

Attention-Aware Remasking: A Training-Free Approach for Enhancing Visual Fidelity in Multimodal Diffusion Language Models

Yechan Cha and Dong-Kyu Chae

Department of Artificial Intelligence, Hanyang University, Seoul, South Korea
{chayechan, dongkyu}@hanyang.ac.kr

Abstract

Multimodal diffusion Large Language Models (dMLLMs) emerge as a compelling alternative to traditional autoregressive paradigms. But they still suffer from hallucinations driven by an inherent language bias. Through an analysis of the dMLLM inference dynamics, we observe that visual tokens frequently receive disproportionately low attention scores. We identify that the conventional low-confidence remasking strategy redundantly leverages logit-based probabilities during the token selection process, thereby exacerbating this language bias. To address this vulnerability, we propose *Attention-Aware Remasking*, a training-free, plug-and-play strategy that incorporates newly defined visual confidence scores to select the final output tokens, effectively promoting visual grounding. To enhance the logical integrity and inference efficiency of dLLMs, we further propose *clue-based answering* and *echo-filling*. Evaluations on representative Visual-Language Natural Language Explanation (VL-NLE) tasks, including VQA-X and e-SNLI-VE, demonstrate that our method achieves the highest BERTScore on both VQA-X and e-SNLI-VE, and the best answer accuracy on VQA-X among the compared methods. Our study reveals that the widely adopted logit-based remasking is not necessarily the optimal approach, highlighting promising new avenues for visually grounded inference in dMLLMs. Our code is available at: <https://github.com/yechnacha/Attention-aware-Remasking>.

1 Introduction

Recent advances in Large Vision-Language Models (LVLMs) (Alayrac et al., 2022; Liu et al., 2023), which integrate visual and linguistic information, have highlighted the importance of modality alignment (Li et al., 2023; Leng et al., 2024). This concern has motivated extensive research on mitigating hallucinations caused by visual-textual misalign-

ment in these frameworks, as such failures reveal a *lack of visual fidelity*: models may generate linguistically plausible responses that are insufficiently grounded in the actual visual evidence (Wang et al., 2025; Zhao et al., 2025b).

Concurrently, diffusion Large Language Models (dLLMs) are rapidly emerging as a compelling alternative to the conventional autoregressive (AR) paradigm (Li et al., 2025b; Nie et al., 2025; Ye et al., 2025). Unlike traditional AR models that generate tokens sequentially from left to right, dLLMs can generate multiple tokens simultaneously in a parallel and non-sequential manner. Leveraging these architectural advantages, recent efforts have extended this approach into diffusion multimodal large language models (dMLLMs), which process visual and linguistic features within a unified token space (You et al., 2026; Li et al., 2025a).

As noted earlier, a significant challenge in LVLMs remains the lack of visual fidelity. Recent studies have observed that these models frequently assign disproportionately low attention scores to visual inputs, leading to hallucinations where they mention non-existent objects or contradict visual facts (Chen et al., 2024; Zhao et al., 2025a). This points to a strong language bias, where the model prioritizes learned linguistic patterns over actual visual evidence. However, whether this vulnerability persists within dMLLMs remains under-explored, highlighting a need to investigate the inference dynamics of dMLLMs and develop effective strategies to enhance their visual alignment.

By tracing the inference process of dMLLMs, we observed that the issue of disproportionately low attention scores being allocated to visual tokens persists in these models as well. During inference, dMLLMs simultaneously generate predictions for all masked tokens, retain only the tokens deemed most confident as outputs, and remask the remainder. As illustrated in Figure 1, the conventional *low-confidence remasking* (Nie et al., 2025; You

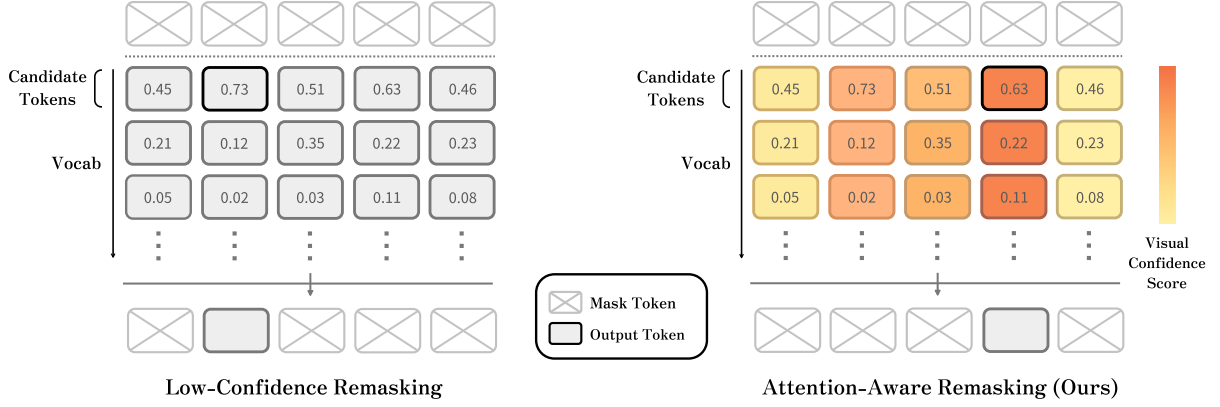


Figure 1: Token selection processes in low-confidence and our attention-aware remasking. In both methods, the token with the highest logit-based softmax probability at each mask index is initially selected as the candidate token. When determining the final output tokens from these candidates, low-confidence remasking redundantly re-evaluates the aforementioned softmax probabilities. In contrast, our attention-aware remasking compares the visual confidence scores, which represent the degree of attention allocated to the visual tokens at this stage.

et al., 2026) strategy redundantly leverages logits during this process, thereby exacerbating the inherent language bias. While concurrent efforts explore spatial attention concentration for visual grounding (e.g., VISAGE (Narnaware et al., 2026)), we tackle the modality imbalance between visual and language tokens during token commitment.

Therefore, we propose Attention-Aware Remasking, a novel strategy designed to explicitly enforce visual grounding during the iterative denoising process. Instead of redundantly relying on logit probabilities, our strategy introduces a visual confidence score, which is computed by averaging the top self-attention scores allocated to visual tokens, to selectively commit tokens grounded in visual evidence while remasking those driven merely by language priors. Furthermore, we propose two additional techniques designed to enhance both logical integrity and inference efficiency. (1) *Clue-based Answering*: We explicitly partition the mask sequence into two distinct blocks—the explanation block and the answer block—based on the strategic placement of pre-filled tokens. By enforcing the complete generation of the explanation block before the answer block is inferred, this strategy effectively mitigates post-hoc rationalization. (2) *Echo-Filling*: This algorithm detects and populates non-informative padding spaces without inference, effectively preventing the waste of computational resources without degrading performance. The proposed techniques are training-free and operate as a plug-and-play module at the inference stage.

To demonstrate the effectiveness of the proposed framework, we evaluated it on two representative

Visual-Language Natural Language Explanation (VL-NLE) tasks: VQA-X (Park et al., 2018) and e-SNLI-VE (Kayser et al., 2021). Our experimental results show that the proposed framework achieves the best performance in semantic alignment for generated explanations (measured by BERTScore) across both benchmarks, as well as the best VQA task accuracy on VQA-X. Our ablation studies verify that the proposed attention-aware remasking effectively promotes visual grounding. These findings reveal that the widely adopted low-confidence remasking strategy is not necessarily the optimal approach, highlighting the potential for further refinements in remasking protocols.

