

Fixing Semantic Blind Spots in Anchor Tokens of dMLLMs

Ruixuan Xu¹, Jiexi Xu², Qiyao Zhao³, Xiaofeng Zhang^{3†}

¹The Chinese University of Hong Kong

²The University of Hong Kong

³Department of Automation and Intelligent Sensing, Shanghai Jiao Tong University

{1155211213@}link.cuhk.edu.hk

Abstract

Recent advances in diffusion-based Multimodal Large Language Models (dMLLMs) offer a compelling alternative to autoregressive counterparts; however, they remain prone to hallucinations. Through information flow analysis on representative dMLLMs (e.g., LLaDA-V and LaViDa), we identify two intertwined factors contributing to this issue. First, although the special *image_newline* tokens serve as semantic anchors for aggregating visual information, they simultaneously induce severe attention sinks, excessively consuming the model’s attention budget. Second, the long-range decay inherent in Rotary Position Embedding (RoPE) leads to semantic blind spots, preventing these anchors from uniformly perceiving the entire visual input. Accordingly, our objective is to moderately alleviate the attention sink effect on semantic anchors while enhancing their ability to aggregate global visual information, thereby eliminating semantic blind spots. To this end, we propose Extrinsic Distance-Aware Regularization (EDAR), a training-free decoding strategy that augments the attention key space with a static, distance-aware matrix. This matrix jointly redistributes excessive attention away from anchors and injects absolute positional bias to ensure uniform visual coverage. Experiments on LLaDA-V and LaViDa demonstrate that EDAR effectively eliminates semantic blind spots and achieves state-of-the-art performance on both hallucination-specific and general multimodal benchmarks.

1 Introduction

Autoregressive models (ARMs) (Liu et al., 2023; Bai et al., 2025b; Chen et al., 2025c) have long dominated the landscape of Multimodal Large Language Models (MLLMs). However, their next-token prediction paradigm enforces strictly sequential text generation, resulting in substantial inference latency and the well-known reversal curse.

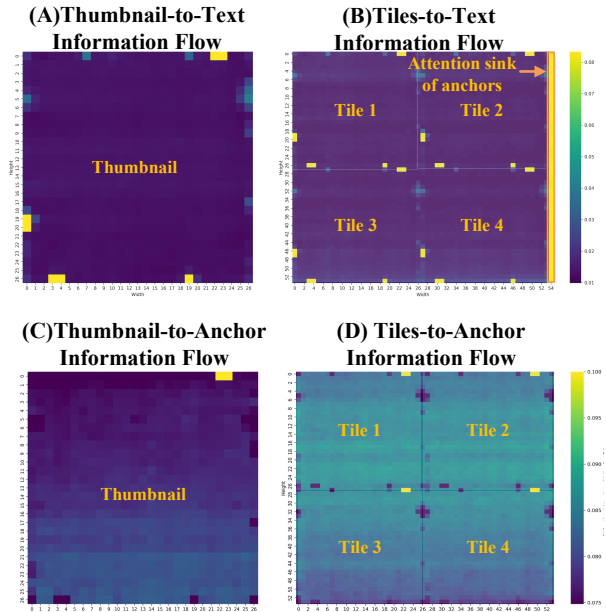


Figure 1: **Information Flow Analysis of LLaDA-V.** From (a) and (b), the *image_newline* tokens emerge as anchors, absorbing attention and compressing visual representations. In (c) and (d), the anchors show a pronounced neglect of the thumbnail, in contrast to its uniform attention to the tile patches.

Recently, an alternative paradigm has emerged in the form of diffusion-based Multimodal Large Language Models (dMLLMs) (You et al., 2025; Yang et al., 2025; Li et al., 2025c). By generating text via parallel iterative masked denoising, dMLLMs depart from rigid token-by-token decoding. This paradigm offers inherent advantages in parallel decoding and more flexible reasoning behaviors, positioning dMLLMs as a promising and forward-looking direction for multimodal generation.

