Unmasking-Time Visual Calibration for Hallucination Mitigation in Multimodal Discrete Diffusion Language Models

Tian Qin[1,2*], Junzhe Chen[1*], Tianshu Zhang[1], and Lijie Wen[1†]

[1] Tsinghua University, Beijing, China
[2] The University of Sydney, Sydney, NSW, Australia
[*]Equal contribution. [†]Corresponding author.
tqin0515@uni.sydney.edu.au
chenjz24@mails.tsinghua.edu.cn
wenlj@tsinghua.edu.cn

Abstract. Despite the recent success of multimodal discrete diffusion language models (multimodal dLLMs) in generating text through iterative demasking with bidirectional attention, object hallucination remains a critical challenge. Existing hallucination mitigation methods are designed for autoregressive generation and encounter fundamental mismatches when applied to multimodal dLLMs, either doubling inference cost through contrastive decoding or suffering from ambiguous step and position alignment under bidirectional context. To address this, we propose Unmasking-Time Visual Calibration (UVC), a lightweight, training-free framework that mitigates hallucinations natively within the demasking process. UVC extracts multi-granularity contrastive activation shift vectors offline by comparing model responses to clean and visually degraded inputs, identifies the most visually informative attention heads via per-head AUC scoring, and injects the pre-computed calibration signals exclusively at still-masked positions during inference, requiring no additional forward passes. Extensive experiments on POPE and MME with two representative multimodal dLLMs demonstrate that UVC consistently reduces hallucinations by a substantial margin while introducing less than one percent inference overhead. Additional evaluation on open-ended captioning with the CHAIR metric further confirms that UVC substantially lowers hallucination rates without affecting generation length, and generalizes effectively across benchmarks without retraining. Code is available at https://github.com/THU-BPM/UVC.

Keywords: Hallucination Mitigation · Multimodal Discrete Diffusion Language Models · Activation Calibration

1 Introduction

Large vision-language models (LVLMs) bridge visual encoders with large language models via visual instruction tuning [25, 40] and typically adopt autoregressive (AR) left-to-right decoding. Despite strong performance, AR-LVLMs

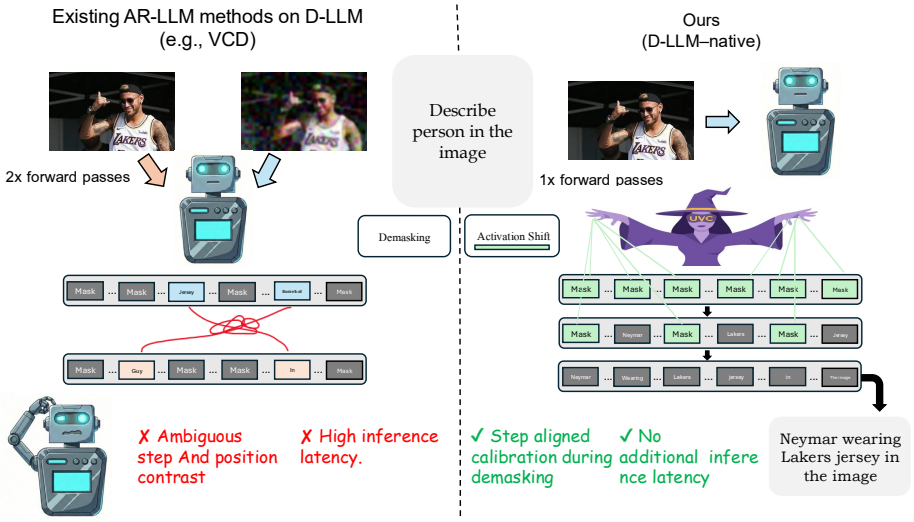

Fig. 1: Comparison of AR-centric contrastive decoding (*e.g.*, VCD [20]) and UVC during multimodal dLLM inference.

suffer from object hallucination, *i.e.* the generation of objects, attributes, or relations absent from the visual input, which remains a persistent challenge [14, 22, 31]. A primary cause is overreliance on language priors: when visual evidence is insufficiently attended, models default to statistical co-occurrence patterns learned from text [19, 34]. The left-to-right causal chain further exacerbates these errors, as one hallucinated token raises the likelihood of subsequent hallucinations [46]. Various mitigation strategies have been proposed, including contrastive decoding [15, 20], training-time alignment [33, 43], post-hoc verification [8, 41], and attention-level calibration [1].

Recently, a distinct generation paradigm has emerged: discrete diffusion language models (dLLMs), which replace sequential decoding with iterative demasking under bidirectional attention [29]. Compared with AR models, dLLMs offer global context awareness from the first generation step, flexible length control with parallel token prediction, and inherent avoidance of sequential error propagation. Advances in masked diffusion modeling [2, 7, 12] and multimodal extensions [38, 39] have demonstrated competitive visual understanding and generation capabilities, attracting growing research interest.

Despite these advantages, multimodal dLLMs are not immune to hallucination. They share the fundamental cause of overreliance on language priors with AR-LVLMs, yet the shift to bidirectional generation introduces structurally distinct failure modes. Bidirectional attention allows the mask configuration, particularly its length, to influence hidden states across the entire sequence from the first demasking step, creating a global vulnerability to drift from language priors. Unlike AR models where hallucination is locally triggered and sequentially amplified, dLLMs exhibit a diffuse pattern in which signals that induce hallu-

cination pervade the representation space simultaneously, making the problem harder to diagnose and to address with local corrections.

These structural differences render existing AR-centric methods ill-suited for multimodal dLLMs. Contrastive decoding methods such as VCD [20] double inference latency via a parallel degraded forward pass, and the confidence-based unmasking schedule yields drastically different demasking trajectories, making contrastive alignment at each step unreliable. Training-time corrections [33, 43] assume sequential token dependencies absent in bidirectional demasking, and penalty-based methods like OPERA [15] rely on prefix-conditioned attention patterns that have no counterpart in simultaneous prediction. These mismatches call for a mitigation approach that operates natively within the dLLM generation mechanism.

We address this challenge from the perspective of internal activations. Prior work on language model interpretability shows that internal activations encode factual associations that can be precisely located and edited [13, 27], suggesting that hallucination signals in multimodal dLLMs can be corrected through targeted activation calibration without retraining or additional forward passes. The key insight is that such calibration must be step-aligned and mask-aware: it applies exclusively to still-masked positions at each demasking step, respecting the generation logic of dLLMs.