2 Related Works

2.1 Diffusion Language Models

While traditional Large Language Models (LLMs) predominantly rely on the autoregressive (AR) paradigm, Diffusion Language Models (DLMs) have emerged as a compelling alternative offering bidirectional context modeling and parallel generation (Li et al., 2025b). To overcome the discretization artifacts inherent in early continuous formulations (Li et al., 2022; Lin et al., 2023), the paradigm shifted toward discrete Masked Diffusion Language Models (MDLMs) (Sahoo et al., 2024). By defining the diffusion process directly at the token level using an absorbing state (e.g., a [MASK] token), MDLMs simplify complex diffusion theories into tractable cross-entropy losses. This streamlined approach has unlocked unprecedented scalability, as demonstrated by large-scale

models like LLaDA (Nie et al., 2025) and Dream (Ye et al., 2025). Driven by this success, pioneering diffusion multimodal large language models (dMLLMs), such as LLaDA-V (You et al., 2026) and LaViDa (Li et al., 2025a), have successfully integrated visual and linguistic features within a unified discrete token space.

2.2 Natural Language Explanation for Vision-Language Task

The generation of Natural Language Explanations (NLE) in Vision-Language (VL) tasks is essential for ensuring system trustworthiness, as it provides transparency into the model’s internal reasoning. Evaluations in this domain typically rely on benchmark datasets that emphasize both visual evidence and logical reasoning, such as VQA-X (Park et al., 2018) and e-SNLI-VE (Kayser et al., 2021).

Despite significant architectural advancements, a fundamental challenge in VL-NLE remains the prevalence of hallucinations and a lack of visual fidelity. Specifically, these models frequently suffer from inherent language and exposure biases, leading to post-hoc rationalization—where they generate plausible, linguistically coherent text that fails to reflect their actual decision-making process or the underlying visual evidence. Consequently, recent research trajectories have focused on mitigating these hallucinations and enforcing strict visual-language alignment. To this end, various strategies, ranging from attention-based visual justification (Park et al., 2018; Sammani et al., 2022) to self-critical learning (Suo et al., 2023) and reinforcement learning (Chang et al., 2024), have been actively explored to ensure that textual explanations are firmly rooted in visual evidence.

3 Analysis of Attention Distribution

It has been recently documented that visual tokens in LVLMMs are frequently assigned disproportionately low attention scores (Chen et al., 2024). Zhao et al. (2025a) empirically demonstrated that this visual attention deficit is the root cause of severe visual hallucinations, compelling the model to ignore genuine visual information and fall back on the language bias of its pre-trained backbone.

Despite these findings, the prevalence of this attention disparity within dMLLMs remains under-investigated. To bridge this gap, we adapted the attention distribution analysis of output tokens introduced by Zhao et al. (2025a) to the dMLLM

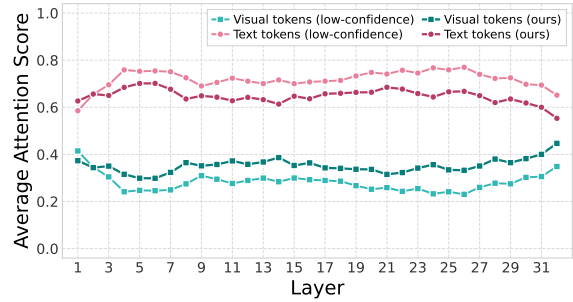


Figure 2: Average attention scores of output tokens toward textual and visual tokens across different layers in LLaDA-V. Compared to conventional low-confidence remasking, our Attention-Aware Remasking substantially narrows the attention gap between textual and visual tokens. The results are averaged over 300 test samples from VL-NLE benchmarks (VQA-X and e-SNLI-VE).

architecture. In contrast to conventional AR models that produce output tokens sequentially in a strict left-to-right fashion, MDMs operate via an iterative denoising process across multiple masked positions to progressively finalize tokens. Accordingly, for our analysis, we treat the tokens finally unmasked by the remasking strategy at a given denoising step as the direct equivalents of the output tokens in an AR decoding step.

As illustrated in Figure 2, our measurements on VL-NLE test samples utilizing LLaDA-V confirm that this attention imbalance is distinctly present in the conventional low-confidence remasking baseline. While MDMs and AR models are driven by fundamentally different training objectives, their architectural blueprints are remarkably similar. This high degree of compatibility is exemplified by Dream-7B (Ye et al., 2025), which was successfully initialized directly from Qwen2.5 (Yang et al., 2024) weights. Therefore, it is reasonable to deduce that this inherent structural flaw has been directly inherited by the diffusion framework.

Importantly, when our Attention-Aware Remasking is applied, this attention disparity between textual and visual tokens is significantly reduced across transformer layers. By incorporating visual confidence into token commitment, our method prevents over-reliance on language priors and encourages the model to allocate higher attention weights to visual tokens. Consequently, this leads to generated explanations that are more faithfully grounded in visual evidence.

4 Preliminary

Traditional autoregressive (AR) paradigm of LLMs focuses on “next-token prediction”. In contrast, Masked Diffusion Models (MDMs) define a model distribution through a forward masking process and a corresponding reverse process for discrete tokens.

The forward process $\{x_t\}$ is indexed by time $t \in [0, 1]$, where information from the clean data x_0 is gradually corrupted by injecting a special mask token $[M]$. At any given time t , the conditional distribution of x_t given x_0 is defined in a fully factorized form:

$$q_{t|0}(x_t|x_0) = \prod_{i=1}^L q_{t|0}(x_t^i|x_0^i) \quad \text{where} \quad (1)$$

$$q_{t|0}(x_t^i|x_0^i) = \begin{cases} 1-t, & x_t^i = x_0^i \\ t, & x_t^i = \mathbf{M} \end{cases}$$

Here, each token is independently masked with probability t . At $t = 1$, the entire sequence is guaranteed to be masked. This linear masking schedule is motivated by the principle that information in text is generally proportional to the number of tokens, making linear information loss a reasonable assumption for the forward process.

The reverse process recovers the data distribution by iteratively predicting masked tokens as t moves from 1 to 0, starting from a fully masked sequence. For a single reverse step where $0 \leq s < t \leq 1$, the conditional transition is defined as:

$$q_{s|t}(x_s|x_t) = \prod_{i=1}^L q_{s|t}(x_s^i|x_t^i) \quad \text{where} \quad (2)$$

$$q_{s|t}(x_s^i|x_t^i) = \begin{cases} 1, & x_t^i \neq \mathbf{M}, x_s^i = x_t^i \\ \frac{s}{t}, & x_t^i = \mathbf{M}, x_s^i = \mathbf{M} \\ \frac{t-s}{t} p_\theta(x_s^i|x_t), & x_t^i = \mathbf{M}, x_s^i \neq \mathbf{M} \\ 0, & \text{otherwise} \end{cases}$$

During inference, to align with the theoretical reverse dynamics, the mask predictor p_θ predicts all masked tokens simultaneously at each step, followed by re-masking a fraction $\frac{s}{t}$ of the tokens. This process adopts the *low-confidence remasking strategy* (Nie et al., 2025; You et al., 2026) inspired by Chang et al. (2022), where tokens with the lowest prediction confidence are selected for re-masking in the subsequent iteration. Specifically, at each step, every masked position is temporarily assigned the token from the vocabulary with the

highest predicted probability, alongside its corresponding softmax score. The model then commits only the top $(1 - \frac{s}{t})$ fraction of tokens—those with the highest softmax scores—to the sequence. The remaining tokens, falling below this confidence threshold, remain masked to be re-predicted in the following step.