Despite their remarkable versatility, dMLLMs still suffer from hallucinations, where models generate content that contradicts the visual input. Prior works have leveraged information-flow analysis, an interpretability approach for black-box models, to explore the relationship between internal model mechanisms and hallucinations, and have proposed

a series of mitigation strategies for ARMs (Xing et al., 2024a; Huang et al., 2024; Zhang et al., 2024, 2025, 2026a; Zhao et al., 2026; Zhang et al., 2026b; Wei and Zhang, 2024). However, these methods fundamentally rely on the unidirectional, causal nature of ARMs and are not transferable to dMLLMs, which operate under a bidirectional, non-causal diffusion paradigm. Consequently, the internal mechanisms driving hallucinations in dMLLMs remain poorly understood—and existing ARMs-based interventions are ineffective in this setting. To this end, we conduct an in-depth investigation of information flow in LLaDA-V (You et al., 2025), and further verify that the same phenomena hold on LaViDa (Li et al., 2025c) (see Appendix A). As illustrated in Figure 1, our analysis is organized around two complementary perspectives:

(1) The information flow from visual tokens to textual tokens. Prior work on ARMs, such as CCA (Xing et al., 2024a) and MCA (Zhao et al., 2025), has analyzed visual-to-text information flow and revealed the adverse effects of RoPE’s long-term decay on multimodal alignment and hallucinations. For dMLLMs, taking LLaDA-V (You et al., 2025) as an example, the input image is processed under the AnyRes scheme (Li et al., 2024; Chen et al., 2025c) into one thumbnail patch and multiple high-resolution tile patches, which are encoded into visual tokens and concatenated with special *image_newline* tokens. As shown in Figure 1 (a) and (b), we analyze the thumbnail-to-text and tiles-to-text information flow to study cross-modal interactions. We observe that text tokens attend only sparsely to visual tokens, while the *image_newline* tokens receive disproportionately high attention, acting as semantic anchors and exhibiting a pronounced attention sink. This behavior leaves the model with only a very limited attention budget for image tokens.

(2) The information flow between anchor token and visual tokens. Anchor tokens associated with attention sink typically serve as critical information hubs, compressing the dense visual and textual semantics into a compact representation to guide the generation process (Tang et al., 2025a; Huang et al., 2024). Furthermore, in Figure 1 (c) and (d), we respectively illustrate the information flow from thumbnail and tile visual tokens to the anchor tokens to examine the comprehensiveness of the semantic aggregation. For the bidirectional attention in dMLLMs, attention computation is likewise

constrained by RoPE’s long-term decay. Consequently, the anchor tokens tend to allocate more attention to visual tokens appearing later in the flattened sequence, as they are closer in relative positional distance. This results in sparse attention to earlier thumbnail tokens, a phenomenon we refer to as "semantic blind spots." This phenomenon limits the anchor tokens’ perception of global image information.

Summary: Attention sink reduces the attention budget for image tokens, while RoPE induces sparse anchor tokens’ attention to the thumbnail. Together, these factors lead to insufficient image perception and hallucinations.

To this end, we propose Extrinsic Distance-Aware Regularization (EDAR), a simple and effective training-free decoding strategy. EDAR directly calibrates the attention mechanism, laterally concatenating a static, distance-aware regularization matrix to the key space of the attention module. This extrinsic matrix serves two purposes: first, it acts as a "probability sink," dynamically offloading excess attention from over-focused attention sink tokens, thereby increasing the relative attention budget for tokens at longer relative distances; second, through a progressive decay factor, it embeds absolute positional information, compelling the model to maintain an ordered and uniform information flow from the semantic anchors to the entire visual canvas. This design effectively improves the information flow allocation of dMLLMs. Our contributions can be summarized as follows:

- Our analysis of the information flow reveals a two-fold bottleneck leading to hallucinations, stemming from the attention sink in *image_newline* tokens and the "semantic blind spots" caused by its failure to aggregate global image information.
- We propose EDAR, a novel and training-free method that rectifies attention distribution by augmenting the key space with an extrinsic regularization matrix, effectively improving the information flow distribution of dMLLMs.
- Through experiments on LLaDA-V and LaViDa models, we demonstrate that EDAR effectively eliminates "semantic blind spots" and achieves great performance on multiple benchmarks.

2 Related Work

2.1 Diffusion Multimodal Large Language Models

The latest diffusion-based large language models (dLLMs) (Nie et al., 2025; Zhu et al., 2025; Ye et al., 2025; Gong et al., 2025) have been successfully scaled to 8B parameters, achieving performance comparable to state-of-the-art autoregressive large language models (Devlin et al., 2018; Grattafiori et al., 2024; DeepSeek-AI et al., 2024; Team, 2024). Furthermore, by integrating visual instruction tuning and architectural extensions, dLLMs have recently been extended to diffusion Multimodal Large Language Models (dM-LLMs) (Yang et al., 2025; Li et al., 2025c; You et al., 2025; Yu et al., 2025), demonstrating promising multimodal capabilities.