Building on this insight, we propose Unmasking-Time Visual Calibration (UVC), a training-free, zero-overhead framework for hallucination mitigation in multimodal dLLMs (Fig. 1). UVC follows a three-stage offline-to-online pipeline: (1) multi-granularity contrastive image pairs extract activation shift vectors encoding the direction from visually degraded to faithful representations (Contrastive Activation Shift Extraction); (2) per-head logistic regressors identify the most visually informative attention heads (Head Selection) [6, 28, 35]; (3) during inference the pre-computed shift vectors are injected into the selected heads at every demasking step, exclusively at still-masked positions (Mask-aware Intervention), requiring no additional forward passes.

We evaluate UVC on two established hallucination benchmarks, POPE [22] and the perception subset of MME [9], using two representative multimodal dLLMs: Lumina-DiMOO [38] and MMaDA [39], both built upon LLaDA [29]. UVC consistently reduces hallucinations across both models and benchmarks, with average F1 improvements of 7.0 and 7.4 points on POPE, while introducing less than 1% inference overhead.

Our contributions are as follows:

1. We present the first hallucination mitigation framework designed natively for multimodal dLLMs, grounded in a systematic analysis of the structural incompatibilities between AR-centric methods and iterative demasking.
2. We propose UVC, a training-free framework that extracts multi-granularity contrastive shift vectors offline, selects visually informative heads via per-head AUC scoring, and injects calibration exclusively at still-masked positions during demasking, requiring no additional forward passes.

3. Experiments on POPE [22] and MME [9] with both models show consistent hallucination reduction (average F1 gains of 7.0 and 7.4 on POPE) with $\leq$1% overhead; CHAIR evaluation further confirms generalization to open-ended captioning.

2 Related Work

2.1 Large Vision-Language Models and Multimodal Discrete Diffusion

Large vision-language models (LVLMs) extend LLMs with visual perception by coupling a pre-trained visual encoder with an LLM through a lightweight projection module [25, 40]. Visual instruction tuning on curated image-text pairs has become the prevailing paradigm for multimodal reasoning, yet recent analyses reveal persistent visual shortcomings [5, 34]: models frequently fail to faithfully attend to visual evidence, defaulting instead to language priors. An alternative paradigm has emerged in discrete diffusion language models (dLLMs) [2, 4, 7, 12, 47], which replace autoregressive decoding with iterative demasking under bidirectional attention. LLaDA [29] introduces a pure masked diffusion architecture competitive with AR baselines, while Lumina-DiMOO [38] and MMaDA [39] extend this paradigm to multimodal inputs. The demasking process differs structurally from AR generation: at each step s, the model predicts all positions in the remaining masked set Ω_s simultaneously with bidirectional context; the mask configuration acts as an implicit conditioning signal; and tokens are "pre-committed" in the representation space before being revealed. These properties fundamentally alter the design space for hallucination mitigation, as AR-centric contrastive decoding, per-token intervention, and prefix-conditioned schedules all require rethinking under evolving mask dynamics.

2.2 Mitigating Hallucinations in LVLMs

A substantial body of work has sought to understand and characterize object hallucination in LVLMs [10, 14, 19, 22, 31, 34, 46]. Existing mitigation strategies can be broadly divided into two categories. **Training-time** methods introduce curated datasets or novel training objectives to reduce hallucination propensity [11,17,33,42,43,45]; while effective, they require extensive training infrastructure and are bounded by the fine-tuning distribution. **Inference-time** methods operate on frozen models and can be organized along four axes: (i) contrastive and modified decoding [15, 18, 20, 21, 24, 30, 36, 44, 48] that contrast clean and degraded output distributions or apply penalty-based retrospection; (ii) detection and revision pipelines [8, 41] that verify and correct hallucinated content after generation; (iii) attention and prompt manipulation [1, 26, 37] that restore visual salience without modifying parameters; and (iv) intervention at the activation level [3, 13, 27], which steers factual knowledge localized in specific model components through targeted edits. Notably, all methods above target the autoregressive paradigm; none addresses the iterative demasking of multimodal

dLLMs, where all masked positions are predicted simultaneously under bidirectional context. Furthermore, the activation space of attention heads during inference remains largely underexplored even within the AR setting. Directly porting AR-centric strategies faces fundamental mismatches: contrastive decoding incurs 2× latency per demasking step, per-token schedules assume a monotonically growing prefix absent from bidirectional generation, and conditioning on mask length introduces a global drift channel with no AR counterpart. This gap motivates our work: we present the first hallucination mitigation framework designed natively for the demasking dynamics of multimodal dLLMs, operating through mask-aware, step-aligned activation interventions that require no additional forward passes.

3 Task Formulation

Consider a multimodal dLLM that receives a visual input $\mathbf{V} = \{v_1, v_2, \ldots, v_{N_v}\}$ and a textual input $\mathbf{X} = \{x_1, x_2, \ldots, x_{N_x}\}$, where N_v and N_x denote the number of visual and textual tokens, respectively. The response region is initialized as N_m mask tokens $\mathbf{M} = \{[\texttt{MASK}]_1, \ldots, [\texttt{MASK}]_{N_m}\}$. After encoding through the visual encoder and text embedding layer respectively, the three components are concatenated into $\mathbf{H}^{(0)} = \text{concat}(\mathbf{V}, \mathbf{X}, \mathbf{M}) \in \mathbb{R}^{N \times D}$, where $N = N_v + N_x + N_m$ is the total sequence length and D the model hidden dimension, and processed by L transformer layers with H attention heads each. For clarity, we focus on the multi-head attention (MHA) output projection and omit LayerNorm and feed-forward sub-layers. At layer l, the hidden states are updated as:

$$\mathbf{H}^{(l+1)} = \mathbf{H}^{(l)} + \sum_{h=1}^{H} \text{Attn}_h^{(l)}\big(\mathbf{H}^{(l)}\big)\, \mathbf{W}_{o,h}^{(l)}, \tag{1}$$

where $\text{Attn}_h^{(l)}(\cdot) \in \mathbb{R}^{N \times d}$ is the output of the h-th attention head, $\mathbf{W}_{o,h}^{(l)} \in \mathbb{R}^{d \times D}$ the corresponding output projection, and $d = D/H$ the per-head dimension.

Unlike autoregressive LVLMs, which generate tokens left-to-right via $p(y_t \mid y_{<t})$, multimodal dLLMs use bidirectional attention and iterative demasking. At each step $s \in \{1, \ldots, S\}$, the model predicts all remaining masked positions in parallel:

$$p_\theta(y_t \mid \mathbf{H}_t^{(L)}) = \text{Softmax}\big(f_c(\mathbf{H}_t^{(L)})\big), \quad \forall\, t \in \Omega_s, \tag{2}$$

where $\Omega_s \subseteq \{1, \ldots, N\}$ is the set of positions still masked at step s. Here $f_c : \mathbb{R}^D \to \mathbb{R}^{|\mathcal{V}|}$ denotes the token classification head that maps the final hidden state to logits over the vocabulary $\mathcal{V}$. A confidence-based subset is then unmasked; this repeats for S steps until all N_m positions are resolved.

