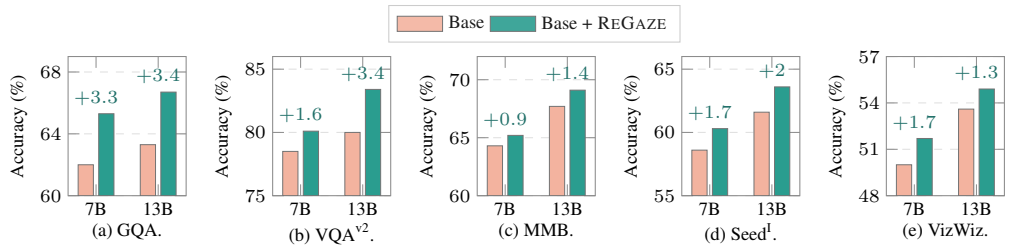
Method	POPE		CHAIR		MME				
	Acc.↑	F1↑	$CHAIR_S$↓	$CHAIR_I$↓	Existence↑	Count↑	Position↑	Color↑	Total↑
LLaVA-1.5-7B									
Greedy	81.38	82.20	49.6	14.4	173.45	122.72	113.39	149.92	559.48
Beam Search	84.66	84.60	46.3	12.9	175.67	124.67	114.00	151.00	565.34
DoLa [7]	84.06	84.62	47.1	13.8	180.10	127.40	119.30	154.60	581.40
VCD [19]	84.66	84.52	49.2	14.8	184.66	137.33	128.67	153.00	603.66
OPERA [13]	84.88	85.21	45.4	12.7	180.67	133.33	111.67	123.33	549.00
SID [16]	84.82	85.50	44.2	12.2	183.90	132.20	127.80	155.90	599.80
EAZY [3]	84.97	85.78	38.8	11.4	182.29	131.73	126.93	155.21	596.16
DaID [39]	85.08	85.92	35.9	11.3	186.14	130.70	154.46	162.38	633.68
REGAZE	**86.62**	**86.84**	**30.2**	**10.5**	**188.32**	**141.56**	**157.84**	**165.10**	**652.82**
LLaVA-NeXT									
Greedy	83.78	82.24	32.8	9.1	179.78	127.40	121.92	151.82	580.92
Beam Search	83.77	81.69	33.0	9.2	180.27	128.90	120.34	155.29	584.80
DoLa [7]	84.53	84.77	31.3	9.0	184.45	133.07	126.72	160.38	604.62
VCD [19]	83.37	83.95	32.8	8.9	187.20	**140.16**	139.91	157.41	624.68
OPERA [13]	84.37	84.21	34.0	8.7	183.98	135.85	118.00	131.72	569.55
SID [16]	84.61	85.32	32.8	8.3	185.03	134.96	137.23	161.74	618.96
EAZY [3]	84.91	85.40	26.8	8.3	184.55	134.18	133.52	158.89	611.14
DaID [39]	85.32	85.76	24.2	8.2	189.47	133.12	156.29	165.52	644.40
REGAZE	**87.15**	**87.22**	**21.8**	**7.4**	**191.20**	137.67	**160.12**	**168.30**	**657.29**
Qwen3-VL-8B									
Greedy	84.72	85.25	28.4	8.2	187.86	138.51	151.28	159.34	636.99
Beam Search	84.92	85.18	28.8	8.4	187.15	139.62	152.08	160.55	639.40
DoLa [7]	85.26	85.44	27.6	8.1	189.50	141.73	154.88	160.29	646.40
VCD [19]	85.37	85.82	28.0	8.1	190.88	145.15	155.22	161.74	652.99
OPERA [13]	85.74	86.42	27.1	8.0	187.98	140.87	151.55	161.82	642.22
SID [16]	85.83	86.64	27.8	7.9	190.13	143.28	157.46	158.40	649.27
EAZY [3]	86.09	86.22	26.9	7.7	191.74	143.91	158.15	160.55	654.35
DaID [39]	86.78	86.42	21.5	7.5	192.34	144.86	160.42	165.28	662.90
REGAZE	**88.89**	**89.08**	**18.1**	**6.3**	**193.17**	**147.62**	**162.80**	**172.82**	**676.41**

Evaluation Benchmarks. We evaluate REGAZE on two categories of visual tasks. The first category focuses on visual hallucination, including POPE [22], CHAIR [29], and MME [9], which are used to assess whether MLLMs can faithfully ground their responses in the visual input. The second category focuses on visual reasoning, including GQA [15], VQA^{v2} [10], MMB [25], $Seed^{I}$ [20], and VizWiZ [11], which are used to evaluate the models' general visual understanding capabilities.

