Seeing to Ground: Visual Attention for Hallucination-Resilient MDLLMs

Vishal Narnaware*, Animesh Gupta*, Kevin Zhai, Zhenyi Wang, and Mubarak Shah

Institute of Artificial Intelligence, University of Central Florida, USA
{firstname.lastname}@ucf.edu

* Equal contribution.

Abstract. Multimodal Diffusion Large Language Models (MDLLMs) achieve high-concurrency generation through parallel masked decoding, yet the architectures remain prone to multimodal hallucinations. This structural vulnerability stems from an algorithmic flaw: the decoder ranks candidate tokens based on textual likelihood without verifying localized visual support. We establish that this language-only ranking induces an objective mismatch, where language probability mass acts as a misspecified proxy for the intended multimodal task. Consequently, we reinterpret hallucination as a localized optimization error, a phenomenon where the decoder exploits language shortcuts to maximize a proxy score at the expense of visual grounding. To address this objective mismatch, we introduce VISAGE, a training-free decoding framework that calibrates the objective at inference time. VISAGE estimates the proxy discrepancy by quantifying the spatial entropy of cross-attention distributions. By enforcing a localization consensus across attention heads, the method penalizes spatially uniform distributions and re-ranks token commitments to favor visually grounded outcomes. We provide an analytical stability guarantee establishing that VISAGE maintains a bounded objective loss under estimation error. Evaluations across hallucination-sensitive and general-purpose benchmarks demonstrate the robustness of the framework, yielding relative gains of 8.59% on MMMU-val and 7.75% on HallusionBench.

Keywords: hallucination · masked decoding · multimodal diffusion large language models

1 Introduction

Multimodal Diffusion Large Language Models (MDLLMs) [16, 36, 39, 44] represent a high-concurrency alternative to autoregressive decoding [2, 31, 35] for vision-language generation. By enabling parallel token generation, these architectures reduce inference latency. However, the architectural efficiency does not address the persistent problem of hallucination: the generation of responses that lack grounding in the visual input. Prior approaches often treat these generative

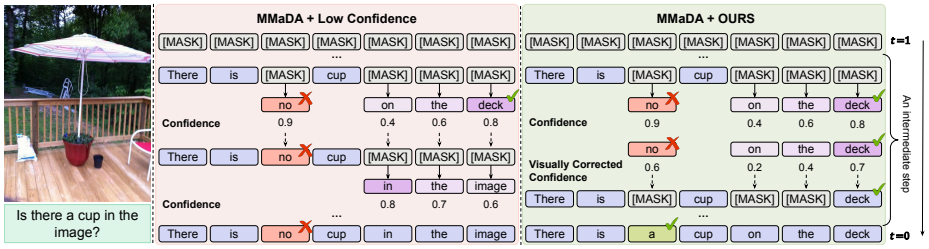
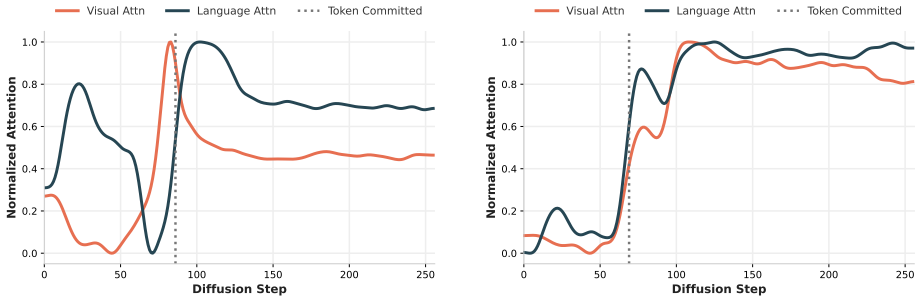


Fig. 1: VISAGE resolves language shortcuts by correcting the objective mismatch during parallel unmasking. Given the query "*Is there a cup in the image?*", the standard decoder (left) assigns a high language-only confidence (0.9) to the statistically plausible token "*no*." Because this proxy score lacks visual verification, the token is finalized prematurely and induces a hallucination. In contrast, VISAGE (right) calculates a visually corrected confidence (0.6) by estimating the proxy discrepancy b_i. This modification to the ranking score prevents the early commitment of the ungrounded token, allowing the model to recover the correct multimodal outcome: "*There is a cup on the deck.*"

failures as a consequence of insufficient model capacity or poor multimodal alignment during pre-training [14]. In this work, we propose a different perspective: in the parallel unmasking setting, hallucination is a localized optimization error induced by the decoding algorithm itself.

This localized optimization error arises because the parallel unmasking mechanism acts as an implicit, position-wise optimization process. By committing tokens based on maximum language likelihood under the denoising distribution, the decoder optimizes a language-only proxy objective rather than the intended multimodal objective. This algorithmic reliance induces an objective mismatch. Consequently, the decoder maximizes the proxy score by exploiting language shortcuts [10], relying on textual priors to finalize predictions without verifying localized visual probability mass. As illustrated in Figure 1 (left), when answering "Is there a cup in the image?", standard masked decoding assigns high probability mass to the statistically plausible token "no". Because the algorithm commits tokens based on the language-only proxy ranking, the ungrounded token is prematurely finalized, propagating as incorrect context for all subsequent parallel decoding steps.

To empirically validate the language shortcut hypothesis, we analyze the normalized peak probability mass for visual and language components across the diffusion steps leading up to commitment. As shown in Figure 2, comparing the probability mass of grounded versus hallucinated token commitments reveals a structural disparity. When the model commits a visually grounded token (left), peak visual probability mass exceeds the language prior leading up to the commitment step. In contrast, during a language shortcut hallucination (right), the visual probability mass remains spatially uniform, resulting in a suppressed peak relative to the textual context throughout the decoding process. This structural disparity demonstrates that standard probability-based ranking optimizes



(a) Grounded Token **(b)** Hallucinated Token

Fig. 2: The disparity between visual and linguistic probability mass across decoding steps suggests objective misspecification. We visualize the normalized peak probability mass for visual (orange) and language (blue) components across the refinement process. (a) In **Visually Grounded Generation**, peak visual probability mass exceeds the language attention prior before the commitment step (vertical dotted line), consistent with a localized, low-entropy distribution over image tokens. (b) In a **Language Shortcut Hallucination**, the visual component retains a uniform spatial distribution, resulting in a suppressed peak that falls below the language attention prior. This structural disparity demonstrates that the decoder optimizes for textual likelihood while bypassing localized visual probability mass, suggesting that the spatial Shannon entropy of cross-attention provides a detectable signal for re-ranking ungrounded commitments.

for textual likelihood while bypassing localized visual evidence, necessitating a re-ranking intervention at decoding time.