Bidirectional attention gives each position a global receptive field, but it also allows structural properties of the mask configuration, particularly its length, to influence hidden states across the entire sequence from the first step, creating a vulnerability to drift from language priors that our method addresses.

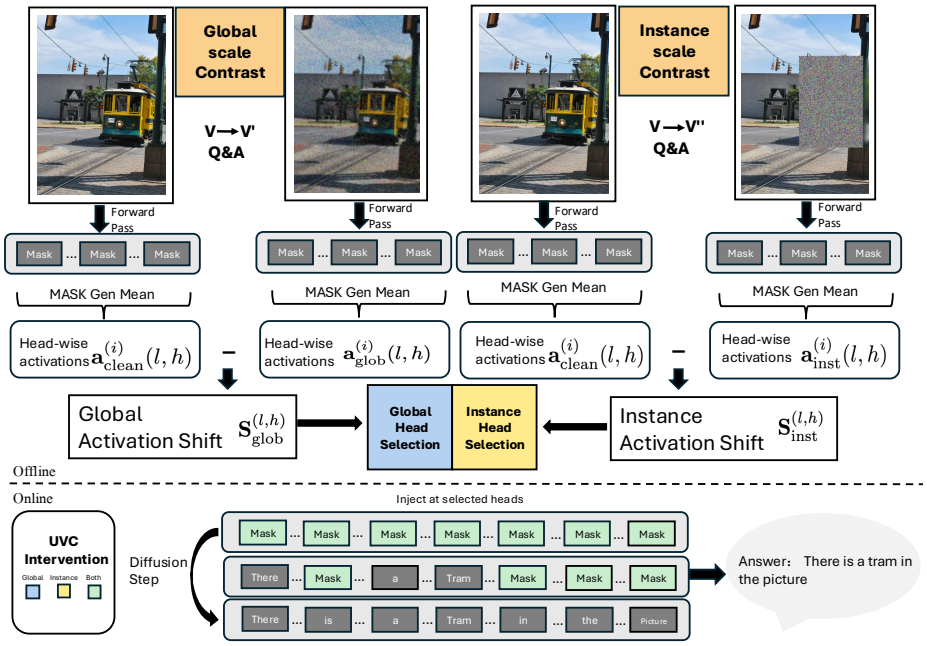

Fig. 2: Overview of UVC. Top: Multi-granularity contrastive activation shift extraction (offline). Middle: Per-head AUC-based head selection (offline). Bottom: Mask-aware shift injection during iterative demasking (online).

4 Method

In this section, we present UVC (Unmasking-Time Visual Calibration). UVC is a training-free framework that mitigates hallucinations in multimodal dLLMs. As illustrated in Fig. 2, UVC follows a three-stage pipeline: shift extraction, head selection, and intervention at masked positions. In the first stage, we construct multi-granularity contrastive pairs and extract activation shift vectors that encode the direction from visually degraded to faithful representations (Sec. 4.1). In the second stage, we fit lightweight logistic regressors to identify, among all $L \times H$ attention heads, those most reliably encoding visual grounding information at the global and instance levels (Sec. 4.2). These two stages are performed offline. In the third stage, during inference we inject the pre-computed shift vectors into the selected heads at every demasking step, exclusively modifying still-masked positions (Sec. 4.3), introducing no additional forward passes.

4.1 Contrastive Activation Shift Extraction

To calibrate the model's internal representations, we require a signal that captures how they change when visual information is degraded. As shown in Fig. 2 (left), we construct contrastive pairs at two complementary granularities, global

and instance, and measure the resulting activation difference across all $L \times H$ attention heads to obtain activation shift vectors.

Global-scale contrast. Consider a batch of n images $\{(q_i, V_i)\}_{i=1}^{n}$ sampled from COCO 2014 train, where each q_i asks about an object present in V_i according to the ground-truth annotations. For each image V_i, we progressively add Gaussian noise following the forward diffusion process to obtain the globally degraded image V_i':

$$p_{\text{fwd}}(V_i^{(t)} \mid V_i^{(t-1)}) = \mathcal{N}\big(V_i^{(t)};\ \sqrt{1-\beta_t}\, V_i^{(t-1)},\ \beta_t \mathbf{I}\big), \tag{3}$$

$$V_i' \sim p_{\text{fwd}}(V_i' \mid V_i) = \prod_{t=1}^{T} p_{\text{fwd}}(V_i^{(t)} \mid V_i^{(t-1)}), \tag{4}$$

where β_t is the noise variance at step t, T controls degradation strength (fixed across samples; see Sec. 5.1), and $\mathbf{I}$ denotes identity covariance. Full-image degradation disrupts global spatial structure (object arrangement, scene layout, and inter-object spatial relations) while preserving partial local texture, creating a contrastive pair that isolates the global-scale visual signal. We construct faithful and degraded pairs for each sample: (q_i, V_i) and (q_i, V_i').

Both pairs are processed by the model. For each sample i, attention head (l, h), and condition $c \in \{\text{clean}, \text{glob}, \text{inst}\}$, we compute a mask-averaged activation vector over the `[MASK]` set Ω_i:

$$\mathbf{a}_c^{(i)}(l,h) = \frac{1}{|\Omega_i|} \sum_{t \in \Omega_i} \mathbf{h}_{c,t}^{(i)}(l,h), \tag{5}$$

where $\mathbf{h}_{c,t}^{(i)}(l,h) \in \mathbb{R}^d$ is the attention output of head h at layer l for position t under condition c, *i.e.* the value-weighted sum after scaled dot-product attention and before the output projection $\mathbf{W}_{o,h}^{(l)}$. This vector is extracted via a forward pre-hook registered on $\mathbf{W}_{o,h}^{(l)}$, capturing the per-head representation before it is projected into the residual stream. Averaging over mask positions aggregates the model's "pre-commitment" across all unresolved tokens, yielding a compact summary of each head's response to the visual input. Let $\mathbf{a}_{\text{clean}}^{(i)}(l,h)$ and $\mathbf{a}_{\text{glob}}^{(i)}(l,h)$ denote the mask-averaged activations under the clean and globally degraded conditions, respectively. The global-scale shift vector for each head is then the mean difference over the n pairs:

$$\mathbf{S}_{\text{glob}}^{(l,h)} = \frac{1}{n} \sum_{i=1}^{n} \big(\mathbf{a}_{\text{clean}}^{(i)}(l,h) - \mathbf{a}_{\text{glob}}^{(i)}(l,h)\big). \tag{6}$$