Implementation Details. We start monitoring inter-layer rank consistency from layer $l_s = 16$ and set the reference size for Kendall-τ computation to $K = 20$. The healthy consistency interval is $[\tau_{\text{low}}, \tau_{\text{high}}] = [0.3, 0.7]$. For gaze aggregation and attention-sink detection, we use a window size of $W = 8$, effective-attention threshold $\epsilon = 0.01$, and sink threshold $\rho = 0.8$. For token partitioning and pathology-aware gaze steering, we set the critical-token scaling factor to $\gamma = 1.2$ and the collapse extraction ratio to $\beta = 0.1$. The reported results are averaged over multiple runs.

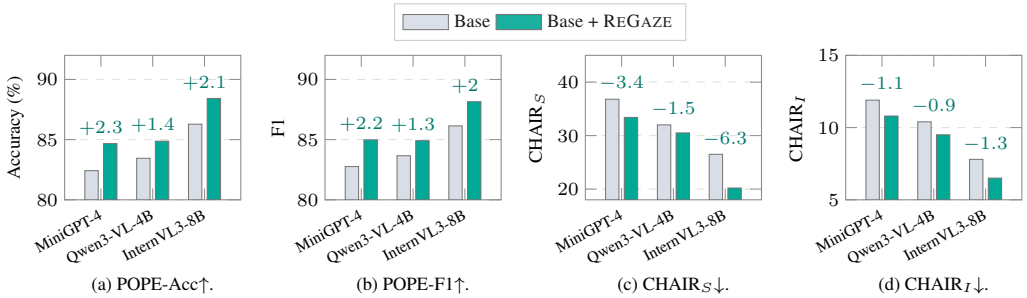
4.2 Main Results

Effectiveness on Hallucination Mitigation. As shown in Table 1, REGAZE significantly mitigates MLLM hallucinations across all evaluated settings. On POPE, REGAZE achieves the best accuracy and F1-score for LLaVA-1.5-7B, LLaVA-NeXT, and Qwen3-VL-8B, outperforming the strongest baseline by up to 2.11% in accuracy and 2.44% in F1-score. On CHAIR, REGAZE substantially reduces both sentence-level and instance-level hallucination rates; for example, on LLaVA-1.5-7B, $CHAIR_S$ decreases from 35.9% to 30.2% and $CHAIR_I$ decreases from 11.3% to 10.5% compared with the best baseline, yielding reductions of 5.7% and 0.8%, respectively. On Qwen3-VL-8B, $CHAIR_S$ decreases from 21.5% to 18.1% and $CHAIR_I$ decreases from 7.5% to 6.3%. On MME, REGAZE also obtains the highest overall score, with total-score gains of 19.14, 12.89, and 13.51,



(a) GQA. (b) VQA^{v2}. (c) MMB. (d) Seed^{I}. (e) VizWiz.

Figure 4: Performance of LLaVA-1.5-7B and LLaVA-1.5-13B on visual reasoning benchmarks.



(a) POPE-Acc↑. (b) POPE-F1↑. (c) CHAIR$_S$↓. (d) CHAIR$_I$↓.

Figure 5: Model generalization analysis of REGAZE across diverse MLLM architectures.

respectively. These results demonstrate that REGAZE can effectively suppress MLLM hallucinations by steering the model's visual gaze during reasoning.

Improvement on Visual Reasoning. As shown in Figure 4, REGAZE also brings consistent gains on general visual reasoning benchmarks. Compared with the base LLaVA-1.5 models, adding REGAZE improves performance on all datasets for both the 7B and 13B scales. For the 7B model, REGAZE increases accuracy by 3.3%, 1.6%, 0.9%, 1.7%, and 1.7% on GQA, VQA^{v2}, MMB, Seed^{I}, and VizWiz, respectively, with an average gain of 1.84%. The improvement is even more evident on the 13B model, where REGAZE achieves an average gain of 2.30%. These results demonstrate that the proposed gaze steering mechanism also enhances the model's visual reasoning ability.

4.3 Ablation Study

Table 2 evaluates the contribution of the two calibration branches in REGAZE, showing that removing either branch consistently degrades performance across all evaluated settings. Without the Attention Locking calibration, POPE accuracy drops by 2.45%, 1.92%, and 2.46% on the three models, while CHAIR$_S$ increases by 4.9%, 5.6%, and 4.0%. Removing the Attention Collapse calibration causes a comparable decline, with POPE accuracy decreasing by 2.49%, 1.73%, and 2.37% and CHAIR$_S$ increasing by 6.0%, 6.8%, and 3.8%. These results confirm that hallucination mitigation benefits from addressing both over-stable gaze locking and unstable gaze collapse.

Table 2: Ablation study of REGAZE.