The core of the MDM is the mask predictor p_θ , which is trained using a cross-entropy loss optimized specifically for masked tokens. The training objective $\mathcal{L}(\theta)$ is defined as (Nie et al., 2025):

$$\mathcal{L}(\theta) \triangleq -\mathbb{E} \left[\frac{1}{t} \sum_{i=1}^L \mathbf{1}[x_t^i = \mathbf{M}] \log p_\theta(x_0^i|x_t) \right], \quad (3)$$

where t is sampled uniformly from $[0, 1]$, and the indicator function $\mathbf{1}[\cdot]$ ensures that the loss is only computed at masked positions. Theoretically, this objective has been proven to be an upper bound on the negative log-likelihood of the model distribution (Shi et al., 2024; Ou et al., 2025), establishing MDMs as principled generative models.

To adapt this for multimodal tasks, visual tokens v and text prompts p act as fully unmasked, persistent conditions (You et al., 2026). The mask predictor p_θ learns to reconstruct a target sequence x of length L via the Multimodal SFT objective:

$$-\mathbb{E} \left[\frac{1}{t} \sum_{i=1}^L \mathbf{1}[x_t^i = \mathbf{M}] \log p_\theta(x_0^i|x_t, v, p) \right], \quad (4)$$

This conditional formulation enables dMLLMs to generate visually grounded text through iterative denoising.

5 Method

We employ one of the most recent dMLLMs, LLaDA-V (You et al., 2026), as our base model. We adapt LLaDA-V to VL-NLE tasks through Supervised Fine-Tuning (SFT) on the target datasets VQA-X (Park et al., 2018) and e-SNLI-VE (Kayser et al., 2021). We employ Low-Rank Adaptation (LoRA) (Hu et al., 2022) for efficient SFT. More training details are provided in Appendix A.1.

5.1 Attention-Aware Remasking

Given the persistent visual attention deficit in dM-LLMs, the standard token selection mechanism is fundamentally vulnerable to language bias. Specifically, conventional low-confidence remasking exacerbates this by favoring high linguistic proba-

bilities, which further suppresses weak visual signals. To actively counteract this issue, we propose *Attention-Aware Remasking*, a novel strategy designed to directly integrate visual attention scores into the token selection process, enforcing strict visual grounding during iterative denoising.

Since low-confidence remasking was shown to be more effective than random remasking (Nie et al., 2025), recent MDM variants (Dong et al., 2025; Zhou et al., 2026; Zhu et al., 2025) have widely adopted this approach. This remasking process consists of two primary stages: (1) identifying the candidate token with the highest probability for each masked position, and (2) selecting the final subset of these candidates to be officially committed to the sequence. As described in the previous section, low-confidence remasking relies on logits as the sole criterion for both stages. We hypothesize that this redundancy exacerbates the model’s language bias, causing it to prioritize linguistic context over visual evidence.

To address this, we incorporate visual confidence scores into the remasking process. While the first stage (predicting the most likely token based on logits) is tied to the model’s fundamental training objective and thus difficult to modify, the second stage (the selection process) is a discretionary procedure appended during inference, offering flexibility for optimization. Therefore, while maintaining the conventional logit-based approach for identifying candidate tokens, we select the final tokens by evaluating the attention scores between the predicted tokens and the visual tokens. The specific procedure is detailed in Algorithm 1.

We extract the attention scores directed toward the visual tokens across the entire generation sequence from the final transformer layer, yielding a tensor $A \in \mathbb{R}^{L \times v \times h}$ (sequence length $\times$ visual tokens $\times$ attention heads). To condense this information into a robust metric for each sequence position, we perform a two-step reduction using top- n and top- m averaging. Specifically, we first calculate the mean of the top- n highest scores across the head dimension, reducing the tensor to $\mathbb{R}^{L \times v}$. Subsequently, we compute the mean of the top- m highest scores across the visual token dimension, yielding a final visual confidence score $c_{\text{vis}} \in \mathbb{R}^L$. During the remasking phase, tokens falling within the bottom $\frac{s}{t}$ fraction of c_{vis} are reverted to [MASK] tokens. To ensure sequence stability, tokens that are not subject to prediction (e.g., pre-filled prompts or tokens already committed in prior steps) are

Algorithm 1 Attention-aware Remasking

Require: mask predictor p_θ , prompt p_0 , image tokens v , answer length L , steps N , attention heads h

- 1: Set r_1 as a fully masked sequence of length L .
- 2: **for** $t = 1 \rightarrow \frac{1}{N}$ **step** $-\frac{1}{N}$ **do**
- 3: $s \leftarrow t - \frac{1}{N}$
- 4: $A \leftarrow \text{LastLayerAttn} \in \mathbb{R}^{L \times v \times h}$
- 5: **for** $i = 1 \dots L$ **do**
- 6: **if** $r_t^i = M$ **then**
- 7: $r_0^i \leftarrow \arg \max_{r_0^i} p_\theta(r_0^i \mid v, p_0, r_t)$
- 8: $A'_i \in \mathbb{R}^v \leftarrow \text{AvgTop}_n(A_i)$
- 9: $c_{\text{vis}}^i \leftarrow \text{AvgTop}_m(A'_i)$
- 10: **else**
- 11: $r_0^i \leftarrow r_t^i, \quad c_{\text{vis}}^i \leftarrow 1$
- 12: **end if**
- 13: **end for**
- 14: $\mathcal{K} \leftarrow \text{LowestIndices}(\{c_{\text{vis}}^j\}_{j=1}^L, \lfloor L \cdot \frac{s}{t} \rfloor)$
- 15: $r_0^i \leftarrow M, \forall i \in \mathcal{K}$
- 16: $r_s \leftarrow r_0$
- 17: **end for**
- 18: **return** r_0

assigned a maximum confidence value of 1, effectively exempting them from remasking. By incorporating this strategy, we prioritize visual fidelity without compromising generation quality.

5.2 Dynamic Explanation Length

All labels in the VL-NLE dataset can be standardized into a fixed template: "Because [Explanation], the answer is [Answer]". Leveraging the unique capability of masked diffusion models to perform inference from partially filled sequences, we insert pre-filled tokens as constant structural anchors to enforce format consistency. However, placing these anchors at fixed positions creates a rigidly sized masked region for the explanation. Because the natural length of reasoning varies across instances, we allocate a maximum fixed capacity for the explanation block and introduce specialized [RESERVED] tokens to pad any remaining space. Through this strategy, the model dynamically terminates text generation when a concise explanation suffices, efficiently filling the residual masked positions with [RESERVED] tokens until the next structural anchor.