To resolve the objective mismatch, we introduce VISAGE (Visual Attention Grounding Entropy), a training-free re-ranking framework. VISAGE intercepts the parallel commitment schedule by deriving a visual re-ranking penalty from the spatial Shannon entropy of cross-attention over image tokens. The spatial entropy computation quantifies the probability mass for localized visual evidence. To ensure robustness, VISAGE aggregates the spatial entropy across attention heads using a discrete β-quantile function. This aggregation operator functions as a localization consensus, ensuring that a token is only deemed grounded if a specific fraction of heads independently exhibit a low-entropy distribution. Tokens exhibiting spatially uniform attention distributions receive a re-ranking penalty and are downweighted, while tokens with a verified consensus of localized visual support retain their original probability mass. As illustrated in Figure 1 (right), substituting the flawed proxy score with the entropy-calibrated probability mass penalizes shortcut-driven tokens and aligns the parallel decoding process toward grounded multimodal outcomes.

To summarize, our main contributions are as follows:

1. **Formalization of the objective mismatch in parallel decoding.** We reinterpret hallucination in MDLLMs as a localized optimization error rather than a generative capacity failure. We establish that probability-based un-

masking operates as a misspecified proxy objective that incentivizes language shortcuts.

2. **A consensus-driven framework for grounded generation.** We introduce VISAGE, a training-free inference framework that calibrates the decoding objective. By estimating the proxy discrepancy via the robust spatial entropy of cross-attention, the framework enforces a localization consensus to penalize visually unsupported commitments.
3. **Analytical stability and empirical hallucination resilience.** We establish a mathematical stability guarantee for the monotonic reweighting, proving that VISAGE bounds the maximum objective loss under estimation error. We validate the framework across diverse multimodal benchmarks, demonstrating consistent improvements in visually grounded generation.

2 Related Work

Multimodal Diffusion Language Models. Multimodal Diffusion Large Language Models (MDLLMs) adapt image diffusion to discrete text spaces by replacing sequential prediction with parallel masked decoding [1,17]. Architectures like LLaDA [21] and Dream [40] improve scalability via mask-prediction objectives, while MMaDA [39] and others [6, 16, 25, 36, 44] extend the MDLLM paradigm to joint vision-language reasoning. These high-concurrency architectures reduce inference latency compared to autoregressive counterparts. However, the discrete proposal space breaks continuous alignment mechanisms, such as classifier guidance [8, 12, 28, 29, 34, 43] and classifier-free guidance [13, 27], because the continuous mechanisms require backpropagating perturbations through a differentiable score function. Within this discrete setting, recent training-free decoding methods for diffusion language models operate at inference without retraining: CORE [46] remasks already-committed tokens to revise unreliable predictions, and DyStruct [30] adapts generation structure and length during decoding. Both, however, operate purely on text and leave visual grounding unaddressed. In contrast, VISAGE intervenes at the commitment step itself: rather than revising tokens after the fact, it re-ranks masked candidates by visual grounding so that ungrounded proposals are never committed, recovering the grounding objective without continuous gradient approximations.

Visual Grounding and Hallucination. Hallucination mitigation relies on training-time alignment [3–5, 19, 22–24, 33] or autoregressive inference-time interventions via external verification or internal calibration. External verification pipelines like Woodpecker [41] and MARINE [48] utilize auxiliary object detectors to validate content post-hoc, introducing multi-pass latency. Internal calibration methods instead utilize the base model's representations to apply contrastive penalties [7,15,37,42], suppress attention heads [26], or constrain information bottlenecks [47] during sequential next-token prediction. While these calibration interventions mitigate linguistic biases, the sequential mechanisms ignore the localized optimization error specific to parallel masked decoding, where

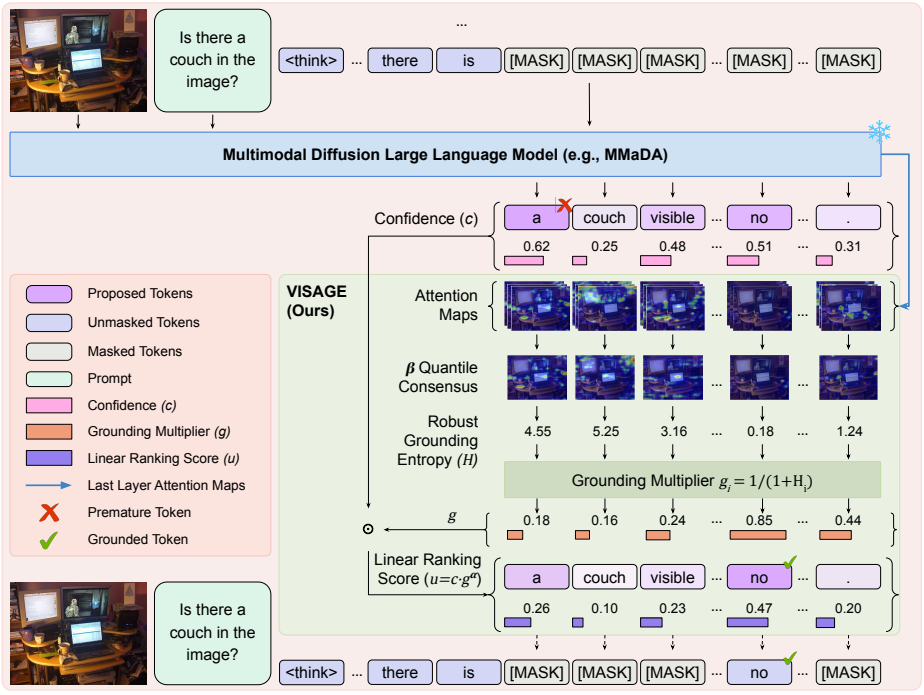


Fig. 3: Overview of VISAGE during parallel masked decoding. At each decoding step, the frozen MDLLM generates candidate tokens alongside their initial confidence c. To verify visual support, we extract each candidate token's last-layer cross-attention weights over the image and compute Shannon entropy for each attention head. We then aggregate these values across heads via a β-quantile operator, yielding Robust Grounding Entropy H, which quantifies the concentration of visual support. A penalty multiplier $g = 1/(1+H)$ is then computed. Tokens are subsequently re-ranked using the linear ranking score $u = c \cdot g^{\alpha}$ ($\alpha = 0.5$). As illustrated, the ungrounded token (*"a"*) is penalized by high entropy, ensuring the visually supported token (*"no"*) attains a high linear ranking score and is successfully committed to the sequence.