Method	POPE		CHAIR	
	Acc.↑	F1↑	CHAIR$_S$↓	CHAIR$_I$↓
LLaVA-1.5-7B				
+ REGAZE	**86.62**	**86.84**	**30.2**	**10.5**
w/o Locking Calib.	84.17	84.24	35.1	12.9
w/o Collapse Calib.	84.13	82.35	36.2	13.0
LLaVA-NeXT				
+ REGAZE	**87.15**	**87.22**	**21.8**	**7.4**
w/o Locking Calib.	85.23	85.45	27.4	9.1
w/o Collapse Calib.	85.42	85.67	28.6	9.2
Qwen3-VL-8B				
+ REGAZE	**88.89**	**89.08**	**18.1**	**6.3**
w/o Locking Calib.	86.43	86.36	22.1	8.8
w/o Collapse Calib.	86.52	86.78	21.9	8.9

4.4 Model Generalization Analysis

Figure 5 evaluates the model generalization capability of REGAZE on MiniGPT-4, Qwen3-VL-4B, and InternVL3-8B. Across the three models, REGAZE consistently improves POPE performance, increasing accuracy by 2.3%, 1.4%, and 2.1%, and F1-score by 2.2%, 1.3%, and 2.0%, respectively. Meanwhile, REGAZE also reduces hallucination rates on CHAIR: CHAIR$_S$ decreases by 3.4%, 1.5%, and 6.3%, while CHAIR$_I$ decreases by 1.1%, 0.9%, and 1.3%. These results indicate that REGAZE provides a broadly applicable mechanism for improving visual grounding and reducing hallucination.

4.5 Hyperparameter Analysis

Impact of the Critical-Token Threshold γ. We study the effect of γ, which controls the threshold for identifying critical visual tokens. As shown in Figure 6, all three models show the same trend: POPE accuracy and F1-score improve as γ increases from 1.0 to 1.2, but decline with larger values. This suggests that a small threshold may include too many peripheral tokens as critical evidence, weakening the distinction between visual anchors and background context. Conversely, an overly large threshold may retain only a few dominant tokens and discard useful supporting cues. The best results at $\gamma = 1.2$ suggest a robust balance between preserving visual anchors and retaining sufficient context.

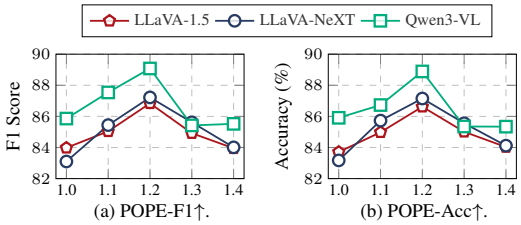


Figure 6: Impact of the critical-token threshold factor γ on the POPE benchmark.

Impact of the Collapse Extraction Ratio β. We further analyze β, which controls the fraction of attention extracted from redundant sources during Attention Collapse calibration. As shown in Figure 7, all three models exhibit the same pattern: performance improves as β increases from 0.05 to 0.10, but declines with larger values. A small β provides insufficient correction for dispersed gaze, leaving critical visual anchors under-emphasized. In contrast, an overly large β may remove too much attention from peripheral tokens, disrupting useful contextual evidence. The best performance at $\beta = 0.10$ indicates that a moderate extraction ratio restores collapsed gaze without over-correcting attention.

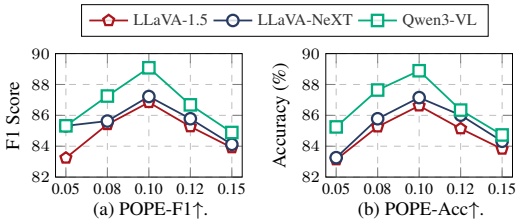


Figure 7: Impact of the collapse extraction ratio β on the POPE benchmark.

5 Related Work

Visual Hallucination. Visual hallucination is a major reliability issue in MLLMs, where models generate objects, attributes, or relations unsupported by the visual input [29, 22, 14]. Existing benchmarks such as CHAIR [29], POPE [22], and MME [9] evaluate this behavior at the output level, showing that hallucination often stems from over-reliance on language priors or dataset co-occurrence biases [26]. Recent studies further examine its internal causes, revealing that visual evidence can be weakened, misplaced, or overwhelmed during multi-layer reasoning [35, 17, 37]. Attention-based analyses identify visual attention sinks [17, 37], insufficient global-local evidence calibration [1], and causal attention patterns that amplify object hallucination [44]. In this paper, we show that hallucination is associated with two pathological forms of layer-wise gaze dynamics.

Inference-Time Intervention. Inference-time intervention has become an appealing way to improve MLLM reliability without costly retraining. Existing methods mainly rely on contrastive decoding or attention calibration. Contrastive decoding suppresses hallucination by comparing predictions under original and perturbed inputs [7, 5, 19, 36], but often incurs additional computational overhead and may impair reasoning capabilities. Attention calibration instead adjusts abnormal attention patterns [37, 1], such as redistributing attention from sink tokens [17] or broadcasting visual-sink head maps [49]. Unlike these methods, which mostly operate on static attention distributions, REGAZE tracks the layer-wise evolution of visual gaze and dynamically steers it during inference.