5.3 Clue-Based Answering

In previous studies (Park et al., 2018; Sammani et al., 2022; Suo et al., 2023), the answer was typically placed before the explanation in the generation sequence. This approach suffers from post-hoc rationalization: hallucinating plausible explanations to justify incorrect answers. This vulnerability necessitates the use of *filtered metrics* (evalu-

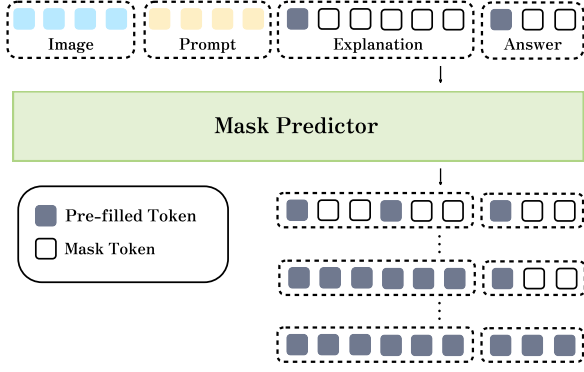


Figure 3: Overview of the clue-based answering process. To enforce a reasoning-first generation, tokens in the answer block are persistently re-masked until the explanation block is completely finalized.

ating explanations only for correct predictions) in standard evaluations. Recognizing this challenge, [Lai et al. \(2024\)](#) have adopted a reasoning-first approach within the AR paradigm by generating the explanation before deriving the final answer. To adapt this concept within the framework of dM-LLMs, we implement our clue-based answering strategy as follows.

We structured our pre-filled tokens to place the explanation before the answer. However, as masked diffusion models inherently generate tokens in a stochastic and non-sequential order across the entire sequence, simply arranging the pre-filled tokens does not guarantee that the explanation will be generated before the answer. Therefore, we explicitly define two distinct generation phases based on the pre-filled tokens as illustrated in Figure 3: the explanation block and the answer block. During the explanation generation phase, all tokens within the answer block, excluding the pre-filled tokens, are strictly re-masked at each iterative step. The model calculates re-masking priorities solely for the tokens within the explanation block, ensuring that the explanation is fully finalized without any influence from potential answer tokens. Only after the explanation block is entirely completed does the model begin the inference process for the answer block. This approach effectively prevents the model from retrofitting an explanation to a pre-determined answer while ensuring the answer is logically grounded in established evidence.

5.4 Echo-Filling

The aforementioned approach for dynamic explanation length shares a similar objective with the padding mechanism proposed in LaViDa ([Li et al.](#),

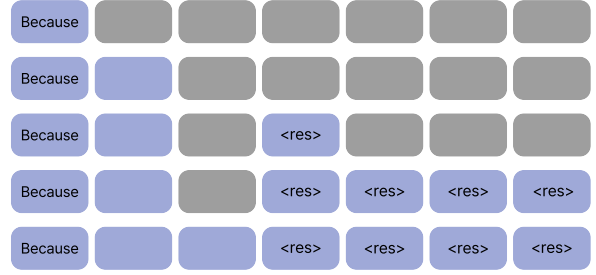


Figure 4: Inference acceleration via Echo-Filling. The mechanism auto-populates the remaining sequence upon detecting an [RESERVED] token, effectively bypassing redundant computations.

[2025a](#)). However, LaViDa-FIM possesses a structural limitation in that it requires continuous, redundant neural network inference operations to fill padding tokens ([S]), even after the actual content generation has effectively concluded.

To address the inference latency inherent in DLMs, several studies recently have introduced acceleration strategies ([Ma et al., 2026](#); [Wu et al., 2025](#)). We identify a unique logical property in our framework regarding the dynamic explanation length: if the model predicts a [RESERVED] token at a specific position, all subsequent consecutive mask tokens must uniformly be [RESERVED] tokens. Leveraging this logical consistency, we propose Echo-Filling, an inference optimization strategy designed to achieve significant acceleration by bypassing redundant computations.

Echo-Filling is triggered specifically when the model internally detects the occurrence of a [RESERVED] token. Once detected, the mechanism immediately populates all following mask tokens up to the next pre-filled token with [RESERVED] tokens, thereby bypassing any further neural network computations for those positions. By eliminating these redundant operations, Echo-Filling contributes to reducing total inference time while maintaining the quality of the generated explanations, as it only streamlines the generation of non-content padding tokens.

6 Experiments

To demonstrate the effectiveness of our methods, we employ VL-NLE tasks, which necessitate the generation of both precise answers and coherent justifications for given image-question pairs. We thus conduct extensive experiments on the most representative two benchmarks: VQA-X ([Park et al., 2018](#)) and e-SNLI-VE ([Kayser et al., 2021](#)).

Method	VQA-X									
	B@1	B@2	B@3	B@4	M	R	C	S	BS	Acc
NLX-GPT (Sammani et al., 2022)	64.2	49.5	37.6	28.5	23.1	51.5	110.6	22.1	86.9	83.0
S3C (Suo et al., 2023)	64.7	50.5	<u>38.8</u>	30.7	23.9	52.1	116.7	23.0	-	85.6
Revise (Ge et al., 2023)	64.6	50.0	<u>37.7</u>	28.2	23.2	51.8	108.9	22.6	88.1	-
MCLE [†] (Lai et al., 2024)	66.6	<u>50.7</u>	38.5	<u>31.2</u>	24.2	53.1	118.3	24.2	92.2	87.7
KICNLE [†] (Xie et al., 2024)	64.9	49.5	38.0	29.4	22.9	51.7	112.5	24.7	90.6	86.7
Rapper (Chang et al., 2024)	65.5	51.6	40.5	31.8	24.3	<u>52.9</u>	124.0	24.5	-	87.3
G2 (Yuan et al., 2025)	<u>66.3</u>	49.5	38.1	28.8	24.0	49.4	108.8	24.1	86.6	-
<i>Ours</i>	55.7	42.8	32.3	23.9	20.7	48.5	100.1	19.9	93.1	90.8

Method	e-SNLI-VE									
	B@1	B@2	B@3	B@4	M	R	C	S	BS	Acc
NLX-GPT (Sammani et al., 2022)	37.0	25.3	17.9	12.9	18.8	34.2	117.4	33.6	80.8	73.9
Revise (Ge et al., 2023)	38.3	26.5	19.0	13.8	19.7	34.7	<u>126.7</u>	34.2	81.5	-
MCLE [†] (Lai et al., 2024)	33.5	22.5	15.7	11.1	18.1	33.0	118.2	33.0	87.2	79.2
KICNLE [†] (Xie et al., 2024)	32.8	22.1	15.4	11.0	17.2	32.1	105.8	31.3	89.2	66.7
Rapper (Chang et al., 2024)	<u>40.5</u>	<u>28.1</u>	<u>20.2</u>	14.7	20.8	35.9	128.6	34.9	-	75.7
G2 (Yuan et al., 2025)	43.3	29.0	20.5	<u>13.9</u>	29.5	41.7	114.5	45.6	<u>90.4</u>	-
<i>Ours</i>	26.7	17.9	12.4	8.7	16.3	32.1	110.2	32.0	92.4	<u>76.1</u>

Table 1: Filtered scores (evaluating explanations only for correct predictions) comparison with state-of-the-art methods. B@N, M, R, C, S, and BS stand for BLEU-N, METEOR, ROUGE-L, CIDEr, SPICE, and BERTScore, respectively. †: we reproduced their results by running their official code, while the results for the remaining methods were taken from their original papers due to the lack of publicly available implementations.

To ensure direct comparability with existing work, we adopt the filtered evaluation protocol (evaluating explanations exclusively for correctly answered predictions) as recent state-of-the-art VL-NLE baselines (Xie et al., 2024; Chang et al., 2024; Yuan et al., 2025) do not report unfiltered explanation metrics. Beyond this protocol alignment, evaluating unfiltered metrics provides little additional insight for our framework; because our Clue-Based Answering strategy generates explanations prior to final answers, our framework yields virtually identical performance across filtered and unfiltered settings, with differences remaining within ± 0.1 percentage points across all metrics.