the unmasking algorithm selects tokens based on the categorical textual distribution. The reliance on the textual distribution enables ungrounded language shortcuts [10] where language priors override visual conditioning. Unlike autoregressive interventions, VISAGE resolves the objective mismatch by deriving a robust grounding entropy from internal attention distributions. The robust entropy ensures token commitments possess high probability mass across localized visual evidence, providing a training-free defense against decoding-induced hallucinations.

3 Problem Formulation

Let $\mathcal{V}$ be the vocabulary and x denote the text prompt. For a visual input v, the intermediate MDLLM sequence at step $t \in \{1, \ldots, T\}$ is represented by $y^{(t)} = [x; v; y_1^{(t)}, \ldots, y_L^{(t)}]$, where each response position $y_i^{(t)} \in \mathcal{V} \cup \{\texttt{[MASK]}\}$. Decoding initializes all L response positions as $y_i^{(1)} = \texttt{[MASK]}$ and replaces these mask tokens with discrete tokens over T steps. At each decoding step t, the set of masked token positions is defined as:

$$C_t \;=\; \{\, i \in \{1, \ldots, L\} : y_i^{(t)} = \texttt{[MASK]} \,\}. \tag{1}$$

For each masked token position $i \in C_t$, the model processes the sequence $y^{(t)}$ to compute a discrete probability distribution over the vocabulary:

$$p_\theta(y_i = \nu \mid y^{(t)}, i), \qquad \forall \nu \in \mathcal{V}.$$

Let $\hat{y}_i \in \mathcal{V}$ denote the token proposal at position i derived via greedy maximization of the categorical distribution. We define the position-wise confidence c_i as the probability mass assigned to the proposal:

$$c_i \;\triangleq\; p_\theta(y_i = \hat{y}_i \mid y^{(t)}, i). \tag{2}$$

Standard decoding optimizes a proxy objective independent of visual context. Governed by an unmasking schedule, the decoder is allocated a token budget k_t for the current step. The decoder must select a subset of positions $U_t \subseteq C_t$ such that $|U_t| = k_t$. The sequence is then updated by committing the proposed tokens at these selected positions:

$$y_i^{(t+1)} \;=\; \begin{cases} \hat{y}_i, & i \in U_t, \\ y_i^{(t)}, & \text{otherwise.} \end{cases} \tag{3}$$

Confidence-based ranking selects this subset by solving the following maximization problem:

$$U_t \triangleq \arg \max_{\substack{U \subseteq C_t \\ |U| = k_t}} \sum_{i \in U} \log c_i. \tag{4}$$

This maximization objective is equivalent to selecting the top-k_t token positions ranked by the scalar c_i.

The formulation in Equation 4 reveals a structural vulnerability. Because Equation 4 optimizes raw text probability, the equation acts as a misspecified proxy objective that omits visual grounding verification. A high confidence score c_i reflects statistical language priors derived from the initial prompt or previously unmasked tokens, rather than grounding in the visual input v. Consequently, the unmasking objective favors the language prior and enables the model to take language shortcuts, committing tokens that lack visual support. The mismatch between the proxy objective and the intended multimodal objective necessitates a decoding-time correction mechanism capable of verifying localized visual grounding before token commitment.

4 VISAGE: Grounding-Aware Decoding

To address the reliance on language-only proxy objectives that induce hallucinations, we introduce VISAGE, a decoding-time re-ranking framework. Our method formalizes hallucination as a localized optimization error under a misspecified ranking score, and introduces a cross-attention spatial entropy estimator to approximate the unobservable multimodal objective.

4.1 Sequential Commitment under Objective Misspecification

As established in Equation 4, standard decoding ranks candidate tokens by the proxy score $r_{\text{proxy}}(i) \triangleq \log c_i$. This score represents the textual likelihood of a candidate token but lacks a mechanism to quantify spatial concentration over the visual input. The intended multimodal decoding objective $r_\star$ incorporates both textual likelihood and localized visual correspondence:

$$r_\star(i) \;=\; \log c_i \;+\; \lambda\, r_{\text{vis}}(i), \tag{5}$$

where $\lambda > 0$ scales the visual grounding strength. Because a perfect oracle for visual correspondence r_{vis} requires computationally expensive external verification that defeats the efficiency of parallel decoding, standard algorithms default to the language-only proxy r_{proxy}. This reliance induces an objective mismatch where $r_{\text{proxy}}(i) \neq r_\star(i)$. We formalize this objective mismatch as the proxy discrepancy $b_i \geq 0$:

$$r_\star(i) \;=\; \log c_i \;-\; b_i. \tag{6}$$

The proxy discrepancy b_i quantifies the magnitude by which textual probability mass overrides localized visual evidence. This mismatch allows the decoder to maximize the proxy score by committing tokens with high textual likelihood even when the proxy discrepancy b_i is large. Because these committed tokens provide the conditioning context for subsequent steps, such ungrounded commitments propagate as incorrect context throughout the parallel decoding process.

The deviation from the intended multimodal objective $r_\star$ defined in Equation 5 is a position-wise phenomenon captured by the proxy discrepancy b_i:

$$b_i \;=\; r_{\text{proxy}}(i) - r_\star(i). \tag{7}$$

Under the discrepancy formulation in Equation 7, hallucinations correspond to candidate positions that exhibit high textual likelihood $r_{\text{proxy}}(i)$ but suffer from a large position-wise discrepancy b_i. By treating hallucination as a localized optimization error rather than a global sequence failure, we apply targeted re-ranking penalties to indices where visual correspondence is insufficient.