6 Conclusion

In this paper, we study MLLM hallucination from an internal gaze perspective. By tracing inter-layer visual attention, we identify a layer-wise gaze bifurcation phenomenon, where hallucinated trajectories drift toward either Attention Locking or Attention Collapse. Based on this observation, we introduce REGAZE, a training-free framework that monitors gaze consistency and performs pathology-aware gaze steering to restore healthy visual grounding. Extensive experiments demonstrate that REGAZE significantly mitigates hallucination while improving visual reasoning capabilities.

Acknowledgments and Disclosure of Funding

This work was supported in part by the Science and Technology Development Fund of Macau (Grant Nos. 0107/2024/RIA2 and 0061/2025/RIB2) and the Multi-Year Research Grant of the University of Macau (Grant Nos. MYRG-GRG2023-00211-IOTSC-UMDF and MYRG-GRG2024-00180-IOTSC). It was also supported by the 2024 Nansha District Science and Technology Program, Guangzhou, China (Grant No. 2024ZD005). The authors declare that there are no competing financial interests or relationships that could be perceived to influence this work.

References

[1] Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, Qianying Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29915–29926, 2025.

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[3] Liwei Che, Tony Qingze Liu, Jing Jia, Weiyi Qin, Ruixiang Tang, and Vladimir Pavlovic. Hallucinatory image tokens: A training-free eazy approach to detecting and mitigating object hallucinations in lvlms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21635–21644, 2025.

[4] Jun Chen, Meng Yang, Wenping Gong, and Yang Yu. Multi-neighborhood guided kendall rank correlation coefficient for feature matching. *IEEE Transactions on Multimedia*, 25:7113–7127, 2023.

[5] Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. HALC: Object hallucination reduction via adaptive focal-contrast decoding. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 7824–7846. PMLR, 21–27 Jul 2024.

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198, June 2024.

[7] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 54158–54183, 2024.

[8] Wenlong Fang, Qiaofeng Wu, Jing Chen, and Yun Xue. Notes-guided mllm reasoning: Enhancing mllm with knowledge and visual notes for visual question answering. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19597–19607, 2025.

[9] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. Mme: A comprehensive evaluation benchmark for multimodal large language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025.

[10] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[11] Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018.

[12] Zhiqiang Hu, Mingxing Xu, and Qixiu Cheng. Multimodal large-language model empowering next-generation autonomous driving systems. *Journal of Intelligent and Connected Vehicles*, 8(2):9210059–1–9210059–3, 2025.

[13] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13418–13427, 2024.

[14] Wen Huang, Hongbin Liu, Minxin Guo, and Neil Gong. Visual hallucinations of multi-modal large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9614–9631, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[15] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6693–6702, 2019.

[16] Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 24272–24295, 2025.

[17] Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 87676–87703, 2025.

[18] Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. Natural language understanding and inference with mllm in visual question answering: A survey. *ACM Comput. Surv.*, 57(8), March 2025.

[19] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13872–13882, 2024.

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

[21] Shuo Li, Jiajun Sun, Guodong Zheng, Xiaoran Fan, Yujiong Shen, Yi Lu, Zhiheng Xi, Yuming Yang, Wenming Tan, Tao Ji, Tao Gui, Qi Zhang, and Xuanjing Huang. Mitigating object hallucinations in MLLMs via multi-frequency perturbations. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1230–1247, Suzhou, China, November 2025. Association for Computational Linguistics.

[22] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore, December 2023. Association for Computational Linguistics.

[23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc., 2023.

[24] Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 125–140, Cham, 2025. Springer Nature Switzerland.

[25] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 216–233, Cham, 2025. Springer Nature Switzerland.

[26] Kyungmin Min, Minbeom Kim, Kang-il Lee, Dongryeol Lee, and Kyomin Jung. Mitigating hallucinations in large vision-language models via summary-guided decoding. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4183–4198, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics.

[27] Yuxin Peng, Zishuo Wang, Geng Li, Xiangtian Zheng, Sibo Yin, and Hulingxiao He. A survey on fine-grained multimodal large language models. *Chinese Journal of Electronics*, 35(2):771–803, 2026.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[29] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[30] Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Arik, and Tomas Pfister. Mitigating object hallucination in mllms via data-augmented phrase-level alignment. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 73796–73846, 2025.

[31] He Sun, Jinrui Zhou, Li Li, Yebo Wu, and Mingjun Xiao. Heimdallm: Efficient cloud-assisted federated fine-tuning with zeroth-order rectification for llms. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, page 4463–4474, New York, NY, USA, 2026. Association for Computing Machinery.