Instance-scale contrast. Global-scale vectors capture coarse spatial information but are insufficient for fine-grained attribute errors (*e.g.* wrong color or shape). As shown in Fig. 2 (left, instance branch), we construct a dataset $\{(q_i, V_i, V_i'', O_i)\}_{i=1}^{n}$, where O_i is the queried object. Using the ground-truth bounding box of O_i from COCO annotations, we selectively apply Gaussian

noise to this region via Eqs. (3) and (4), producing a locally degraded image V_i'' that corrupts object attributes while preserving surrounding context. From the faithful pair $(q_i + O_i, V_i)$ and the degraded pair $(q_i + O_i, V_i'')$, we compute the instance-scale shift vector $\{\mathbf{S}_{\text{inst}}^{(l,h)}\}_{l=1,h=1}^{L,H}$:

$$\mathbf{S}_{\text{inst}}^{(l,h)} = \frac{1}{n}\sum_{i=1}^{n}\big(\mathbf{a}_{\text{clean}}^{(i)}(l,h) - \mathbf{a}_{\text{inst}}^{(i)}(l,h)\big), \tag{7}$$

where $\mathbf{a}_{\text{inst}}^{(i)}(l,h) \in \mathbb{R}^d$ denotes the mask-averaged activation under the locally degraded condition.

The resulting two sets of shift vectors, $\{\mathbf{S}_{\text{glob}}^{(l,h)}\}$ and $\{\mathbf{S}_{\text{inst}}^{(l,h)}\}$, cover all $L \times H$ heads: global vectors recover spatial layout grounding, while instance vectors restore fine-grained attribute fidelity. Since not every head carries useful visual information, indiscriminate application would introduce noise, motivating the targeted head selection described next.

4.2 Head Selection

Not all $L \times H$ heads encode visual grounding information; many are dominated by syntactic or positional patterns irrelevant to perception. Since different heads in multi-head attention specialize in distinct types of information [6, 28, 35], targeted selection is essential. As shown in Fig. 2 (middle), we quantify per-head linear separability between clean and degraded conditions by fitting a logistic regressor on the head activations and reporting cross-validated AUC as the selection criterion.

For each head (l,h), we train a logistic regressor $f^{(l,h)}(\cdot)$ that receives the mask-averaged activation vector $\mathbf{a}_c^{(i)}(l,h) \in \mathbb{R}^d$ as input and predicts whether it originates from the clean or degraded condition. If a head's activations shift consistently and separably between the two conditions, that head encodes visual information disrupted by degradation, precisely the signal we wish to restore. Conversely, if a head's activations are indistinguishable across conditions, it carries no useful visual signal and should be excluded from intervention. We measure discriminative power using 2-fold cross-validated AUC, denoted $\text{AUC}(f^{(l,h)})$, and rank all heads accordingly. The top-K most discriminative heads are selected for each scale independently:

$$\mathcal{S}_{\text{glob}} = \big\{(l,h) \,\big|\, (l,h) \in \text{TopK}\big(\text{AUC}(f_{\text{glob}}^{(l,h)})\big)\big\}, \tag{8}$$

$$\mathcal{S}_{\text{inst}} = \big\{(l,h) \,\big|\, (l,h) \in \text{TopK}\big(\text{AUC}(f_{\text{inst}}^{(l,h)})\big)\big\}, \tag{9}$$

where K is a hyperparameter controlling the number of intervention heads, and $f_{\text{glob}}^{(l,h)}$, $f_{\text{inst}}^{(l,h)}$ denote the logistic regressors trained on global-scale and instance-scale contrastive pairs, respectively. In practice, we observe that $\mathcal{S}_{\text{glob}}$ and $\mathcal{S}_{\text{inst}}$ are mostly disjoint: global-scale heads tend to reside in middle layers where spatial layout is consolidated, while instance-scale heads concentrate in later

layers where fine-grained object features are refined. This functional specialization confirms that multi-granularity selection captures complementary aspects of visual perception, and reduces the total intervention scope from $L \times H$ to at most $2K$ precisely targeted heads. The shift vectors and selected head sets are subsequently combined in the online inference stage.

4.3 Mask-aware Intervention

Given the pre-computed shift vectors from Sec. 4.1 and the selected head sets from Sec. 4.2, UVC injects the calibration signal into the selected heads at inference time during the iterative demasking process. Crucially, the injection is restricted to still-masked positions only: tokens that have already been unmasked carry committed semantics and should not be perturbed, whereas masked positions have not yet been resolved and are therefore the natural targets for visual calibration. This mask-aware design respects the generation logic of multimodal dLLMs, preventing over-intervention on resolved tokens while concentrating the correction where it is most needed. To formalize this, we define a position-aware indicator $\mathbb{I}^{(l,s)}_{\text{glob},h,t}$ that equals 1 iff both $(l,h) \in \mathcal{S}_{\text{glob}}$ and $t \in \Omega_s$, and 0 otherwise; the instance-scale indicator $\mathbb{I}^{(l,s)}_{\text{inst},h,t}$ is defined analogously with $\mathcal{S}_{\text{inst}}$.

The modified forward pass at layer l and demasking step s is:

$$\mathbf{H}^{(l+1)}_t = \mathbf{H}^{(l)}_t + \sum_{h=1}^{H} \Big(\mathbf{o}^{(l)}_{h,t} + \mathbb{I}^{(l,s)}_{\text{glob},h,t}\, \lambda_{\text{glob}}\, \mathbf{S}^{(l,h)}_{\text{glob}} + \mathbb{I}^{(l,s)}_{\text{inst},h,t}\, \lambda_{\text{inst}}\, \mathbf{S}^{(l,h)}_{\text{inst}} \Big) \mathbf{W}^{(l)}_{o,h}, \quad (10)$$

where $\mathbf{o}^{(l)}_{h,t} = \text{Attn}^{(l)}_h(\mathbf{H}^{(l)})_t \in \mathbb{R}^d$ is the attention output of head h at position t, $\mathbf{W}^{(l)}_{o,h} \in \mathbb{R}^{d \times D}$ is the per-head output projection slice as defined in Eq. (1), and λ_{glob}, λ_{inst} control the global and instance intervention strengths, respectively. The shift vectors are added before the output projection, so the calibration signal is projected into the residual stream jointly with attention; for $t \notin \Omega_s$, both indicators are zero, and overlapping heads receive additive shifts. Setting λ_{inst}=0 or λ_{glob}=0 recovers the global-only (UVC-Global) or instance-only (UVC-Instance) variants as special cases. Intuitively, the intervention steers the selected heads' representations toward stronger visual grounding, amplifying the model's attention to both global scene context and fine-grained object attributes at still-masked positions. As tokens are progressively unmasked, the set Ω_s shrinks, so the calibration naturally concentrates on the remaining uncertain positions where drift from language priors is most pronounced, while already-committed tokens are left intact. Since the shift vectors and head sets are extracted once offline from contrastive pairs of other images, inference requires only the injection of these pre-computed vectors and no ground-truth annotations for the target input.