4.2 Grounding-Aware Objective Correction

Spatial entropy quantifies the localization of visual correspondence. Approximating the unobservable multimodal objective $r_\star(\cdot)$ in Equation 6 requires a computable estimate of the proxy discrepancy b_i. We do not assume

that cross-attention weights represent the exact causal influence of visual features; instead, we use the spatial concentration of normalized cross-attention score distributions as a proxy for visual evidence. As observed in the attention trajectories of Figure 2, grounded token commitments coincide with high peak probability mass over specific image regions, consistent with a localized, low-entropy distribution. Conversely, when the model relies on textual priors, the visual probability mass is distributed across the image tokens, resulting in a spatially uniform distribution with a suppressed peak. We formalize this relationship by deriving a computable estimator $\hat{b}_i$ from the spatial entropy of the cross-attention distribution, penalizing tokens that lack concentrated visual probability mass.

Cross-attention entropy yields the grounding-aware ranking score. Let $\mathcal{I}_{\text{img}}$ be indices of image tokens of visual input v and $N := |\mathcal{I}_{\text{img}}|$ denote the number of image tokens. At step t, let $A_{i,j}^{(t,h)}$ denote the last-layer cross-attention from the query representation at masked text position i to image token $j \in \mathcal{I}_{\text{img}}$ in head $h \in \{1, \ldots, M\}$. We renormalize attention over image tokens to obtain a normalized spatial distribution $\tilde{A}_{i,(\cdot)}^{(t,h)}$:

$$\tilde{A}_{i,j}^{(t,h)} \;=\; \frac{A_{i,j}^{(t,h)} + \frac{\delta}{N}}{\sum_{j' \in \mathcal{I}_{\text{img}}} A_{i,j'}^{(t,h)} + \delta}, \qquad j \in \mathcal{I}_{\text{img}}, \tag{8}$$

where $\delta > 0$ is a numerical smoothing constant for stability. We interpret the distribution $\tilde{A}_{i,(\cdot)}^{(t,h)}$ as the spatial localization of visual evidence for the candidate token $\hat{y}_i$. To quantify this concentration of probability mass, we define the Shannon entropy $H_i^{(t,h)}$ for the h-th attention head at token position i:

$$H_i^{(t,h)} \;=\; -\sum_{j \in \mathcal{I}_{\text{img}}} \tilde{A}_{i,j}^{(t,h)} \log\big(\tilde{A}_{i,j}^{(t,h)}\big). \tag{9}$$

In standard vision-language architectures, individual attention heads occasionally collapse to spurious artifacts, yielding artificially low entropy even when the token is ungrounded. To prevent a single spuriously sharp head from bypassing the grounding penalty, we aggregate the M head-wise entropies into a single robust grounding entropy scalar H_i. We define the discrete quantile function $q_\beta(\cdot)$ as an operator that selects the $\lceil \beta M \rceil$-th smallest value from a set. This operator functions as a *localization consensus*: rather than trusting the head with the lowest entropy, we require a specific fraction of heads to agree that a token is grounded. Let $H_{i,(1)} \leq H_{i,(2)} \leq \cdots \leq H_{i,(M)}$ denote the head-wise entropy values $\{H_i^{(t,h)}\}_{h=1}^{M}$ sorted in ascending order. We define the aggregate robust grounding entropy H_i as the specific value at index $\lceil \beta M \rceil$ in the sorted set:

$$H_i \;\triangleq\; q_\beta\big(\{H_i^{(t,h)}\}_{h=1}^{M}\big) \;=\; H_{i,(\lceil \beta M \rceil)}, \qquad \beta \in (0, 1]. \tag{10}$$

This quantile selection ensures that the aggregate entropy H_i takes a numerically small value only if at least βM individual heads also exhibit low entropy. For

instance, setting $\beta = 0.5$ defines H_i as the median head entropy. This median requirement enforces a majority consensus: the resulting discrepancy estimator $\hat{b}_i$ remains small, thereby preserving the token's ranking, only if at least half of the attention heads independently yield concentrated probability mass over the image tokens.

We define $\hat{b}_i$ as the computable estimator for the true proxy discrepancy b_i. We convert the robust grounding entropy H_i into this estimator $\hat{b}_i$, which increases with higher entropy:

$$\hat{b}_i \triangleq \alpha \log(1 + H_i), \qquad \alpha \geq 0, \tag{11}$$

where α acts as a coefficient to scale the penalty relative to the log-probability. This penalty yields the corrected log-score $\hat{r}(i) = \log c_i - \hat{b}_i$. Defining the grounding multiplier $g_i \triangleq \frac{1}{1+H_i}$, we transform the log-space score into the linear ranking score u_i:

$$u_i = c_i \cdot g_i^{\alpha} = c_i \cdot (1 + H_i)^{-\alpha}. \tag{12}$$

Under the k_t budget constraint, the optimal commitment set U_t is the subset that maximizes the sum of corrected scores:

$$U_t = \arg \max_{\substack{U \subseteq C_t \\ |U|=k_t}} \sum_{i \in U} \hat{r}(i). \tag{13}$$

Because the objective is separable across candidates, the exact optimizer is given by the top-k_t positions ranked by u_i:

$$U_t = \mathrm{TopK}\left(\{u_i\}_{i \in C_t}, k_t\right). \tag{14}$$

Bounded estimation error ensures stable and grounded commitments. We provide a detailed description of the VISAGE procedure in the Supplementary Material. In practice, the entire spatial entropy derivation reduces to a highly efficient inference mechanism: VISAGE reweights token proposals monotonically via the multiplier g_i^{α}. For a fixed confidence c_i, the ranking score u_i decreases as the robust grounding entropy H_i increases. This monotonic relationship ensures that the decoder commits a token only when the token is both confident and validated by localized visual correspondence. Analytically, the VISAGE framework provides a stability guarantee against estimation error between the entropy-based estimator $\hat{b}_i$ and the true proxy discrepancy b_i. If the spatial entropy estimator $\hat{b}_i$ approximates the proxy discrepancy b_i such that the absolute error satisfies $|\hat{b}_i - b_i| \leq \varepsilon_t$ for all candidates $i \in C_t$, the resulting objective loss is at most $2k_t\varepsilon_t$ relative to the optimal grounded subset. This $2k_t\varepsilon_t$ bound arises from the potential for a worst-case rank reversal: an overestimated suboptimal token (inflated by $+\varepsilon_t$) can displace an underestimated optimal token (deflated by $-\varepsilon_t$) in the index, resulting in a maximum per-token objective loss of $2\varepsilon_t$. We provide the formal derivation of this stability bound in the Supplementary. By directly addressing the proxy discrepancy, this bounded-error framework ensures the commitment process remains stable and visually grounded.