[32] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented RLHF. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[33] Wei Tang, Yanpeng Sun, Qinying Gu, and Zechao Li. Visual position prompt for mllm based visual grounding. *IEEE Transactions on Multimedia*, 28:3739–3754, 2026.

[34] David Wan, Han Wang, Ziyang Wang, Elias Stengel-Eskin, Hyunji Lee, and Mohit Bansal. Multimodal fact-level attribution for verifiable reasoning. In Tong Zhang, Miroslav Dudik, Martin Jaggi, Alekh Agarwal, Sharon Li, Dale Schuurmans, Jerry Zhu, Felix Berkenkamp, Hanze Dong, and Alberto Bietti, editors, *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 124706–124736. PMLR, 06–11 Jul 2026.

[35] Sudong Wang, Yunjian Zhang, Yao Zhu, Enci Liu, Jianing Li, Yanwei Liu, and Xiangyang Ji. Shift: Smoothing hallucinations by information flow tuning for multimodal large language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3639–3649, 2025.

[36] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15840–15853, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[37] Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1927–1951, Vienna, Austria, July 2025. Association for Computational Linguistics.

[38] Jing Wu, Suiyao Chen, Alexander Gutfraind, Inseok Heo, Shengjie Liu, Chen Li, Jeremy Curuksu, and Michael Sharps. Building more accountable multi-modal LLMs through spatially-informed visual reasoning. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025.

[39] Yebo Wu, Han Jin, Zhijiang Guo, and Li Li. Spotlight and shadow: Attention-guided dual-anchor introspective decoding for MLLM hallucination mitigation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens, editors, *Findings of the Association for Computational Linguistics: ACL 2026*, pages 13219–13233, San Diego, California, United States, July 2026. Association for Computational Linguistics.

[40] Yebo Wu, Jingguang Li, Zhijiang Guo, and Li Li. Developmental federated tuning: A cognitive-inspired paradigm for efficient LLM adaptation. In *The Fourteenth International Conference on Learning Representations*, 2026.

[41] Yebo Wu, Jingguang Li, Zhijiang Guo, and Li Li. Don't reinvent the wheel, just realign the spokes: Resource-efficient federated fine-tuning via rank-wise expert assembly. In Tong Zhang, Miroslav Dudik, Martin Jaggi, Alekh Agarwal, Sharon Li, Dale Schuurmans, Jerry Zhu, Felix Berkenkamp, Hanze Dong, and Alberto Bietti, editors, *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 136193–136214. PMLR, 06–11 Jul 2026.

[42] Yebo Wu, Jingguang Li, Chunlin Tian, KaHou Tam, Zhijiang Guo, and Li Li. Beyond end-to-end: Dynamic chain optimization for private LLM adaptation on the edge. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens, editors, *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18419–18435, San Diego, California, United States, July 2026. Association for Computational Linguistics.

[43] Yebo Wu, Feng Liu, Ziwei Xie, Changwang Zhang, Jun Wang, and Li Li. Tsembed: Unlocking task scaling in universal multimodal embeddings. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari, editors, *Computer Vision – ECCV 2026*, pages 1–19, Cham, 2026. Springer Nature Switzerland.

[44] Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. Mitigating object hallucination via concentric causal attention. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 92012–92035. Curran Associates, Inc., 2024.

[45] Naen Xu, Jiayi Sheng, Changjiang Li, Chunyi Zhou, Yuyuan Li, Tianyu Du, Jun Wang, Zhihui Fu, Jinbao Li, and Shouling Ji. "I see what you did there": Can large vision-language models understand multimodal puns? In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens, editors, *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9786–9805, San Diego, California, United States, July 2026. Association for Computational Linguistics.

[46] Naen Xu, Jinghuai Zhang, Changjiang Li, Hengyu An, Chunyi Zhou, Jun Wang, Boyu Xu, Yuyuan Li, Tianyu Du, and Shouling Ji. Bridging the copyright gap: Do large vision-language models recognize and respect copyrighted content? *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(42):35949–35957, Mar. 2026.

[47] Naen Xu, Jinghuai Zhang, Changjiang Li, Zhi Chen, Chunyi Zhou, Qingming Li, Tianyu Du, and Shouling Ji. VideoEraser: Concept erasure in text-to-video diffusion models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5954–5983, Suzhou, China, November 2025. Association for Computational Linguistics.

[48] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: hallucination correction for multimodal large language models. *Sci. China Inf. Sci.*, 67(12), 2024.

[49] Xiaofeng Zhang, Yihao Quan, Chen Shen, Chaochen Gu, Xiaosong Yuan, Shaotian Yan, Jiawei Cao, Hao Cheng, Kaijie Wu, and Jieping Ye. Shallow focus, deep fixes: Enhancing shallow layers vision attention sinks to alleviate hallucination in LVLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3512–3534, Suzhou, China, November 2025. Association for Computational Linguistics.