2.2 Hallucination

The term "hallucination" (Yu et al., 2024; Sarkar et al., 2024; Xiao et al., 2024; Xing et al., 2024b; Ma et al., 2024; Gong et al., 2024; Chen et al., 2024a; Kim et al., 2024; Liu et al., 2024; Zhou et al., 2023; Zhai et al., 2023; Wang et al., 2023; Huang et al., 2023; Zhu et al., 2024; Jiang et al., 2024; Zhou et al., 2025; Bai et al., 2025a; Suo et al., 2025; Lymperaiou et al., 2025; Wang et al., 2025; Li et al., 2025b; Chen et al., 2025b; Che et al., 2025; Chen et al., 2025a; Tu et al., 2025; Mao et al., 2025; Duan et al., 2025; Yin et al., 2025; Li et al., 2025d; Bae et al., 2025; Li et al., 2025a; Xie et al., 2025a; Zheng et al., 2024; Kan et al., 2024; Zhao et al., 2024; Wang et al., 2024a,b; He et al., 2024; Yang et al., 2024; Gu et al., 2024) refers to the fact that multimodal models, when processing multimodal inputs, sometimes produce content that does not correspond to the actual inputs or is even fictitious. Among them, RLHF (Ziegler et al., 2020; Ouyang et al., 2022; Lee et al., 2024) is an approach that relies on human feedback reinforcement learning techniques, which manually evaluates and guides model outputs, prompting the model to pay more attention to factual basis and logical consistency.

2.3 RoPE decay methods

Recent works have identified RoPE’s *long-term decay*—i.e., exponentially diminishing attention between tokens at large relative distances—as a root cause of visual neglect and object hallucination in LVLMs (Xing et al., 2024a; Zhao et al., 2025; Liu et al., 2025b; Yip et al., 2026a; Balazević et al.,

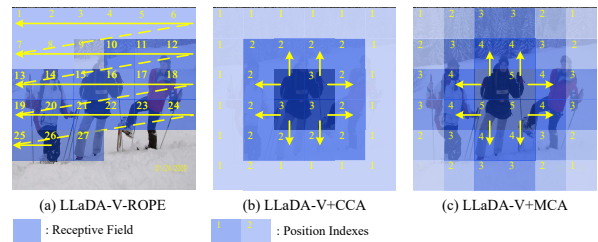


Figure 2: RoPE decay and previous methods of reindexing position ids.

2024; Yip et al., 2026b). To counteract this, several position-encoding interventions have been proposed, which we categorize along two orthogonal axes: (1) re-indexing vs. architecture augmentation, and (2) training-required vs. training-free.

First, CCA-LLaVA (Xing et al., 2024a) introduces concentric causal attention, reassigning image-patch positions in a ring-wise inward order so that instruction tokens are closer (in 1D sequence distance) to more visual tokens. In parallel, MCA-LLaVA (Zhao et al., 2025) and DAPE (Xie et al., 2025b) propose a Manhattan-based reordering: by assigning 2D coordinates from all four image corners and summing them to form new position indices $\mu(m) = x_m + y_m$, the 1D unidirectional decay of is extended to a 2D multidirectional spatial decay (see Figure 2). Farsight (Tang et al., 2025b) proposes to use causal masks to establish information propagation between multimodal words.

Second, approaches like ComeMo (Liu et al., 2025a) and VRoPE (Liu et al., 2025b) avoid sequence reordering and instead augment the architecture. ComeMo adopts a dual-path design: a standard autoregressive context path and a cross-attention memory path (with position-aware in both query and key), thereby bypassing decay in the causal chain.

3 Motivation

3.1 Preliminary

AnyRes High-Resolution Processing. To accommodate high-resolution inputs while balancing computational efficiency, LLaDA-V employs the AnyRes (Li et al., 2024; Chen et al., 2025c) strategy. The input image is processed through two parallel streams to capture both holistic structure and fine-grained details. For the thumbnail, which provides the global view, the entire original image is directly resized to a fixed resolution and encoded into visual

tokens. Simultaneously, for the localized tiles, the original image is cropped into fixed-size patches. Each patch is then resized to the exact same resolution as the thumbnail and encoded independently. These local feature maps are subsequently concatenated to form a high-resolution representation. Finally, the special separator *image_newline* tokens are inserted at the end of each row of visual tokens.

3.2 Core Motivation

To deconstruct the internal mechanisms of information propagation within dMLLMs, we conducted a comprehensive empirical analysis on the MSCOCO (Lin et al., 2015) dataset. We tracked the attention dynamics across 1000 samples during the inference process. Our investigation reveals fundamental misalignments in how semantic information is anchored and distributed.