5 Experiments

5.1 Experimental Setup

Datasets and Metrics. We evaluate UVC on two hallucination benchmarks. **POPE** [22] tests object presence judgment through balanced Yes/No questions, drawing images from MS COCO [23], A-OKVQA [32], and GQA [16] with three negative sampling strategies (Random, Popular, Adversarial), yielding 27,000 pairs. We report Accuracy and F1. **MME** [9] evaluates perception across four categories: Existence and Count (object-level) and Position and Color (attribute-level), each scored on a 200-point scale.
Models. We apply UVC to two multimodal dLLMs built upon LLaDA [29]: Lumina-DiMOO [38] and MMaDA [39], both generating text through iterative demasking.
Implementation Details. We sample 500 images from COCO 2014 train and use ground-truth annotations to construct contrastive pairs: the entire image is degraded for global-scale pairs, while only the bounding-box region is degraded for instance-scale pairs. Per-head logistic regressors are fit with 2-fold cross-validated AUC; the top-K heads are selected, and shift vectors are unit-normalized and scaled by the projection standard deviation. We set $\lambda_{\text{glob}} = \lambda_{\text{inst}}$ and determine the optimal shared intervention strength together with K via grid search. All experiments run on 3× A100 GPUs.

5.2 Results on POPE

Table 1 reports results on nine POPE subsets spanning three data sources and three split types. **1)** UVC-Both improves the average F1 by 7.0 and 7.4 points for Lumina and MMaDA, with Accuracy gains of 3.4 and 12.6 points, respectively. The larger Accuracy gain on MMaDA correlates with its lower Vanilla Accuracy (61.1% vs. 73.9%), where weaker visual grounding yields more directionally consistent shifts that better align with the hallucination direction. **2)** UVC-Global and UVC-Instance individually improve average F1 by 4.8/3.4 on Lumina and 7.4/6.4 on MMaDA. On MMaDA, UVC-Instance leads Accuracy in all nine settings (+16.4), while UVC-Global leads F1 in five of nine; UVC-Both integrates both signals and achieves the best F1 across all nine settings on Lumina. **3)** UVC generalizes across different data distributions: under UVC-Both, F1 improvements on COCO, GQA, and A-OKVQA reach 8.4, 6.8, and 5.9 points for Lumina and 7.0, 7.9, and 7.3 points for MMaDA. MMaDA's Accuracy gains remain stable across splits (12.7, 12.6, 12.5 for Random, Popular, Adversarial), confirming generalization rather than distribution fitting.

Table 1: POPE results across MS COCO, A-OKVQA, and GQA. ΔA/ΔF: absolute Accuracy/F1 gains over Vanilla. **Bold**: best variant per row.

MMaDA

	Vanilla		UVC-Global				UVC-Instance				UVC-Both			
Dataset	Acc	F1	Acc	ΔA	F1	ΔF	Acc	ΔA	F1	ΔF	Acc	ΔA	F1	ΔF
COCO Random	66.57	74.81	76.17	+9.60↑	**83.24**	+8.43↑	**81.37**	+14.80↑	81.22	+6.41↑	79.83	+13.26↑	82.16	+7.35↑
COCO Popular	61.73	72.10	75.73	+14.00↑	**81.06**	+8.96↑	**79.83**	+18.10↑	79.85	+7.75↑	76.40	+14.67↑	80.21	+8.11↑
COCO Adversarial	58.00	70.24	67.97	+9.97↑	**76.54**	+6.30↑	**74.00**	+16.00↑	75.40	+5.16↑	71.57	+13.57↑	75.79	+5.55↑
GQA Random	64.53	73.75	76.30	+11.77↑	**83.64**	+9.89↑	**80.40**	+15.87↑	80.73	+6.98↑	75.17	+10.64↑	82.81	+9.06↑
GQA Popular	61.53	72.04	72.53	+11.00↑	79.14	+7.10↑	**79.00**	+17.47↑	79.51	+7.47↑	70.30	+8.77↑	**80.77**	+8.73↑
GQA Adversarial	58.13	70.29	66.63	+8.50↑	75.80	+5.51↑	**73.33**	+15.20↑	75.36	+5.07↑	68.50	+10.37↑	**76.21**	+5.92↑
A-OKVQA Random	63.23	73.06	75.30	+12.07↑	**82.44**	+9.38↑	**79.17**	+15.94↑	79.63	+6.57↑	77.33	+14.10↑	81.03	+7.97↑
A-OKVQA Popular	60.03	71.29	73.70	+13.67↑	78.88	+7.59↑	**78.27**	+18.24↑	78.79	+7.50↑	74.27	+14.24↑	**79.94**	+8.65↑
A-OKVQA Adversarial	56.13	69.34	70.03	+13.90↑	73.12	+3.78↑	**71.67**	+15.54↑	74.04	+4.70↑	69.80	+13.67↑	**74.54**	+5.20↑

Lumina-DiMOO

	Vanilla		UVC-Global				UVC-Instance				UVC-Both			
Dataset	Acc	F1	Acc	ΔA	F1	ΔF	Acc	ΔA	F1	ΔF	Acc	ΔA	F1	ΔF
COCO Random	73.70	64.95	76.90	+3.20↑	70.72	+5.77↑	75.90	+2.20↑	68.88	+3.93↑	**78.40**	+4.70↑	**73.74**	+8.79↑
COCO Popular	73.47	64.75	76.57	+3.10↑	70.55	+5.80↑	75.67	+2.20↑	68.67	+3.92↑	**78.00**	+4.53↑	**73.39**	+8.64↑
COCO Adversarial	72.37	63.81	75.10	+2.73↑	69.27	+5.46↑	74.27	+1.90↑	67.45	+3.64↑	**75.50**	+3.13↑	**71.48**	+7.67↑
GQA Random	75.20	68.45	77.03	+1.83↑	71.82	+3.37↑	77.20	+2.00↑	72.01	+3.56↑	**79.00**	+3.80↑	**75.79**	+7.34↑
GQA Popular	71.20	65.13	73.70	+2.50↑	69.99	+4.86↑	72.73	+1.53↑	68.27	+3.14↑	**73.73**	+2.53↑	**71.41**	+6.28↑
GQA Adversarial	70.27	64.26	72.73	+2.46↑	69.41	+5.15↑	71.90	+1.63↑	67.59	+3.33↑	**73.10**	+2.83↑	**71.17**	+6.91↑
A-OKVQA Random	78.20	73.15	81.13	+2.93↑	77.96	+4.81↑	79.90	+1.70↑	76.02	+2.87↑	**81.90**	+3.70↑	**79.73**	+6.58↑
A-OKVQA Popular	76.87	71.97	78.87	+2.00↑	75.25	+3.28↑	78.53	+1.66↑	74.80	+2.83↑	**80.03**	+3.16↑	**77.97**	+6.00↑
A-OKVQA Adversarial	73.73	69.41	**76.37**	+2.64↑	74.30	+4.89↑	75.67	+1.94↑	72.86	+3.45↑	75.80	+2.07↑	**74.47**	+5.06↑