[50] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond multimodal hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2025.

[51] Deyao Zhu, jun chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, editors, *International Conference on Learning Representations*, volume 2024, pages 18378–18394, 2024.

Table 3: Hyperparameter sensitivity of REGAZE on CHAIR using LLaVA-1.5-7B.

Parameter	Value	$\text{CHAIR}_S\downarrow$	$\text{CHAIR}_I\downarrow$
γ	1.0	34.2	11.6
	1.1	32.9	10.7
	1.2	**30.2**	**10.5**
	1.3	31.9	10.8
	1.4	33.7	11.5
β	0.05	35.0	12.9
	0.08	31.2	11.0
	0.10	**30.2**	**10.5**
	0.12	31.9	11.1
	0.15	32.8	12.1
W	0	36.9	13.4
	4	32.1	11.5
	8	**30.2**	**10.5**
	12	33.6	12.0

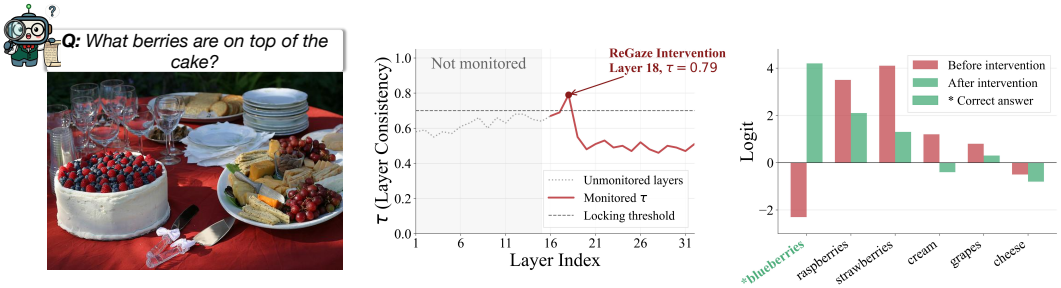


Figure 9: Case study of Attention Locking. Given the question "What berries are on top of the cake?", the model initially over-commits to an incorrect visual cue. REGAZE detects a high-consistency locking state at layer 18 and applies gaze steering, after which the model shifts its prediction from hallucinated berry types toward the correct answer, blueberries.

A More Experimental Results

Impact of the Sliding Window Size W. We analyze the effect of the sliding window size W, which controls how many recent layers are aggregated when estimating token importance. As shown in Figure 8, performance improves when increasing W from 0 to 8 across all three models, indicating that aggregating recent attention patterns provides a more stable estimate than relying on a single layer. However, further enlarging the window to 12 leads to performance degradation, likely because an overly long history mixes outdated attention states with the current layer's attention. The best results are consistently achieved at $W = 8$, suggesting that a moderate window effectively captures persistent visual anchors while suppressing transient layer-wise fluctuations.

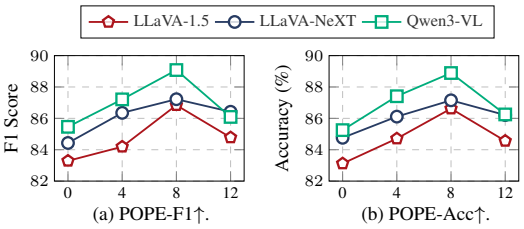


Figure 8: Impact of the sliding window size W on the POPE benchmark.

Hyperparameter Sensitivity on CHAIR. To further validate the robustness of the selected hyperparameters, we report CHAIR results on LLaVA-1.5-7B in Table 3. The trends are consistent with the POPE analysis: $\gamma = 1.2$, $\beta = 0.10$, and $W = 8$ achieve the lowest CHAIR_S and CHAIR_I. This shows that the default setting not only improves discriminative hallucination evaluation, but also reduces object hallucination in image captioning.

Case Study of Attention Locking. Figure 9 illustrates how REGAZE corrects an Attention Locking failure. In this example, the model is asked to identify the berries on top of the cake. Before