Attention Concentration on Semantic Anchors.

We first quantify the visual-to-text information flow, formally defined as the aggregate attention a textual token receives from the visual context. Let \mathcal{T} denote the set of visual tokens and $A_{i,j}$ represent the attention weight from a query token i to a key token j . For a given text token j , the flow $I_{\text{vis} \rightarrow \text{txt}}^{(j)}$ is calculated as:

$$I_{\text{vis} \rightarrow \text{txt}}^{(j)} = \sum_{i \in \mathcal{T}} A_{i,j},$$

where the summation aggregates contributions from the entire visual sequence. Contrary to the expectation that textual tokens would attend uniformly across visual features, we observe a distinct attention collapse. As visualized in Figure 1 (a) and (b), the direct attention from text to standard image content tokens is negligible. Instead, the information flow is disproportionately aggregated by *image_newline* tokens. This phenomenon suggests that *image_newline* tokens function as critical semantic anchors, compressing and encapsulating visual and textual semantics into a compact representation.

Bidirectional Visual-Semantic Interaction and Semantic Blind Spots. Identifying *image_newline* tokens as the primary carriers of semantic information naturally prompts an investigation into how this compressed information is broadcast to the rest of the image. Consequently, we examine the bidirectional information flow between these carrier tokens and the standard

image content tokens, leveraging the bidirectional attention characteristics of LLaDA-V (You et al., 2025). Let \mathcal{N} be the *image_newline* tokens and \mathcal{V}_{img} be the set of standard image content tokens. We define the interaction flow $I_{\text{bi}}^{(j)}$ for a content token $j \in \mathcal{V}_{\text{img}}$ as the sum of mutual attention between itself and *image_newline* tokens:

$$I_{\text{bi}}^{(j)} = \sum_{k \in \mathcal{N}} (A_{k,j} + A_{j,k}).$$

This bidirectional interaction is crucial for unpacking the semantic information stored in the anchors back into the spatial features. Ideally, the information flow should have a uniform distribution on all visual tokens. However, our analysis uncovers a severe long-term decay within the thumbnail map, forming the "semantic blind spots." As illustrated in Figure 1 (c), tokens located earlier in the flattened sequence of the thumbnail receive significantly lower bidirectional attention scores compared to those appearing later.

This finding highlights a fundamental misalignment: while *image_newline* tokens function as semantic hubs, successfully compressing dense semantics, the underlying attention mechanism fails to aggregate the information uniformly from visual tokens. The observed semantic blind spots will compromise the model’s ability to maintain global consistency.

4 Methodology

Building on the above analysis, we propose Extrinsic Distance-Aware Regularization (EDAR). EDAR addresses the issue by laterally augmenting the key space, injecting an extrinsic regularization matrix to calibrate the attention distribution explicitly, forcing the model to maintain an ordered information flow across long sequences.

4.1 Extrinsic Key Augmentation

In the standard dMLLM architecture, attention logits $\mathbf{L} \in \mathbb{R}^{N \times N}$ are computed within a closed system of queries and keys. EDAR extends the key space by laterally concatenating a static regularization matrix $\mathbf{R} \in \mathbb{R}^{N \times M}$. Here, N denotes the sequence length and M represents the capacity of the extrinsic buffer. The augmented attention operation transforms the standard Softmax into a gated probability mechanism:

$$\mathbf{A}_{\text{aug}} = \text{Softmax}([\mathbf{L} \parallel \mathbf{R}]), \quad (1)$$