For linguistic evaluation of generated explanations, we employ BLEU@N (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al.). However, these lexical-based metrics have a limitation in capturing paraphrasing and semantic nuances. We supplement our evaluation with BERTScore (Zhang et al., 2019) to measure contextual and semantic consistency. The accuracy of VQA tasks is also measured to evaluate the correctness of the predicted answers. We note that there is a trade-off between BERTScore and the VQA task accuracy: optimizing for either semantic explanation quality or answer correctness alone can compromise the other, making it particu-

larly challenging to achieve strong performance on both dimensions simultaneously.

We compare our framework against a set of state-of-the-art AR baselines in the VL-NLE domain, including NLX-GPT (Sammani et al., 2022), S3C (Suo et al., 2023), Revise (Ge et al., 2023), MCLE (Lai et al., 2024), KICNLE (Xie et al., 2024), Rapper (Chang et al., 2024), and G2 (Yuan et al., 2025). We also employ two recent dMLLMs, LLaDA-V (You et al., 2026) and LaViDa (Li et al., 2025a).

6.1 Performance Comparisons

6.1.1 Benchmarking on VL-NLE Tasks

As illustrated in Table 1, while our model yields suboptimal scores in traditional lexical metrics and moderate accuracy on e-SNLI-VE, it achieves the highest BERTScore across both benchmarks, alongside the highest answer accuracy on VQA-X among all compared methods. Notably, on the e-SNLI-VE dataset, the contrasting performance of models like MCLE (which favors accuracy) and KICNLE (which prioritizes BERTScore) indicates the inherent difficulty of optimizing both semantic richness and logical task performance. This discrepancy between moderate lexical scores and high-level semantic alignment highlights a key strength of our framework: our model transcends surface-level word matching to achieve deeper semantic alignment and content validity. Prior studies indi-

cate that traditional lexical metrics often correlate poorly with actual human judgment, particularly in complex reasoning tasks (Kayser et al., 2021; Salemi et al., 2025). Furthermore, evaluating models strictly on lexical overlap often fails to account for linguistic biases and semantically distinct “hard negatives” (Dumpala et al., 2024). In this context, our strong BERTScore and task accuracy indicate that our model successfully decouples semantics from rigid linguistic patterns, generating explanations that are genuinely grounded in logic rather than superficial text correlation.

Although it is not included in Table 1, the diffusion-based baseline LLaDA-V (You et al., 2026) serves as the backbone of our work after VL-NLE-specific fine tuning. Its accuracy is indirectly compared in Table 4. LaViDa (Li et al., 2025a) is additionally evaluated in Appendix A.3 as an alternative diffusion backbone to assess the generalizability of our proposed methods.

6.1.2 Human Evaluation

To verify whether the observed improvements in BERTScore correspond to explanation quality judged as valid justifications by humans, we conducted a human evaluation following widely-adopted protocols (Sammani et al., 2022; Lai et al., 2024; Yuan et al., 2025). We randomly sampled 200 instances from the common subset of test samples correctly answered by all compared models (NLX-GPT, MCLE, and ours) across both VQA-X and e-SNLI-VE datasets. For each dataset, we hired three annotators; they are graduate students at Hanyang University. They evaluated each generated explanation in a randomized, double-blind setup, rating whether it appropriately justifies the correct answer on a 4-point scale: Yes (1), Weak Yes (2/3), Weak No (1/3), and No (0). In the end, each annotator was paid a \$75 reward.

As shown in Table 2, our method achieves the highest human evaluation scores on both datasets. This result provides complementary evidence that our high BERTScore is not merely an artifact of semantic similarity metrics. Human annotators judged our explanations to better justify the predicted answers than strong VL-NLE baselines, supporting our claim that the proposed decoding strategy enhances explanation quality in a manner faithfully aligned with human judgment.

Method	VQA-X	e-SNLI-VE
NLX-GPT	0.881	0.735
MCLE	0.935	0.825
Ours	0.946	0.832

Table 2: Human evaluation scores on VQA-X and e-SNLI-VE benchmark subsets. Scores reflect the degree to which generated explanations appropriately justify the correct answers.

Remasking Method	CHAIR _s ↓	CHAIR _i ↓	Recall ↑
Low-confidence	7.2	5.7	36.0
Ours	5.7	4.4	36.9

Table 3: Object hallucination evaluation using CHAIR metrics without fine-tuning. Lower values are better for CHAIR_s and CHAIR_i, while higher is better for Recall.

6.1.3 Evaluation on Visual Grounding and Object Hallucination

To provide direct quantitative evidence for improved visual grounding beyond semantic metrics and attention visualizations, we additionally evaluate object hallucination on image captioning outputs using the CHAIR benchmark (Rohrbach et al., 2018). For a fair assessment of the decoding strategy itself, this evaluation is conducted using the base LLaDA-V model without VL-NLE fine-tuning, comparing conventional low-confidence remasking against our Attention-Aware Remasking. Table 3 reports the result. Our approach effectively mitigates both sentence-level (CHAIR_s) and instance-level (CHAIR_i) object hallucinations, while simultaneously improving object recall. These results demonstrate that substituting pure logit-based token commitment with visual-confidence-guided commitment effectively enhances physical visual fidelity and reduces hallucinated content, beyond improving surface-level semantic similarity to reference explanations.

6.2 Ablation Studies

In terms of VL-NLE performance, Table 4 summarizes the ablation results for our core components. First, we examine the impact of replacing our *attention-aware remasking* with the conventional low-confidence strategy. As shown in the table, reverting to the conventional strategy leads to a substantial degradation in explanation quality across both datasets. This demonstrates that, without explicit visual attention guidance, the model falls back on its pre-trained language bias, losing

Method	VQA-X		e-SNLI-VE	
	BS	Acc	BS	Acc
Full Framework	93.1	90.8	92.4	76.1
w/o Attention-aware Remask	85.4	69.3	73.1	74.3
w/o Clue-based answering	93.2	79.6	90.1	62.2
w/o Echo-filling	92.9	92.3	90.5	76.6

Table 4: Ablation study on the core components of our framework. "w/o Attention-aware Remasking" indicates reverting to the conventional low-confidence remasking strategy performed by LLaDA-V.

Method	Generation Length			
	64	128	256	512
w/ Echo-filling	3.07s	3.21s	4.96s	7.46s
w/o Echo-filling	15.61	30.90	64.84	145.37s

Table 5: Inference time across various sequence lengths. The results are averaged over 100 samples.

connection to the visual evidence.

Beyond enforcing visual grounding at the token level, the structural sequence of generation also plays an important role in the model’s reasoning capabilities. The ablation results demonstrate that disabling the *clue-based answering* strategy leads to a significant decline in prediction accuracy. This suggests that when the model is permitted to generate an answer before gathering sufficient evidentiary support through the explanation block, it frequently produces incorrect predictions. In other words, prioritizing the reasoning process as a prerequisite for inference ensures that the final answer is logically grounded in visual and textual clues.