5 Experiments

Datasets. To evaluate hallucination mitigation, we utilize POPE [18] to probe basic object existence and HallusionBench [11] to assess language-driven hallucinations and visual illusions. Furthermore, to ensure the grounding interventions do not compromise overall generation quality, we benchmark general multimodal capabilities on MMMU-val [45] for college-level, multi-discipline question answering, and MME [9] for comprehensive vision-language tasks.

Baseline Methods. We evaluate VISAGE against two primary decoding strategies. Our base model is MMaDA [39], which employs standard confidence-based unmasking. Because standard decoding selects tokens based on predicted textual likelihood, standard confidence-based unmasking remains vulnerable to language shortcut bias. As a strong inference-time baseline, we adapt Visual Contrastive Decoding (VCD) [15] to the masked diffusion framework. Originally designed for autoregressive models, we implement VCD by contrasting logit distributions from the original image against a visually distorted counterpart at each parallel unmasking step. This contrast penalizes reliance on linguistic priors when visual evidence is intentionally degraded.

Implementation Details. We build VISAGE decoding upon the MMaDA architecture. Our method uses a fixed spatial entropy quantile threshold $\beta = 0.25$ across all experiments. For the grounding penalty, we apply a default of $\alpha = 0.5$ for most visually intensive benchmarks. However, for MME, we reduce the penalty to $\alpha = 0.3$. Because MME uniquely evaluates both fine-grained perception and complex cognition tasks such as commonsense reasoning and translation, this lower penalty ensures the model retains the necessary linguistic and cognitive priors without over-regularizing the visual grounding. We justify this in Section 6.1 via ablation in Table 3. We omit restrictive post-prompts and extend the maximum generation lengths to elicit intermediate token sequences. All evaluated methods share identical generation configurations within a given benchmark for fair comparison. While method-specific hyperparameters remain constant, generation-specific parameters vary by task complexity. For POPE, HallusionBench, and MME, we set the generation length and diffusion steps to 256, with a block length of 32. For MMMU-val, we maintain a generation length of 256, but utilize 128 diffusion steps and a block length of 64.

Evaluation. We evaluate model performance across the extended generation trajectories using two strategies. For POPE and MME, we use string-matching rules, comparing the final responses against the ground-truth labels to compute F1 metrics and the MME score. For HallusionBench and MMMU-val, we adopt the Qwen3-8B [38] as an external LLM-as-a-judge. The LLM judge is prompted to analyze whether the generated output aligns with the ground truth.

Table 1: VISAGE improves visual reasoning across subject-specific and diagnostic benchmarks.. MMaDA + VISAGE (Ours) consistently outperforms the baseline and VCD across multiple benchmarks. We observe substantial gains over the base model, notably +2.65 on HallusionBench and +2.33% on MMMU-val.

Method	**MMMU-val** (Acc % ↑)	**HallusionBench** (Acc % ↑)	**POPE** (F1 ↑)	**MME** (Score ↑)
MMaDA (Base)	27.11	34.18	75.97	**1383.29**
MMaDA + VCD	28.44	34.80	75.85	1342.21
MMaDA + VISAGE (Ours)	**29.44**	**36.83**	**76.17**	1372.05
Relative Gain % (Ours vs. Base)	+8.59%	+7.75%	+0.26%	-0.81%

6 Main Results

Linear Ranking Score Reduces Language Shortcut Exploitation. We present results on hallucination and general purpose benchmarks in Table 1. Our method, VISAGE, yields improvements on hallucination-sensitive benchmarks, with a +7.75% relative gain on HallusionBench, +8.59% on MMMU-val, and +0.26% on POPE over the MMaDA baseline, while sustaining only a negligible relative decrease of -0.81% on MME. The large improvement on HallusionBench reflects its unique design. The benchmark exposes models that over-rely on textual priors when faced with deceptive visual contexts, this gain validates our core hypothesis. By penalizing ungrounded tokens, VISAGE prevents the decoder from defaulting to statistically plausible but unverified language shortcuts, forcing it to ground its generation in actual image content. Notably, these gains are obtained purely at decoding time, without any additional data, fine-tuning, or architectural modification to the underlying MMaDA model.

While POPE evaluates object-level hallucination, the relative gain (+0.26%) is compressed by systemic label noise in the underlying dataset. Because we enable intermediate reasoning traces, VISAGE performs exhaustive spatial verification before committing to a token. This rigorous grounding identifies valid objects that are absent from the ground-truth annotations. POPE is constructed using MS COCO [20], which is known to have missing labels and false positive object assertions [32]. Qualitative analysis reveals numerous instances where VISAGE rightly refutes incorrect ground-truth assertions (penalized as false negatives). Consequently, these inherited label inaccuracies artificially cap the measurable improvement on this benchmark. Detailed qualitative examples of these ground-truth failures are provided in the supplementary material.

Robustness to Deceptive Visual Contexts. Figure 4 illustrates the domains driving our overall performance gain on HallusionBench. VISAGE substantially expands the accuracy in highly spatial subsets, such as *illusion* (≈9%) and *map* (12.5%).