5.3 Results on MME

Figure 3 shows MME perception scores across four categories. **1)** UVC-Both achieves the highest Total on both models: 475.00 for Lumina (+50.00) and 546.67 for MMaDA (+53.34), with the largest gains on each model's weakest Vanilla category (Position +22.2% on Lumina; Count +46.5% on MMaDA). **2)** UVC-Global and UVC-Instance show complementary strengths with comparable Totals (461.67 vs. 463.33 on Lumina; 528.33 vs. 516.67 on MMaDA): on Lumina, UVC-Instance peaks on Count while UVC-Global peaks on Position; the pattern reverses on MMaDA. **3)** Since each variant corrects distinct perception errors, UVC-Both aggregates their gains, achieving the best Existence, Color, and overall Total on both models.

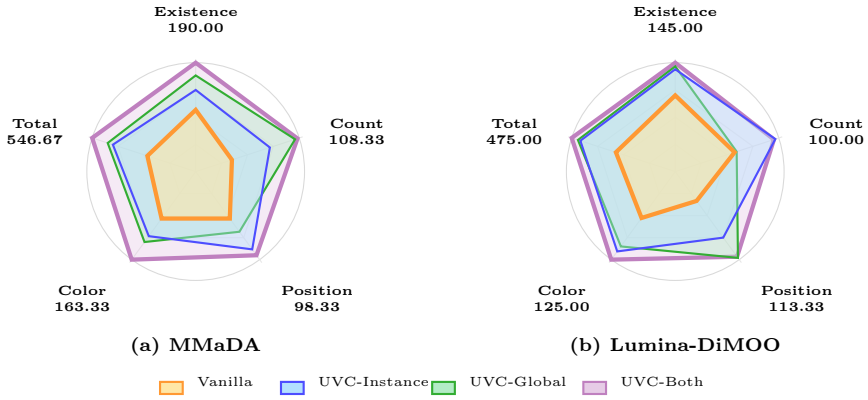

Fig. 3: MME perception scores across four categories and the Total for each UVC variant.

6 Discussions

6.1 Qualitative Case Study

Figure 4 illustrates the open-ended generation behavior behind the CHAIR results: the vanilla model produces a self-contradictory answer, while UVC keeps the response aligned with the visible content.

Fig. 4: Qualitative case study. The vanilla model generates a self-contradictory response, while UVC produces a correct answer.

6.2 Quantitative Analysis

We further assess UVC on open-ended captioning using the CHAIR metric [31], which measures the proportion of hallucinated objects in generated captions:

$$\text{CHAIR}_S = \frac{|\text{hallucinated objects}|}{|\text{all mentioned objects}|}, \quad \text{CHAIR}_I = \frac{|\text{hallucinated captions}|}{|\text{all captions}|}, \quad (11)$$

where CHAIR_S measures hallucination at the instance level and CHAIR_I at the sentence level; lower values indicate fewer hallucinations. We randomly sample 500 images from the COCO 2014 validation set [23], prompt each multimodal dLLM with "Please describe this image in detail," and generate captions with a maximum length of 128 tokens over 64 demasking steps.

Table 2: CHAIR evaluation on 500 COCO 2014 val images (↓: lower is better). Avg Len: mean generated tokens.

	Lumina-DiMOO			MMaDA		
	$C_S \downarrow$	$C_I \downarrow$	Avg Len	$C_S \downarrow$	$C_I \downarrow$	Avg Len
Vanilla	37.40	18.14	126.1	35.20	15.21	62.1
UVC-Both	**21.40**	**12.13**	125.4	**15.40**	**8.78**	63.1

As shown in Table 2, UVC substantially reduces hallucination rates: on Lumina-DiMOO, CHAIR_S drops by 42.8% relative and CHAIR_I by 33.1%; on MMaDA, the reductions are 56.3% and 42.3%. Average caption lengths remain virtually unchanged, confirming that UVC selectively corrects object-level content rather than suppressing generation.

Table 3: Inference latency of Vanilla and UVC-Both across generation lengths.

	Lumina-DiMOO			MMaDA		
(N_m, S)	Van. (ms)	UVC (ms)	Slow.	Van. (ms)	UVC (ms)	Slow.
$(16, 8)$	6552	6560	1.00×	2094	2116	1.01×
$(64, 32)$	26610	26696	1.00×	4247	4295	1.01×
$(128, 64)$	55118	55445	1.01×	9118	9206	1.01×

Furthermore, as shown in Table 3, UVC introduces at most 1.01× slowdown across all generation lengths, confirming negligible computational overhead.

6.3 Head Selection Analysis

Figure 5 reveals that the two scales select largely disjoint head sets: global-scale heads concentrate in middle layers (14 to 20), while instance-scale heads cluster in deeper layers (21 to 26), confirming complementary visual perception. Despite architectural differences, both models exhibit similar layer-wise specialization patterns, and the selected heads are sparsely distributed across layers rather than clustered, corroborating the design choice of selecting across all $L \times H$ heads.

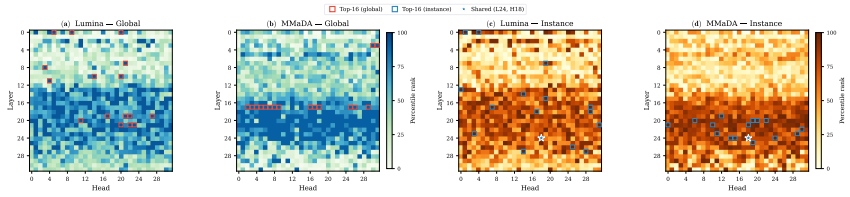

Fig. 5: Per-head AUC heatmaps for global-scale (a, b) and instance-scale (c, d) on Lumina-DiMOO and MMaDA.