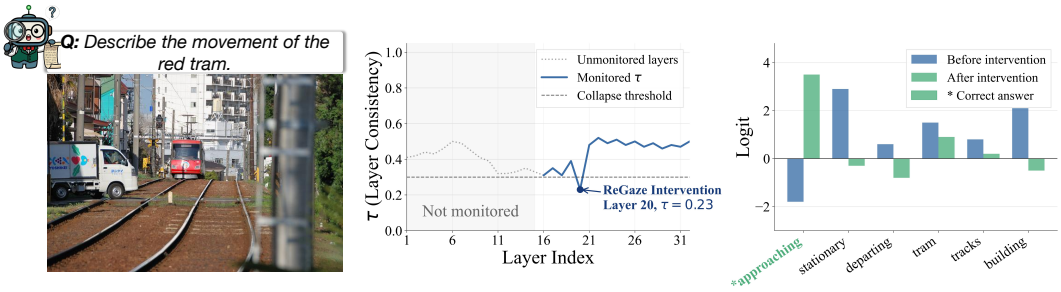
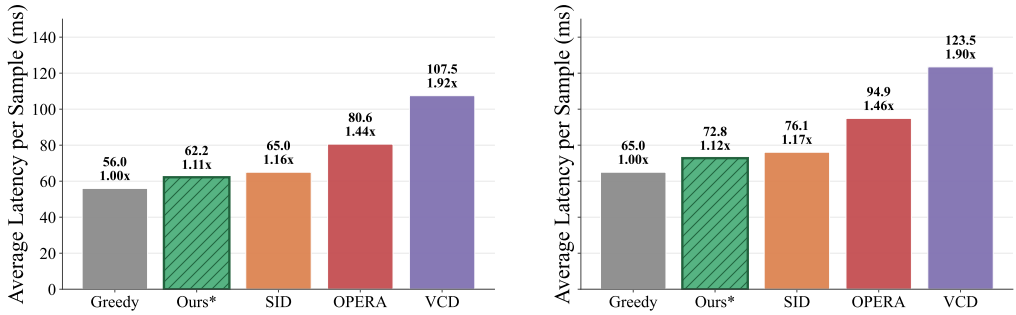


Figure 10: Case study of Attention Collapse. Given the question "Describe the movement of the red tram", the model initially produces dispersed evidence over several distractors. REGAZE detects a low-consistency collapse state at layer 20 and calibrates the scattered attention, increasing the logit of the correct answer, approaching.



(a) LLaVA-1.5-7B. **(b) Qwen3-VL-8B.**

Figure 11: Inference latency comparison on an NVIDIA H800 GPU with batch size 1. REGAZE introduces only modest overhead over greedy decoding while remaining substantially faster than existing inference-time hallucination mitigation methods.

intervention, the prediction logits assign high confidence to incorrect answers such as raspberries and strawberries, indicating that the model has locked onto misleading visual evidence. The middle panel shows that the inter-layer consistency rises above the locking threshold at layer 18, suggesting an overly stable gaze trajectory. After REGAZE intervenes by diffusing the over-concentrated attention, the logit of the correct answer, blueberries, increases substantially, while the hallucinated alternatives are suppressed. This case shows that high consistency can correspond to a rigid but incorrect visual focus, and that relaxing this locked gaze helps the model recover more faithful grounding.

Case Study of Attention Collapse. Figure 10 shows an example where REGAZE mitigates an Attention Collapse failure. The question asks about the movement of the red tram, whose correct description is approaching. Before intervention, the model assigns high logits to distracting concepts such as stationary, tram, tracks, and building, suggesting that its visual evidence is scattered across multiple regions rather than concentrated on the tram's motion. The middle panel shows that the inter-layer consistency falls below the collapse threshold at layer 20, indicating an unstable gaze trajectory. After REGAZE gathers attention from redundant sources and redirects it toward critical visual anchors, the logit of approaching becomes dominant while irrelevant candidates are reduced. This case demonstrates that low consistency can reflect a failure to maintain coherent visual grounding, and that concentrating scattered attention helps recover the correct visual interpretation.

B Overhead Analysis

We further evaluate the inference overhead of REGAZE on an NVIDIA H800 GPU with a batch size of 1. As shown in Figure 11, REGAZE incurs only a small latency increase over greedy decoding. On LLaVA-1.5-7B, greedy decoding requires 56.0 ms per sample, while REGAZE takes 62.2 ms, corresponding to a $1.11\times$ latency ratio. On Qwen3-VL-8B, REGAZE increases latency from 65.0 ms

to 72.8 ms, yielding a similar $1.12\times$ ratio. This modest overhead mainly comes from online gaze monitoring and attention redistribution, both of which are lightweight and do not require additional model forward passes.

Compared with existing inference-time baselines, REGAZE is also substantially more efficient. For LLaVA-1.5-7B, SID, OPERA, and VCD require 65.0 ms, 80.6 ms, and 107.5 ms per sample, respectively, whereas REGAZE remains close to greedy decoding. A similar trend is observed on Qwen3-VL-8B, where SID, OPERA, and VCD take 76.1 ms, 94.9 ms, and 123.5 ms per sample. These results show that REGAZE effectively mitigates hallucinations with negligible inference overhead, making it practical for deployment on existing MLLMs.