High accuracy in these categories requires precise feature localization to resolve visually deceptive or complex layouts. VCD offers negligible improvement in the *illusion* subset, suggesting that while contrastive decoding penalizes generic statistical priors, it fails to enforce the concentrated spatial verification needed to overcome deceptive visual prompts. Furthermore, we observe consistent improvements across dense, structured formats such as *figure*, *math*, and *OCR*, where fine-grained semantic details are easily hallucinated by language priors. However, the baseline retains an advantage in the *video* category. Video reasoning requires aggregating and grounding information across multiple frames, whereas our method derives grounding from frame-level attention without explicitly modeling temporal consistency. Extending the objective correction mechanism to capture cross-frame grounding remains an important direction for future work.

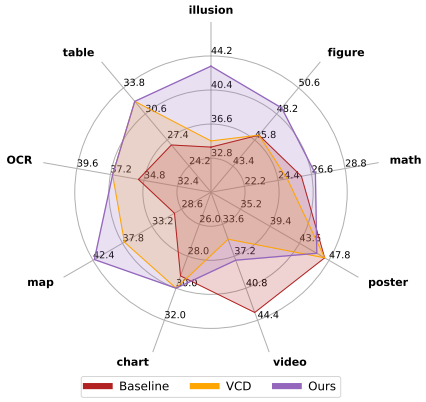


Fig. 4: HallusionBench Category Analysis. Radar chart comparing VISAGE against the Baseline and VCD. Our method achieves robust improvements across *illusion* and spatial-reasoning (*map*, *figure*) categories.

Grounding-Aware Re-ranking Stabilizes Multi-Step Reasoning. VISAGE achieves a +8.59% relative gain on MMMU-val (Table 1), which requires college-level, discipline specific multi-step reasoning. In such rigorous tasks, premature commitment to an ungrounded token during intermediate reasoning steps often triggers a cascading failure, where subsequent generations condition on the hallucinated prior rather than the visual input. By re-ranking candidate tokens via our corrective penalty to the confidence, VISAGE ensures that each step of the reasoning remains anchored to the image. This continuous regularization prevents the decoder from drifting into language shortcuts, stabilizing the generation of complex, long-form answers.

VISAGE Generalizes Across MDLLMs. To verify that the grounding-aware correction is not specific to MMaDA, we apply VISAGE to two additional multimodal diffusion LLMs, LLaDA-V and Lumina-DiMOO, without any architecture specific tuning (Table 2). On LLaDA-V, VISAGE improves MMMU-val from 46.67 to 47.11, and MME from

Table 2: VISAGE consistently improves performance on LLaDA-V and Lumina-DiMOO without any architecture-specific tuning, indicating that grounding-aware token commitment addresses a decoding-level issue.

Backbone	Method	MMMU-val (↑)	MME (↑)
LLaDA-V	Base	46.67	1883.57
	+ VISAGE	**47.11**	**1900.64**
Lumina-DiMOO	Base	29.33	1125.23
	+ VISAGE	**30.13**	**1142.16**

1883.57 to 1900.64. On Lumina-DiMOO, it improves MMMU-val from 29.33 to 30.13 and MME from 1125.23 to 1142.16. These consistent gains indicate that VISAGE is *model-agnostic* within the class of MDLLMs that expose token confidence and visual attention during masked decoding, supporting our central claim that grounding-aware token commitment addresses a broader decoding-level mismatch rather than limited to one architecture.

Overcoming Prompt-Driven Hallucinations. VISAGE neutralizes hallucinations induced by the language bias of the user's prompt. Figure 5 illustrates this token-level correction. In the top example, the baseline defaults to a statistical heuristic, asserting the spoon is "placed outside the bowl" rather than verifying the actual spatial coordinates. Our method successfully anchors the generation to the correct visual relationship. The bottom example highlights a more severe lexical trigger: the prompt asked if there are *"carrots"* in an image of engineering tools, both methods correctly conclude with "*No.*" However, this lexical trigger (*"carrots"*) causes MMaDA to undergo a language shortcut. It ignores the visual evidence of engineering tools and confidently hallucinates "*kitchen tools, including spatulas*" to satisfy the food-related prompt. Our method suppresses this prompt-driven language shortcuts, forcing the model to correctly identify "*pliers*" and "*wrenches.*"

6.1 Ablative Studies.

Effect of Penalty Factor α. The penalty factor α scales the penalty applied to the language model's confidence (log-probability). It interpolates between language-driven confidence and the grounded multimodal objective, controlling how strongly visually ambiguous tokens are downweighted during commitment.

The ablation results in Table 3 support this interpretation. On the MME benchmark, which includes cognition tasks such as Commonsense Reasoning, Text Translation, Numerical Calculation, Code Reasoning, a moderate penalty ($\alpha = 0.3$) achieves the best score (1372.05). It outperforms a stronger penalty ($\alpha = 0.5$, yielding 1320.77), suggesting that excessive penalization can degrade performance when visual grounding is not the primary source of hallucination. Conversely, applying a weak penalty ($\alpha = 0.1$, yielding 1362.68) fails to suppress premature token commitments. Without sufficient regularization of the model's confidence, the decoder is not forced to seek concentrated visual evidence, leaving it vulnerable to the language shortcuts our method aims to prevent.

Table 3: Effect of α. Sensitivity analysis of the grounding penalty on MME scores.

Penalty (α)	**MME** (Score ↑)
0.5	1320.77
0.3	**1372.05**
0.1	1362.68

Robustness via β-Quantile Consensus. Table 4 validates the necessity of our β-quantile consensus (q_β) for robust grounding entropy. On the MMMU-val

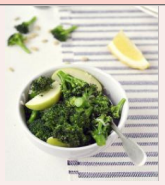
Prompt	Generation
 **Question:** Is the spoon in the bowl? Please answer yes or no.	**MMaDA:** <think> The image shows a bowl of broccoli with a lemon slice and a lemon wedge in the background. The spoon is not inside the bowl, as it is placed outside the bowl.</think> No **OUR Method:** <think> The image shows a bowl of broccoli with a lemon slice and a lemon wedge in the background. The spoon is placed in the bowl, indicating that it is in the bowl.</think> Yes
	Ground Truth: yes
Prompt	**Generation**
 **Question:** Are there any carrots in this figure?	**MMaDA:** <think> The image shows a row of kitchen tools, including spatulas, spatulas, and a tape measure. There are no carrots visible in the image. The answer is no.</think> No **OUR Method:** <think> The image shows a row of tools, including pliers, wrenches, and other hand tools, arranged on a white background. There are no carrots visible in the image. The tools appear to be related to construction or repair work.</think> No
	Ground Truth: No, all the items here are engineering tools

Fig. 5: Qualitative comparison of visually grounded yes/no questions. We compare MMaDA and our method on two examples where hallucination arises from language shortcuts. **Top:** MMaDA incorrectly answers "*No*" by over-relying on the textual prior that a spoon rests outside a bowl. Our method correctly grounds its "*Yes*" decision in localized visual evidence. **Bottom:** While both methods correctly answer "*No*", MMaDA hallucinates irrelevant objects in its intermediate thinking trace. In contrast our method maintains strictly visually consistent descriptions.

benchmark, the q_β strategy achieves 29.44% accuracy, outperforming both mean (28.78%) and minimum (28.56%) aggregation.