In terms of inference time, Table 5 demonstrates the inference acceleration achieved by Echo-Filling. The experiments were conducted using a single NVIDIA H100 GPU, with results averaged over 100 samples across varying maximum token lengths (L). For all trials, the number of inference steps was configured at $\frac{L}{2}$, and the length of the explanation block was designated as 90% of the maximum token limit. Without Echo-Filling, inference time increases linearly in proportion to the maximum generation length. In contrast, Echo-filling maintains a nearly constant inference time. Importantly, as shown in Table 4, this speedup does not come at the expense of generation quality. The VL-NLE performance metrics remain tightly comparable, confirming that our strategy safely bypasses redundant padding computations without incurring notable loss in accuracy.

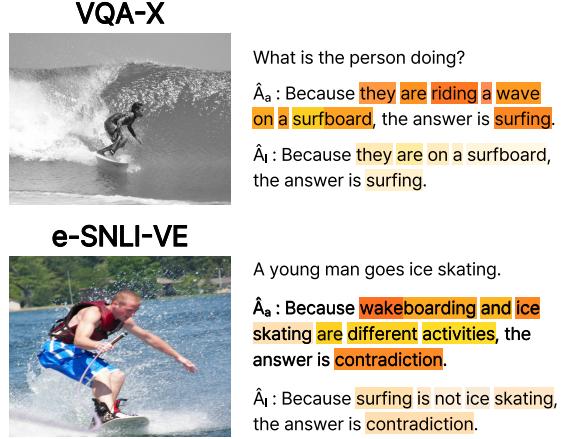


Figure 5: Qualitative examples generated by our framework. $\hat{A}_a$ denotes the model predictions obtained by applying attention-aware remasking, while $\hat{A}_l$ represents the results under low-confidence remasking.

6.3 Qualitative Analysis

Figure 5 presents several qualitative results from both datasets, accompanied by a visualization of the visual confidence scores associated with tokens generated under different remasking strategies. By tracking the visual confidence of tokens produced via the conventional low-confidence remasking strategy, we observe that visual tokens provide only a marginal direct contribution to the formation of token representations. Appendix A.6 shows more examples.

7 Conclusion

In this paper, we identified and addressed the inherent visual attention deficit that severely degrades the visual fidelity of dMLLMs. We revealed that the widely adopted low-confidence remasking strategy heavily relies on text-based logits, thereby trapping the model in its pre-trained language bias. We propose *Attention-Aware Remasking* that can successfully recalibrate the token selection mechanism to prioritize actual visual evidence without requiring any additional training. Combined with the proposed Clue-Based Answering to prevent post-hoc rationalization and Echo-Filling for lossless inference acceleration, our framework demonstrated highly effective visual grounding and semantic explanation quality, outperforming prior baselines in BERTScore and VQA-X accuracy. Our results demonstrate that anchoring the generation process to visual attention is an effective approach to resolving hallucinations and promoting multimodal grounding within dMLLMs.

Limitations

A limitation of this work is that the specific hyperparameters utilized in our framework remain to be fully optimized across diverse settings. To address this, we provide detailed sensitivity analyses for some hyperparameters in Appendix A.4. Crucially, our findings demonstrate that even with these non-exhaustive parameters, our approach consistently surpasses the performance of the de facto standard, low-confidence remasking. This underscores our core argument: that the current industry-standard remasking protocols are not necessarily the most effective, highlighting substantial potential for further refinement within diffusion language models

Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(**RS-2025-24535423**) and the Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT)(**RS-2025-25422680**: Metacognitive AGI Framework and its Applications, and **No. RS-2026-25548925**).

References

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.
- Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer.
- Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.
- Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. 2022. Maskgit: Masked generative image transformer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Kai-Po Chang, Chi-Pin Huang, Wei-Yuan Cheng, Fu-En Yang, Chien-Yi Wang, Yung-Hsuan Lai, and Yu-Chiang Frank Wang. 2024. **RAPPER: Reinforced rationale-prompted paradigm for natural language explanation in visual question answering**. In *The Twelfth International Conference on Learning Representations*.
- Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer.
- Xuanzhao Dong, Wenhui Zhu, Xiwen Chen, Zhipeng Wang, Peijie Qiu, Shao Tang, Xin Li, and Yalin Wang. 2025. Llada-medv: Exploring large language diffusion models for biomedical image understanding. *arXiv preprint arXiv:2508.01617*.
- Sri H Dumpala, Aman Jaiswal, Chandramouli Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. 2024. Sugarcreeper++ dataset: Vision-language model sensitivity to semantic and lexical alterations. *Advances in Neural Information Processing Systems*, 37:17972–18018.
- Jiaxin Ge, Sanjay Subramanian, Trevor Darrell, and Boyi Li. 2023. From wrong to right: A recursive approach towards vision-language explanation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1185.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. **LoRA: Low-rank adaptation of large language models**. In *International Conference on Learning Representations*.
- Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. E-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1244–1254.
- Chengen Lai, Shengli Song, Shiqi Meng, Jingyang Li, Sitong Yan, and Guangneng Hu. 2024. Towards more faithful natural language explanation using multi-level contrastive learning in vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2849–2857.
- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.
- Shufan Li, Konstantinos Kallidromitis, Hritik Bansal, Akash Gokul, Yusuke Kato, Kazuki Kozuka, Jason Kuen, Zhe Lin, Kai-Wei Chang, and Aditya Grover. 2025a. Lavidia: A large diffusion model for vision-language understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Tianyi Li, Mingda Chen, Bowei Guo, and Zhiqiang Shen. 2025b. A survey on diffusion language models. *arXiv preprint arXiv:2508.10875*.

- Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. 2022. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35:4328–4343.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Zhenghao Lin, Yeyun Gong, Yelong Shen, Tong Wu, Zhihao Fan, Chen Lin, Nan Duan, and Weizhu Chen. 2023. Text generation with diffusion language models: A pre-training approach with continuous paragraph denoise. In *International Conference on Machine Learning*, pages 21051–21064. PMLR.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.
- Xinyin Ma, Runpeng Yu, Gongfan Fang, and Xinchao Wang. 2026. dkv-cache: The cache for diffusion language models. *Advances in Neural Information Processing Systems*, 38:149009–149033.
- Vishal Narnaware, Animesh Gupta, Kevin Zhai, Zhenyi Wang, and Mubarak Shah. 2026. Seeing to ground: Visual attention for hallucination-resilient mllms. *arXiv preprint arXiv:2603.25711*.
- Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, JUN ZHOU, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. [Large language diffusion models](#). In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Jingyang Ou, Shen Nie, Kaiwen Xue, Fengqi Zhu, Jiacheng Sun, Zhenguo Li, and Chongxuan Li. 2025. [Your absorbing discrete diffusion secretly models the conditional distributions of clean data](#). In *The Thirteenth International Conference on Learning Representations*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Dong Park, Lisa Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. [Multimodal explanations: Justifying decisions and pointing to the evidence](#). pages 8779–8788.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 4035–4045.
- Subham S Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. 2024. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184.
- Alireza Salemi, Julian Killingback, and Hamed Zamani. 2025. Expert: Effective and explainable evaluation of personalized long-form text generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17516–17532.
- Fawaz Sammani, Tanmoy Mukherjee, and Nikos Deligiannis. 2022. Nlx-gpt: A model for natural language explanations in vision and vision-language tasks. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8322–8332.
- Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. 2024. [Simplified and generalized masked diffusion for discrete data](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Wei Suo, Mengyang Sun, Weisong Liu, Yiqi Gao, Peng Wang, Yanning Zhang, and Qi Wu. 2023. S3c: Semi-supervised vqa natural language explanation via self-critical learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2646–2656.
- Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.
- Xiyao Wang, Jiuhai Chen, Zhaoyang Wang, Yuhang Zhou, Yiyang Zhou, Huaxiu Yao, Tianyi Zhou, Tom Goldstein, Parminder Bhatia, Taha Kass-Hout, et al. 2025. Enhancing visual-language modality alignment in large vision language models via self-improvement. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 268–282.
- Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. 2025. Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding. *arXiv preprint arXiv:2505.22618*.
- Jiayuan Xie, Yi Cai, Jiali Chen, Ruohang Xu, Jiexin Wang, and Qing Li. 2024. Knowledge-augmented visual question answering with natural language explanation. *IEEE Transactions on Image Processing*, 33:2652–2664.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu,

Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*.