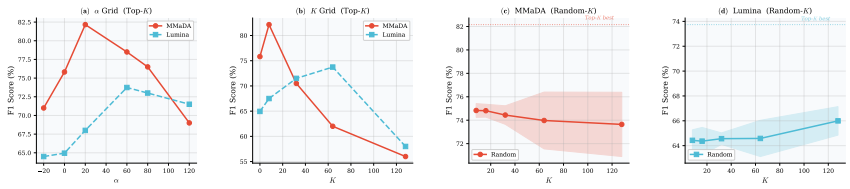

Fig. 6: Hyperparameter sensitivity (λ, K) and top-K vs. random head selection on the POPE COCO Random subset.

Figure 6(a, b) shows that performance is stable across a range of λ and K values, with optimal regions around K=10–20, indicating low sensitivity to hyperparameter choice. As shown in Figure 6(c, d), AUC-guided top-K selection consistently outperforms random selection across all values of K, confirming that the strategy identifies genuinely informative heads rather than benefiting from indiscriminate intervention.

6.4 Demasking Divergence

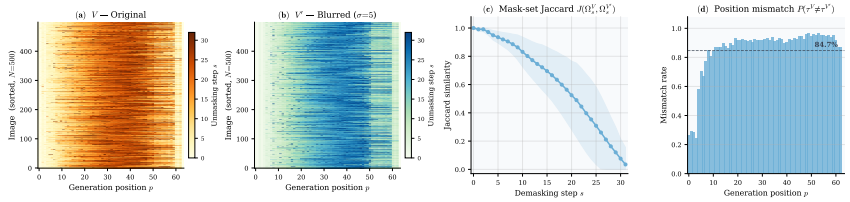

Fig. 7: Demasking order divergence between original and blurred inputs over 500 images.

Figure 7 records per-position unmasking steps for 500 COCO images under both original and Gaussian-blurred inputs. The Jaccard similarity between mask sets Ω_s^V and $\Omega_s^{V'}$ decays rapidly within the first few steps, and 84.7% of positions are unmasked at entirely different steps. This severe divergence exposes a

fundamental obstacle for contrastive decoding methods such as VCD [20], which assume positional alignment at each step, an assumption that collapses under the confidence-based unmasking schedule of dLLMs. UVC avoids this issue entirely by injecting pre-computed shift vectors into still-masked positions at each demasking step, requiring no alignment across trajectories and naturally adapting to the evolving mask configuration.

7 Conclusion and Limitations

In this paper, we tackle the problem of object hallucination in multimodal dLLMs, where existing AR-centric mitigation methods are structurally incompatible with the iterative demasking paradigm. We propose Unmasking-Time Visual Calibration (UVC), a training-free method that performs visual calibration at the activation level natively within the demasking process. UVC extracts multi-granularity contrastive shift vectors offline, selects visually informative attention heads via per-head AUC scoring, and injects calibration signals exclusively at still-masked positions during inference, introducing no additional forward passes. Extensive experiments on POPE and MME with two representative multimodal dLLMs confirm that UVC significantly reduces object hallucinations across discriminative evaluation settings while preserving generation fluency and incurring negligible computational overhead. Additional evaluation with the CHAIR metric on open-ended captioning further validates its effectiveness in generative settings.

Limitations. Shift extraction currently relies on Gaussian-only degradations; exploring semantically guided perturbations is a promising direction. In addition, UVC requires access to intermediate activations, which constrains its applicability to open-weight models.

Acknowledgements

This work is supported by the National Key Research and Development Program of China (No.2024YFB3309702), the National Nature Science Foundation of China (No.62021002), Tsinghua BNRist and Beijing Key Laboratory of Industrial Big Data System and Application, Beijing Natural Science Foundation(QY25044).

References

1. An, W., Tian, F., Leng, S., Nie, J., Lin, H., Wang, Q., Chen, P., Zhang, X., Lu, S.: 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). pp. 29915–29926. IEEE (jun 2025)
2. Arriola, M., Chiu, J., Gokaslan, A., Kuleshov, V., Marroquin, E., Rush, A., Sahoo, S., Schiff, Y.: Simple and effective masked diffusion language models. In: Advances in Neural Information Processing Systems 37. p. 130136–130184. NeurIPS 2024, Neural Information Processing Systems Foundation, Inc. (NeurIPS) (2024)

3. Azaria, A., Mitchell, T.: The internal state of an llm knows when it's lying. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 967–976. Association for Computational Linguistics (2023)
4. Bi, W., Chen, L., Gao, J., Gong, S., Jiang, X., Kong, L., Li, Z., Shi, H., Wu, C., Ye, J., Zheng, L.: Diffusion of thought: Chain-of-thought reasoning in diffusion language models. In: Advances in Neural Information Processing Systems 37. p. 105345–105374. NeurIPS 2024, Neural Information Processing Systems Foundation, Inc. (NeurIPS) (2024)
5. Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., Chang, B.: An Image is Worth 1/2 Tokens After Layer 2: Plug-and-Play Inference Acceleration for Large Vision-Language Models, pp. 19–35. Springer Nature Switzerland (nov 2024)
6. Clark, K., Khandelwal, U., Levy, O., Manning, C.D.: What does bert look at? an analysis of bert's attention. In: Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. pp. 276–286 (2019)
7. Doucet, A., Han, K., Shi, J., Titsias, M., Wang, Z.: Simplified and generalized masked diffusion for discrete data. In: Advances in Neural Information Processing Systems 37. p. 103131–103167. NeurIPS 2024, Neural Information Processing Systems Foundation, Inc. (NeurIPS) (2024)
8. Favero, A., Zancato, L., Trager, M., Choudhary, S., Perera, P., Achille, A., Swaminathan, A., Soatto, S.: Multi-modal hallucination control by visual information grounding. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14303–14312. IEEE (jun 2024)
9. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R., Shan, C., He, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models (2023)
10. Gunjal, A., Yin, J., Bas, E.: Detecting and preventing hallucinations in large vision language models. Proceedings of the AAAI Conference on Artificial Intelligence **38**(16), 18135–18143 (mar 2024)
11. Guo, H., Li, C., Wang, J., Wu, B., Xiao, X., Zhou, X.: Seeing the image: Prioritizing visual correlation by contrastive alignment. In: Advances in Neural Information Processing Systems 37. pp. 30925–30950. NeurIPS 2024, Neural Information Processing Systems Foundation, Inc. (NeurIPS) (2024)
12. He, Z., Sun, T., Tang, Q., Wang, K., Huang, X., Qiu, X.: Diffusionbert: Improving generative masked language models with diffusion models. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). p. 4521–4534. Association for Computational Linguistics (2023)
13. Heimersheim, S., Nanda, N.: How to use and interpret activation patching (2024)
14. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., Liu, T.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems **43**(2), 1–55 (jan 2025)
15. Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., Yu, N.: Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13418–13427. IEEE (jun 2024)
16. Hudson, D.A., Manning, C.D.: 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). pp. 6693–6702. IEEE (jun 2019)