Table 4: Ablative Study on head aggregation. Quantile aggregation improves robustness over mean pooling.

Aggregation	**MMMU-val** (Acc% ↑)
Min over Heads	28.56
Mean over Heads	28.78
q_β (Ours)	**29.44**

In multimodal transformers, individual attention heads can produce spuriously sharp distributions even when a token is not grounded. A minimum-entropy is therefore vulnerable to a single sharp head falsely signaling grounding. Conversely, mean pooling dilutes genuine localized signals if most heads remain diffuse. The q_β consensus provides a principled compromise: it requires that visual grounding is consistently concentrated across a non-trivial consensus of heads before assigning grounding penalty. By filtering out both isolated artifacts and uninformative diffuse heads, q_β consensus ensures stable and reliable spatial verification during token commitment.

Inference Cost and Entropy Stride. Because VISAGE only reweights existing token proposals using attention already computed in the forward pass, it

introduces a controllable inference-time cost rather than an architectural overhead. Table 5 reports latency averaged over MMMU-val samples with generation length 256 and 256 diffusion steps on an NVIDIA GH200. Computing the attention-entropy grounding signal at every denoising step raises latency from 8.22 to 10.84 s/image (1.31×).

Since the grounding signal varies slowly across adjacent steps, VISAGE can instead estimate it only at a subset of steps, which we call the *entropy stride*. Increasing the stride steadily reduces overhead: stride 2 takes 9.56 s/image (1.16×), stride 4 takes 8.58 s/image (1.04×), and stride 8 takes 8.23 s/image (1.00×), nearly identical to the baseline. The entropy stride therefore offers a practical accuracy-latency trade-off while retaining the same decoding framework.

Table 5: Latency in seconds per image, averaged over MMMU-val samples (generation length 256, 256 diffusion steps). Computing the grounding signal only every few denoising steps reduces overhead to nearly the baseline cost.

Method	Latency (s/image) ↓	Relative ↓
MMaDA (Base)	8.22	1.00×
VISAGE (every step)	10.84	1.31×
VISAGE (stride 2)	9.56	1.16×
VISAGE (stride 4)	8.58	1.04×
VISAGE (stride 8)	8.23	1.00×

7 Conclusion

We identify a structural limitation in Multimodal Diffusion Large Language Models (MDLLMs): probability-based masked decoding optimizes a language-only proxy objective, inducing hallucination through language shortcuts. By reframing parallel unmasking as a localized optimization error under objective misspecification, we establish that hallucinated tokens arise from a quantifiable proxy discrepancy rather than generative capacity failure. To resolve this objective mismatch, we introduce VISAGE, a training-free re-ranking framework that calibrates the decoding objective at inference time. By deriving a discrepancy estimator from the spatial entropy of cross-attention distributions and enforcing a localization consensus, VISAGE downweights tokens lacking concentrated visual probability mass. Evaluations across hallucination-sensitive and general-purpose benchmarks demonstrate the robustness of the framework's robustness. The results establish that structural interventions on the discrete decoding objective, without parameter updates, resolve multimodal misalignments in parallel decoding architectures.

Limitations and Future Work. VISAGE addresses visual grounding during parallel masked decoding for image-to-text generation. The current framework does not explicitly incorporate the temporal token dimensions characteristic of video-based MDLLMs. While spatial entropy generalizes to spatio-temporal attention maps, VISAGE does not enforce cross-frame temporal consistency. Extending the entropy-based consensus to explicitly model motion dynamics and temporal grounding in video-based parallel decoding represents a promising direction for future research.

References

1. Austin, J., Johnson, D.D., Ho, J., Tarlow, D., Van Den Berg, R.: Structured denoising diffusion models in discrete state-spaces. Advances in neural information processing systems **34**, 17981–17993 (2021)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Deng, F., Wang, Q., Wei, W., Hou, T., Grundmann, M.: Prdp: Proximal reward difference prediction for large-scale reward finetuning of diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7423–7433 (2024)
4. Dong, H., Xiong, W., Goyal, D., Zhang, Y., Chow, W., Pan, R., Diao, S., Zhang, J., Shum, K., Zhang, T.: Raft: Reward ranked finetuning for generative foundation model alignment. arXiv preprint arXiv:2304.06767 (2023)
5. Fioresi, J., Dave, I.R., Shah, M.: Ted-spad: Temporal distinctiveness for self-supervised privacy-preservation for video anomaly detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13598–13609 (2023)
6. Fioresi, J., Dave, I.R., Shah, M.: Albar: Adversarial learning approach to mitigate biases in action recognition. In: The Thirteenth International Conference on Learning Representations (2025)
7. Fioresi, J., Dave, I.R., Shah, M.: Privacy beyond pixels: Latent anonymization for privacy-preserving video understanding. In: International Conference on Learning Representations (2026)
8. Fioresi, J., Kulkarni, P.P., Vayani, A., Wang, S., Shah, M.: Learning to share: Selective memory for efficient parallel agentic systems. In: Proceedings of the International Conference on Machine Learning (ICML) (2026)
9. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025)
10. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: 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. pp. 6904–6913 (2017)
11. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., et al.: Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14375–14385 (2024)
12. Gupta, A., Parmar, J., Dave, I.R., Shah, M.: From play to replay: Composed video retrieval for temporally fine-grained videos. Advances in Neural Information Processing Systems **38** (2026)
13. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
14. Kalai, A.T., Nachum, O., Vempala, S.S., Zhang, E.: Why language models hallucinate. arXiv preprint arXiv:2509.04664 (2025)
15. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding.