Zebin You, Shen Nie, Xiaolu Zhang, JUN ZHOU, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. 2026. Llada-v: Large language diffusion models with visual instruction tuning. *CVPR*.

Fan Yuan, Xiaoyuan Fang, Rong Quan, Jing Li, Wei Bi, Xiaogang Xu, and Piji Li. 2025. Generative visual commonsense answering and explaining with generative scene graph constructing. *arXiv preprint arXiv:2501.09041*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Haozhe Zhao, Shuzheng Si, Liang Chen, Yichi Zhang, Maosong Sun, Baobao Chang, and Minjia Zhang. 2025a. Looking beyond text: Reducing language bias in large vision-language models via multimodal dual-attention and soft-image guidance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19677–19701.

Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. 2025b. Mitigating object hallucination in large vision-language models via image-grounded guidance. In *Forty-second International Conference on Machine Learning*.

Jiaming Zhou, Hongjie Chen, Shiwan Zhao, Jian Kang, Jie Li, Enzhi Wang, Yujie Guo, Haoqin Sun, Hui Wang, Aobo Kong, et al. 2026. Diffa: Large language diffusion models can listen and understand. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 35112–35120.

Fengqi Zhu, Zebin You, Yipeng Xing, Zenan Huang, Lin Liu, Yihong Zhuang, Guoshan Lu, Kangyu Wang, Xudong Wang, Lanning Wei, et al. 2025. Llada-moe: A sparse moe diffusion language model. *arXiv preprint arXiv:2509.24389*.

A Appendix

A.1 Training Details

We initialize our model with the publicly available weights of LLaDA-V and perform Supervised Fine-Tuning (SFT) tailored for the target Natural Language Explanation (NLE) task. To ensure computational efficiency during training, we employ Low-Rank Adaptation (LoRA) as proposed by Hu

Backbone	VQA-X		e-SNLI-VE	
	BS	Acc	BS	Acc
LLaDA-V	93.1	90.8	92.4	76.1
LaViDa	87.7	91.8	84.5	73.0

Table 6: Performance of our framework using different backbone models (LLaDA-V vs. LaViDa).

Remasking Strategy	CHAIR _s ↓	CHAIR _i ↓	Recall ↑
Low-confidence	9.5	9.8	38.1
Ours	6.0	5.5	40.6

Table 7: Object hallucination evaluation using CHAIR metrics on the LaViDa backbone without fine-tuning.

et al. (2022). Specifically, we keep the vision tower frozen to preserve its pre-established visual representation capabilities and apply LoRA only to the language model backbone. For the optimization process, we utilize the 8-bit AdamW optimizer to further reduce the memory footprint. The training is conducted with a peak learning rate of 1×10^{-5} , governed by a cosine annealing learning rate schedule. To prevent overfitting while ensuring adequate adaptation to the target domain, the model is fine-tuned for only a single epoch on each respective dataset.

A.2 Inference and Decoding Configurations

Since our proposed Attention-Aware Remasking, Clue-Based Answering, and Echo-Filling are strictly training-free techniques, we configure several key inference hyperparameters to govern decoding stability and budget allocation:

- **Attention Aggregation Parameters (n, m):** We set $n = 16$ and $m = 4$. These parameters are chosen heuristically rather than strictly tuned; we selected moderately large values to prevent instability arising from overly narrow attention aggregation.
- **Number of Denoising Steps:** Set to L , where L denotes the total generation length. (While the default LLaDA-V implementation uses $L/2$, generating two tokens per step, to balance speed and performance, we set the number of steps equal to L to generate one token per step. Since diffusion language models can suffer from instability when decoding multiple tokens simultaneously, and the VL-NLE task features relatively short sentences where generation quality is the strict priority, we

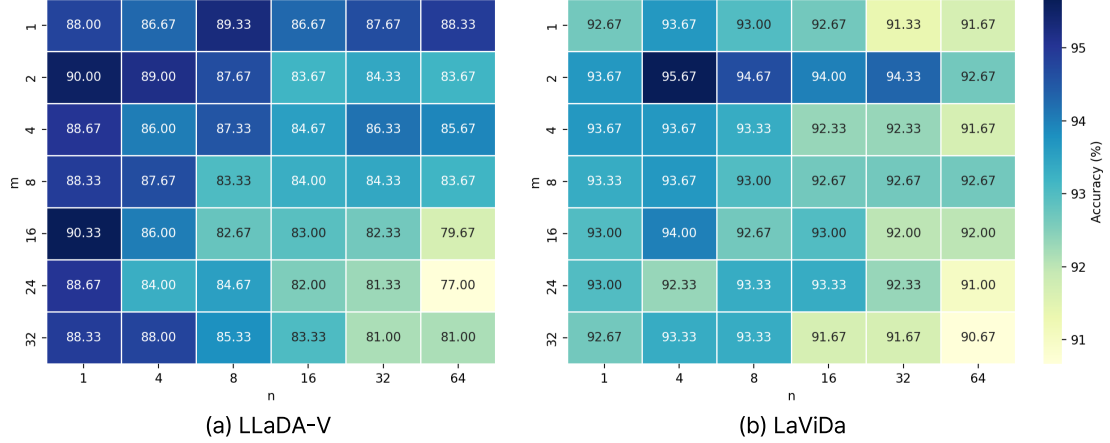


Figure 6: Sensitivity analysis of hyperparameters n and m presented as accuracy heatmaps on the VQA-X subset for (a) LLaDA-V (left) and (b) LaViDa (right).

Backbone \ Layer	0	3	7	11	15	19	23	27	31
LLaDA-V	0.70	0.85	0.77	0.78	0.81	0.86	0.87	0.87	0.89
LaViDa	0.87	0.84	0.91	0.84	0.88	0.89	0.86	0.92	0.95

Table 8: Task accuracy across different transformer layers for visual attention extraction ($n = 4, m = 2$).

adopted this conservative schedule to maximize stability and text quality.)

- **Maximum Generation Length:** Fixed at 128 tokens, which is determined based on the empirical length distribution of target explanation labels in the benchmark datasets.
- **Block Budget Allocation:** We allocate 90% of the total generation length to the explanation block and the remaining 10% to the answer block. This ensures sufficient budget for detailed reasoning prior to final answer commitment.

A.3 Generalizability Across Different Backbones

A key advantage of our proposed strategies is their training-free nature. We herein conduct additional experiments to evaluate the generalizability of our framework by comparing our default LLaDA-V backbone with LaViDa (Li et al., 2025a).