17. Jiang, C., Xu, H., Dong, M., Chen, J., Ye, W., Yan, M., Ye, Q., Zhang, J., Huang, F., Zhang, S.: Hallucination augmented contrastive learning for multimodal large language model. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 27036–27046. IEEE (jun 2024)
18. Kim, H., Kim, J., Kim, Y., Ro, Y.: Code: Contrasting self-generated description to combat hallucination in large multi-modal models. In: Advances in Neural Information Processing Systems 37. pp. 133571–133599. NeurIPS 2024, Neural Information Processing Systems Foundation, Inc. (NeurIPS) (2024)
19. Lee, K.i., Kim, M., Yoon, S., Kim, M., Lee, D., Koh, H., Jung, K.: Vlind-bench: Measuring language priors in large vision-language models. In: Findings of the Association for Computational Linguistics: NAACL 2025. pp. 4129–4144. Association for Computational Linguistics (2025)
20. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13872–13882. IEEE (jun 2024)
21. Li, X.L., Holtzman, A., Fried, D., Liang, P., Eisner, J., Hashimoto, T., Zettlemoyer, L., Lewis, M.: Contrastive decoding: Open-ended text generation as optimization. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12286–12312. Association for Computational Linguistics (2023)
22. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 292–305. Association for Computational Linguistics (2023)
23. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context, pp. 740–755. Springer International Publishing (2014)
24. Liu, B., Zhang, F., Chen, G., Cheng, J.: Multi-frequency contrastive decoding: Alleviating hallucinations for large vision-language models. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 28556–28572. Association for Computational Linguistics (2025)
25. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26286–26296. IEEE (jun 2024)
26. Liu, S., Zheng, K., Chen, W.: Paying More Attention to Image: A Training-Free Method for Alleviating Hallucination in LVLMs, pp. 125–140. Springer Nature Switzerland (nov 2024)
27. Meng, K., Bau, D., Andonian, A., Belinkov, Y.: Locating and editing factual associations in gpt (2022)
28. Michel, P., Levy, O., Neubig, G.: Are sixteen heads really better than one? (2019)
29. Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., Zhou, J., Lin, Y., Wen, J.R., Li, C.: Large language diffusion models (2025)
30. Qin, T., Chen, J., Shi, Y., Zhang, T., Ju, Q., Wen, L.: Do we really need external tools to mitigate hallucinations? sira: Shared-prefix internal reconstruction of attribution. arXiv preprint arXiv:2605.14621 (2026)
31. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 4035–4045. Association for Computational Linguistics (2018)

32. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-OKVQA: A Benchmark for Visual Question Answering Using World Knowledge, pp. 146–162. Springer Nature Switzerland (2022)
33. Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L., Wang, Y.X., Yang, Y., Keutzer, K., Darrell, T.: Aligning large multimodal models with factually augmented rlhf. In: Findings of the Association for Computational Linguistics ACL 2024. pp. 13088–13110. Association for Computational Linguistics (2024)
34. Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., Xie, S.: Eyes wide shut? exploring the visual shortcomings of multimodal llms. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9568–9578. IEEE (jun 2024)
35. Voita, E., Talbot, D., Moiseev, F., Sennrich, R., Titov, I.: Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 5797–5808. Association for Computational Linguistics (2019)
36. Wang, X., Pan, J., Ding, L., Biemann, C.: Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In: Findings of the Association for Computational Linguistics ACL 2024. pp. 15840–15853. Association for Computational Linguistics (2024)
37. Woo, S., Kim, D., Jang, J., Choi, Y., Kim, C.: Don't miss the forest for the trees: Attentional vision calibration for large vision language models. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 1927–1951. Association for Computational Linguistics (2025)
38. Xin, Y., Qin, Q., Luo, S., Zhu, K., Yan, J., Tai, Y., Lei, J., Cao, Y., Wang, K., Wang, Y., Bai, J., Yu, Q., Jiang, D., Pu, Y., Chen, H., Zhuo, L., He, J., Luo, G., Li, T., Hu, M., Ye, J., Ye, S., Zhang, B., Xu, C., Wang, W., Li, H., Zhai, G., Xue, T., Fu, B., Liu, X., Qiao, Y., Liu, Y.: Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding (2025)
39. Yang, L., Tian, Y., Li, B., Zhang, X., Shen, K., Tong, Y., Wang, M.: Mmada: Multimodal large diffusion language models (2025)
40. Ye, Q., Xu, H., Ye, J., Yan, M., Hu, A., Liu, H., Qian, Q., Zhang, J., Huang, F.: mplug-owi2: Revolutionizing multi-modal large language model with modality collaboration. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13040–13051. IEEE (jun 2024)
41. Yin, S., Fu, C., Zhao, S., Xu, T., Wang, H., Sui, D., Shen, Y., Li, K., Sun, X., Chen, E.: Woodpecker: hallucination correction for multimodal large language models. Science China Information Sciences **67**(12) (dec 2024)
42. Yu, Q., Li, J., Wei, L., Pang, L., Ye, W., Qin, B., Tang, S., Tian, Q., Zhuang, Y.: Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12944–12953. IEEE (jun 2024)
43. Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.T., Sun, M.: Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13807–13816. IEEE (jun 2024)
44. Yue, Z., Zhang, L., Jin, Q.: Less is more: Mitigating multimodal hallucination from an eos decision perspective. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 11766–11781. Association for Computational Linguistics (2024)

45. Zhang, J., Wang, T., Zhang, H., Lu, P., Zheng, F.: Reflective Instruction Tuning: Mitigating Hallucinations in Large Vision-Language Models, pp. 196–213. Springer Nature Switzerland (nov 2024)
46. Zhong, W., Feng, X., Zhao, L., Li, Q., Huang, L., Gu, Y., Ma, W., Xu, Y., Qin, B.: Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 11991–12011. Association for Computational Linguistics (2024)
47. Zhou, K., Li, Y., Zhao, X., Wen, J.R.: Diffusion-nat: Self-prompting discrete diffusion for non-autoregressive text generation. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). p. 1438–1451. Association for Computational Linguistics (2024)
48. Zhu, L., Ji, D., Chen, T., Xu, P., Ye, J., Liu, J.: Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. In: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1615–1624. IEEE (jun 2025)