In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13872–13882 (2024)
16. Li, S., Kallidromitis, K., Bansal, H., Gokul, A., Kato, Y., Kozuka, K., Kuen, J., Lin, Z., Chang, K.W., Grover, A.: Lavida: A large diffusion language model for multimodal understanding. arXiv preprint arXiv:2505.16839 (2025)
17. Li, X., Thickstun, J., Gulrajani, I., Liang, P.S., Hashimoto, T.B.: Diffusion-lm improves controllable text generation. Advances in neural information processing systems **35**, 4328–4343 (2022)
18. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.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 (2023)
19. Liang, Z., Yuan, Y., Gu, S., Chen, B., Hang, T., Li, J., Zheng, L.: Step-aware preference optimization: Aligning preference with denoising performance at each step. arXiv preprint arXiv:2406.04314 **2**(5), 7 (2024)
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
21. 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. arXiv preprint arXiv:2502.09992 (2025)
22. Peters, J., Schaal, S.: Reinforcement learning by reward-weighted regression for operational space control. In: Proceedings of the 24th international conference on Machine learning. pp. 745–750 (2007)
23. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. Advances in neural information processing systems **36**, 53728–53741 (2023)
24. Robinson, D., Gupta, A., Clark, E., Melnik, O., Fu, Q., Shah, M.: Enhancing box and block test with computer vision for post-stroke upper extremity motor evaluation. arXiv preprint arXiv:2603.29101 (2026)
25. Robinson, D., Gupta, A., Qureshi, R., Fu, Q., Shah, M.: Strokevision-bench: A multimodal video and 2d pose benchmark for tracking stroke recovery. In: 2025 IEEE 35th International Workshop on Machine Learning for Signal Processing (MLSP). pp. 1–6. IEEE (2025)
26. Sarkar, S., Che, Y., Gavin, A., Beerel, P.A., Kundu, S.: Mitigating hallucinations in vision-language models through image-guided head suppression. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 12492–12511 (2025)
27. Shen, D., Song, G., Xue, Z., Wang, F.Y., Liu, Y.: Rethinking the spatial inconsistency in classifier-free diffusion guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9370–9379 (2024)
28. Shen, Y., Jiang, X., Yang, Y., Wang, Y., Han, D., Li, D.: Understanding and improving training-free loss-based diffusion guidance. Advances in Neural Information Processing Systems **37**, 108974–109002 (2024)
29. Shi, C., Ni, H., Li, K., Han, S., Liang, M., Min, M.R.: Exploring compositional visual generation with latent classifier guidance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 853–862 (2023)
30. Sun, B., Zhai, K., Shah, M., Wang, Z.: Dystruct: Dynamically structured diffusion language model decoding via bayesian inference (2026), `https://arxiv.org/abs/2605.09820`
31. Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M.S., Love, J., et al.: Gemma: Open models based on gemini research and technology. arXiv preprint arXiv:2403.08295 (2024)

32. Tkachenko, U., Thyagarajan, A., Mueller, J.: Objectlab: Automated diagnosis of mislabeled images in object detection data. arXiv preprint arXiv:2309.00832 (2023)
33. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8228–8238 (2024)
34. Wallace, B., Gokul, A., Ermon, S., Naik, N.: End-to-end diffusion latent optimization improves classifier guidance. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7280–7290 (2023)
35. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
36. Xin, Y., Qin, Q., Luo, S., Zhu, K., Yan, J., Tai, Y., Lei, J., Cao, Y., Wang, K., Wang, Y., et al.: Lumina-dimoo: An omni diffusion large language model for multimodal generation and understanding. arXiv preprint arXiv:2510.06308 (2025)
37. Xu, X., Chen, H., Lyu, M., Zhao, S., Xiong, Y., Lin, Z., Han, J., Ding, G.: Mitigating hallucinations in multi-modal large language models via image token attention-guided decoding. In: Chiruzzo, L., Ritter, A., Wang, L. (eds.) Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 1571–1590. Association for Computational Linguistics, Albuquerque, New Mexico (Apr 2025). https://doi.org/10.18653/v1/2025.naacl-long.75, https://aclanthology.org/2025.naacl-long.75/
38. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
39. Yang, L., Tian, Y., Li, B., Zhang, X., Shen, K., Tong, Y., Wang, M.: MMaDA: Multimodal large diffusion language models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=wczmXLuLGd
40. Ye, J., Xie, Z., Zheng, L., Gao, J., Wu, Z., Jiang, X., Li, Z., Kong, L.: Dream 7b: Diffusion large language models. arXiv preprint arXiv:2508.15487 (2025)
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), 220105 (2024)
42. Yousaf, A., Fioresi, J., Beetham, J., Bedi, A.S., Shah, M.: Safer-clip: Mitigating nsfw content in vision-language models while preserving pre-trained knowledge. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 36012–36020 (2026)
43. Yu, J., Wang, Y., Zhao, C., Ghanem, B., Zhang, J.: Freedom: Training-free energy-guided conditional diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23174–23184 (2023)
44. Yu, R., Ma, X., Wang, X.: Dimple: Discrete diffusion multimodal large language model with parallel decoding. arXiv preprint arXiv:2505.16990 (2025)
45. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9556–9567 (2024)
46. Zhai, K., Mollah, S., Wang, Z., Shah, M.: CORE: Context-robust remasking for diffusion language models. In: Forty-third International Conference on Machine Learning (2026), https://openreview.net/forum?id=bmKHxLWkz9

47. Zhang, F., Wu, Y., Wang, Z., Wang, X., Lv, C., Huang, X., Zheng, X.: Vib-probe: Detecting and mitigating hallucinations in vision-language models via variational information bottleneck. arXiv preprint arXiv:2601.05547 (2026)
48. Zhao, L., Deng, Y., Zhang, W., Gu, Q.: Mitigating object hallucination in large vision-language models via image-grounded guidance. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=w0xYx9CJhY