VL-NLE Task Performance. The overall task performance using different backbones is reported in Table 6. While our framework successfully operates on the alternative backbone, there is an observable performance variance. Specifically, although LaViDa achieves a higher VQA accuracy on the VQA-X dataset, its metrics on the e-SNLI-VE

dataset are lower compared to those of LLaDA-V. We attribute this performance discrepancy primarily to the fact that the LaViDa backbone was evaluated under the exact same hyperparameter configurations originally tuned for LLaDA-V, without undergoing model-specific optimization. Nevertheless, the successful deployment of our training-free inference strategies on a distinct diffusion language backbone, without requiring any structural modifications, demonstrates their architecture-agnostic nature.

Object Hallucination Evaluation across Backbones. To further verify the generalizability of our Attention-Aware Remasking in improving visual fidelity, we evaluate object hallucination on image captioning outputs without fine-tuning. The results for our default LLaDA-V backbone are presented in Table 3 (Section 6.1.3). To demonstrate generalizability, we additionally evaluate LaViDa in Table 7. As shown, substituting conventional low-confidence remasking with our Attention-Aware Remasking on LaViDa similarly achieves substantial reductions in both sentence-level (CHAIR_s) and instance-level (CHAIR_i) object hallucinations while improving recall. These consistent improvements across distinct backbones confirm that our visual-confidence-guided decoding strategy effectively enhances visual fidelity and

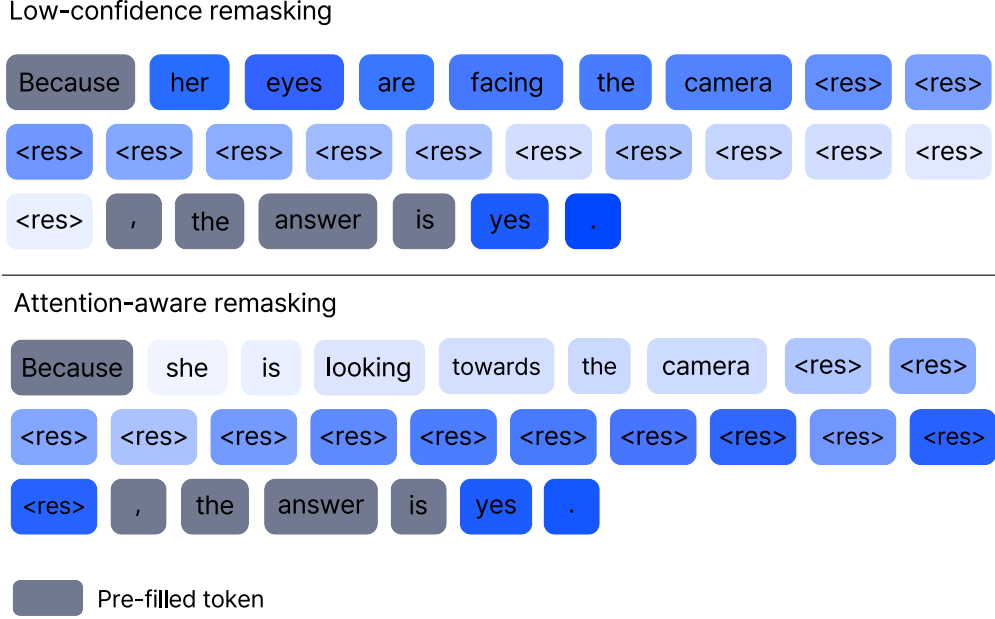


Figure 7: Comparison of token generation trajectories. The visualization tracks the order of token unmasking for low-confidence remasking versus attention-aware remasking. The transition from light to dark colors signifies the progression from early to late generation steps. Echo-filling is deactivated to facilitate the precise observation of generation priorities.



Figure 8: Reasoning failure cases where fine-grained semantic relationships (spatial relations and action predicates) are incorrectly matched.

mitigates hallucinations regardless of the underlying dMLLM architecture.

A.4 Sensitivity Analysis on Hyperparameters

To investigate the robustness of our framework and validate our design choices across different backbones, we present sensitivity analyses regarding the attention aggregation parameters (n, m) and the transformer layer selection for visual confidence extraction. The evaluations are conducted on a 300 randomly sampled subset of the VQA-X dataset using both base models, LLaDA-V and LaViDa.

Sensitivity to n and m . Figure 6 illustrates the accuracy heatmaps across varying values of n and m for both LLaDA-V and LaViDa. We observe

that overly broad aggregation tends to dilute fine-grained visual signals, leading to a minor performance decline. Crucially, across both backbones, the optimal values of n and m reside within a very similar parameter regime. This suggests that extensive, backbone-specific hyperparameter retuning is generally unnecessary when applying our proposed inference-stage framework to different dMLLM backbones.

Sensitivity to Attention Layer Choice. Table 8 displays the task accuracy when extracting attention maps from different transformer layers ($l \in \{0, 3, 7, \dots, 31\}$) with $n = 4, m = 2$. For both LLaDA-V and LaViDa, we observe a general upward trend in accuracy as the layer depth increases, with the final layer ($l = 31$) achieving the highest accuracy. This confirms that deeper layers capture richer, semantically aligned visual-textual cross-attentions, validating our choice of utilizing final-layer attention for computing visual confidence.

A.5 Analysis of Token Generation Trajectories

Through auxiliary experiments, we identified that the choice of remasking strategy significantly influences the temporal order in which tokens are generated during inference. Figure 7 illustrates this



Figure 9: More examples of model results.

phenomenon by tracking the token generation trajectories for both the traditional low-confidence remasking approach and our proposed attention-aware remasking strategy.

When employing low-confidence remasking, the model exhibits a pronounced tendency to prioritize the generation of padding tokens, such as [reserved token], in the early stages of inference. This behavior is primarily attributed to an inherent frequency bias established during the training phase, where padding tokens are disproportionately prevalent. Because the model encounters these tokens most frequently, it tends to assign them high logit values (and thus high confidence) by default. Consequently, the generation process often begins by filling in non-informative structural elements before addressing the substantive content of the explanation, reflecting a reliance on linguistic statistics rather than visual evidence.

In contrast, our proposed strategy leads the model to prioritize the generation of content-rich textual tokens that carry substantive information. Because attention-aware remasking shifts the selection criterion from raw logit probabilities to the degree of visual interaction with the image, padding tokens that lack meaningful correspondence with visual evidence are naturally deprioritized.

This approach effectively allows the model to bypass linguistic bias, ensuring that the reasoning process is grounded in visual evidence from the onset and resulting in explanations that are more faithful to the actual visual context.

A.6 Case Studies

To further illustrate the effectiveness of our proposed Attention-Aware Remasking strategy, we

present additional case studies in Figure 9. Specifically, these qualitative examples compare the generations produced by our visual confidence-based approach against those generated using the default logit-based remasking strategy (i.e., Low-confidence Remasking). As demonstrated in the examples, the low-confidence remasking occasionally misses critical visual details.

While Attention-Aware Remasking effectively anchors token generation to relevant visual regions, **failure cases** often arose where fine-grained semantic relationships were incorrectly matched. Looking at Figure 8, the primary reasoning failures stem from *spatial and contextual relation errors*, where the model correctly identifies visual entities but hallucinates incorrect spatial arrangements, and *fine-grained action misinterpretations*, where the target object is successfully grounded but assigned an inaccurate action predicate.