C Broader Impacts

Positive Impact. This work aims to improve the reliability of MLLMs by reducing hallucinated visual content and encouraging responses to be better grounded in the input image. Such improvements can benefit applications that require faithful visual understanding, including assistive technologies, educational tools, content moderation, medical or scientific data interpretation, and human-AI collaboration systems. By mitigating unsupported object, attribute, or relation descriptions, REGAZE may help users obtain more trustworthy model outputs and reduce the cost of manually verifying generated responses. Since the proposed method is applied during inference without additional training, it can also provide a practical way to enhance existing MLLMs without requiring large-scale retraining or extra annotated data.

Negative Impact. The proposed method is an inference-time technique and does not require collecting additional user data or training new models, so it is unlikely to introduce substantial new societal risks beyond those already associated with existing MLLMs. Nevertheless, it does not completely eliminate incorrect or unsupported responses, and the model may still inherit biases or limitations from the underlying backbone. Therefore, users should avoid treating the improved outputs as fully reliable in sensitive applications, and standard evaluation and human oversight remain necessary when deploying the system.

D Limitations

First, although we evaluate REGAZE on multiple representative MLLMs and model scales, experiments on even larger backbones, such as models beyond the 13B scale, would further validate its scalability. Second, our experiments mainly focus on open-source MLLMs with accessible intermediate states, which allows us to analyze and steer the visual gaze during inference. Extending the evaluation to more recently released or proprietary MLLMs, when their internal attention states are available, would further demonstrate the compatibility of our method.

E LLM Usage Declaration

We used large language models only for language polishing, grammar checking, and improving the clarity of the manuscript. The research idea, method design, experimental implementation, result analysis, and final writing decisions were completed and verified by the authors.

F Compute Reporting

In this section, we detail the computational resources utilized for all experiments. No models were trained from scratch, nor did we perform any fine-tuning or grid searches over learnable parameters. All experiments were executed on a single local GPU.

Local Computing Environment. All experiments were conducted on a single NVIDIA H800 GPU (80 GB VRAM). The host machine was equipped with an Intel Xeon processor, 256 GB of RAM, and ample storage for model checkpoints and evaluation datasets. Neither distributed computing nor multi-GPU configurations were employed. All models were loaded in FP16 precision.

MLLM Inference Setup. REGAZE is a training-free, inference-time intervention method. We evaluated our approach across a suite of open-source MLLMs, including LLaVA-1.5-7B, LLaVA-1.5-13B,

LLaVA-NeXT-7B, Qwen3-VL-4B, Qwen3-VL-8B, MiniGPT-4, and InternVL3-8B. All inference processes were executed locally without the use of commercial API services. For each model, we ran standard greedy decoding, beam search, and our proposed REGAZE method. Concurrently, we reproduced several baseline methods (DoLa, VCD, OPERA, SID, EAZY, and DalD) for comparison. Evaluations were conducted on the primary hallucination benchmarks: POPE, CHAIR, and MME.

Average Inference Latency. On the NVIDIA H800, standard greedy decoding for LLaVA-1.5-7B requires approximately 56.0 ms per sample (averaged across the POPE dataset, generating roughly 16 tokens). REGAZE introduces only a marginal overhead, requiring about 62.2 ms per sample (1.11× compared to greedy decoding). Qwen3-VL-8B exhibits a similar pattern (65.0 ms for greedy vs. 72.8 ms for REGAZE). Under identical settings, other baseline methods incur higher latencies on LLaVA-1.5-7B: approximately 65.0 ms for SID, 80.6 ms for OPERA, and 107.5 ms for VCD.

Total Computational Time. The comprehensive suite of experiments—including main results, ablation studies, hyperparameter analyses, model generalization evaluations, overhead measurements, and case studies—consumed approximately 110 GPU hours in total on a single H800 GPU. The majority of this time was dedicated to reproducing and evaluating the baseline methods; given the lightweight design of REGAZE, its specific operational overhead accounts for only a minimal fraction of the total compute time.

Training, Fine-Tuning, and API Usage. As this study introduces a training-free method, we conducted no model training, fine-tuning, or grid searches involving learnable parameters. All results were generated entirely via local GPU inference, resulting in zero commercial API calls.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our claims made in the abstract and introduction sections accurately reflect our paper's scope and contributions.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations in the Appendix D.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed descriptions of our experimental setup in Section 4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: All datasets and pre-trained models used in this paper are publicly available. We are currently organizing and cleaning our codebase and will make the code publicly available in the future.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide a dedicated section, Section 4.1, to describe our experimental setup in detail.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All experimental results are averaged over multiple runs to ensure stability and reliability.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Computational resources are thoroughly detailed and evaluated across the experimental setup and results sections, with a comprehensive compute report provided in Appendix F.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper fully conforms to the NeurIPS Code of Ethics in every respect.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss both the potential positive and negative societal impacts of our work in Appendix C.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the original paper that produced the models or datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The LLM is only used for writing and not as part of the core methods.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.