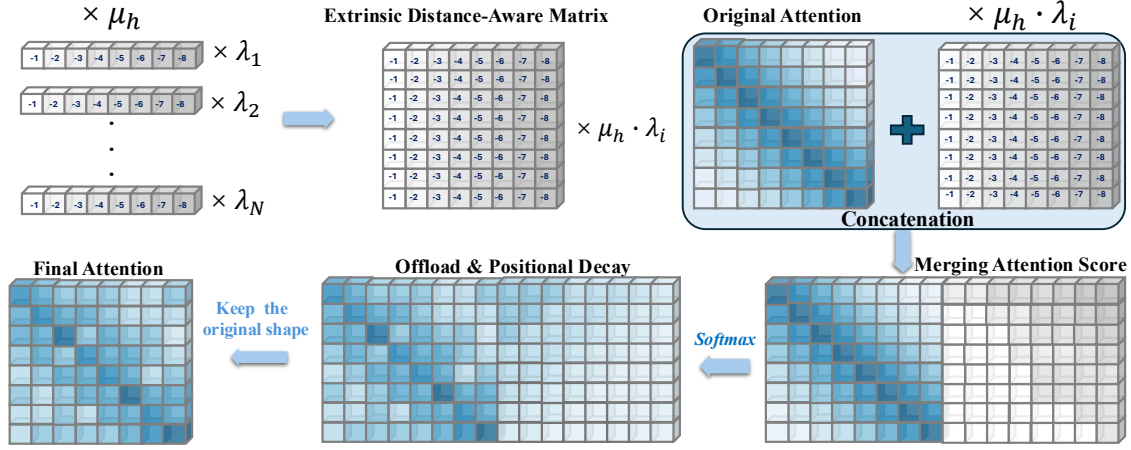


Figure 3: The structure of our method.

where \parallel denotes concatenation along the key dimension. The residual probability mass is offloaded to the extrinsic keys defined by \mathbf{R} , which acts as a dynamic probability sink for redundant attention.

The regularization matrix \mathbf{R} is not learned as parameters but is a distance-aware static coefficient matrix. The matrix is designed as a structured field where each row corresponds to a query position i and each column to an extrinsic register k . For the h -th attention head, the logit value at position (i, k) is formulated as:

$$R_{i,k}^{(h)} = -\mu_h \cdot (k + 1) \cdot \lambda_i, \quad (2)$$

for $0 \leq i < N$, $0 \leq k < M$.

This formulation integrates three structural components. First, the term k serves as a linear index to ensure the regularization field maintains full rank, creating a gradient across the extrinsic dimension rather than a uniform value. Second, following the ALiBi (Press et al., 2022) paradigm, the head-specific coefficient μ_h applies a geometric scaling across heads to capture multi-scale dependencies, ensuring that each head operates at a distinct positional decay rate. Finally, we introduce a progressive decay rectification factor λ_i .

$$\lambda_i = \alpha + \frac{i}{N-1}(\beta - \alpha). \quad (3)$$

We empirically set the bounds as $\alpha = 0.01$ and $\beta = 10.0$. By modulating the penalty magnitude relative to the token index i , this design creates a position-dependent potential field that dynamically adjusts the attractiveness of the extrinsic offload.

The resulting augmented logit row for a query at position i in head h , after concatenation, takes the form:

$$\mathbf{a}_i^{(h)} = \left[\mathbf{q}_i \mathbf{K}^\top, -\mu_h \lambda_i, -2\mu_h \lambda_i, \dots, -M\mu_h \lambda_i \right]. \quad (4)$$

This row-wise progression ensures the matrix embeds positional hierarchies without altering the intrinsic logits.

The calibrated attention weights for the sequence are thereby allowed to sum to less than unity. After computing the augmented attention weights, the final calibrated attention matrix is obtained by slicing to retain only the intrinsic dimensions, thereby preserving the original matrix shape:

$$\mathbf{A}_{\text{calib}} = \mathbf{A}_{\text{aug}}[:, :N]. \quad (5)$$

4.2 Effectiveness Analysis

EDAR incorporates absolute positional information to maintain an ordered information flow across positions. We analyze this through the partition function of the Softmax operation. For a given query i , the total energy Z_i is the sum of the visual component Z_{real} and the regularization component Z_{reg} . Substituting the definition of $R_{i,k}^{(h)}$, the regularization component $Z_{\text{reg}}^{(i)}$ forms a geometric series:

$$Z_{\text{reg}}^{(i)} = \sum_{k=1}^M \exp(R_{i,k}^{(h)}) = \sum_{k=1}^M \exp(-\mu_h \lambda_i \cdot k). \quad (6)$$

Using the closed form of the geometric series, the sum can be expressed as:

$$Z_{\text{reg}}^{(i)} = \frac{\exp(-\mu_h \lambda_i) (1 - \exp(-\mu_h \lambda_i)^M)}{1 - \exp(-\mu_h \lambda_i)}. \quad (7)$$

Since λ_i increases monotonically with the sequence position i from α to β , the term $-\mu_h \lambda_i$ becomes increasingly negative. Consequently, the base term decays exponentially toward zero as $i \rightarrow N$, implying:

$$\frac{\partial Z_{\text{reg}}^{(i)}}{\partial i} < 0. \quad (8)$$

This decay embeds absolute positional information by making the regularization strength inversely proportional to position. Early tokens, with small i and large $Z_{\text{reg}}^{(i)}$, offload more probability mass to the extrinsic sink, suppressing over-reliance on adjacent context and enforcing a baseline order. As i increases, $Z_{\text{reg}}^{(i)}$ diminishes, releasing probability mass back to intrinsic keys and enabling later tokens to draw more heavily from preceding positions.

In essence, the extrinsic matrix acts as a probability sink to offload excessive attention from the anchor, thereby increasing the relative attention budget for other image tokens. A complete elimination of the anchor mechanism would be counterproductive, as it is essential for information aggregation. Therefore, improving the image-to-anchor information flow is crucial. The embedded absolute positional information enforces a more uniform information flow, which mitigates RoPE’s long-term decay and eliminates the "semantic blind spots." It ensures that while semantic anchors retain their vital role as information hubs, they aggregate information across the entire visual canvas more uniformly and comprehensively.

5 Experiments

5.1 Experimental Setup Details

Baselines. Our experiments are conducted on two representative dMLLMs: LLaDA-V (You et al., 2025) and LaViDa (Li et al., 2025c). We evaluate our EDAR method against the original LLaDA-V as the primary baseline. Furthermore, we adapt and implement several existing methods—CCA (Xing et al., 2024a), MCA (Zhao et al., 2025), and Far-sight (Tang et al., 2025b) onto the LLaDA-V architecture for comparative purposes.

Evaluation Benchmarks. We evaluate our method on a diverse set of benchmarks to assess

Algorithm 1 Extrinsic Distance-Aware Regularization (EDAR)

Require: Real attention logits $\mathbf{L} \in \mathbb{R}^{B \times H \times N \times N}$

Ensure: Calibrated attention $\mathbf{A}_{\text{calib}} \in \mathbb{R}^{B \times H \times N \times N}$

- 1: **Initialize:** Extrinsic field $\mathbf{R} \in \mathbb{R}^{1 \times H \times N \times M}$
 - 2: // Step 1: Construct base distance matrix
 - 3: $\mathbf{k} \leftarrow [1, 2, \dots, M]$
 - 4: $\mathbf{v}_{\text{base}} \leftarrow -\mathbf{k}$
 - 5: $\mathbf{R}_{\text{tmp}} \leftarrow \text{Repeat}(\mathbf{v}_{\text{base}}, \text{rows} = N)$
 - 6: // Step 2: Apply progressive scaling along rows
 - 7: **for** $i = 0$ to $N - 1$ **do**
 - 8: $\text{progress} \leftarrow i / \max(N - 1, 1)$
 - 9: $\lambda_i \leftarrow 0.01 + \text{progress} \cdot (10.0 - 0.01)$
 - 10: $\mathbf{R}_{\text{tmp}}[i, :] \leftarrow \mathbf{R}_{\text{tmp}}[i, :] \cdot \lambda_i$
 - 11: **end for**
 - 12: // Step 3: Apply head-specific slopes
 - 13: **for** $h = 0$ to $H - 1$ **do**
 - 14: $\mu_h \leftarrow 2^{-8(h+1)/H}$
 - 15: $\mathbf{R}[0, h, :, :] \leftarrow \mathbf{R}_{\text{tmp}} \cdot \mu_h$
 - 16: **end for**
 - 17: // Step 4: Broadcast and Concatenate
 - 18: $\mathbf{R} \leftarrow \mathbf{R}.\text{expand}(B, -1, -1, -1)$ ▷ Match batch dimension
 - 19: $\mathbf{L}_{\text{aug}} \leftarrow \text{Concat}([\mathbf{L}, \mathbf{R}], \text{dim} = -1)$
 - 20: // Step 5: Compute Attention and Slice
 - 21: $\mathbf{A}_{\text{aug}} \leftarrow \text{Softmax}(\mathbf{L}_{\text{aug}}, \text{dim} = -1)$
 - 22: $\mathbf{A}_{\text{calib}} \leftarrow \mathbf{A}_{\text{aug}}[:, :, :, : N]$ ▷ Retain intrinsic visual attention
 - 23: **return** $\mathbf{A}_{\text{calib}}$
-

its performance on hallucination mitigation and general multimodal capabilities. The benchmarks are divided into two categories: (1) Hallucination Benchmarks: To measure the model’s ability to suppress object-level hallucinations specifically, we use benchmarks POPE (Li et al., 2023b) and CHAIR (Rohrbach et al., 2018). (2) General Multimodal Benchmarks: To verify that our method does not compromise—and ideally improves—general vision-language understanding, we also conduct evaluations on a series of general multimodal benchmarks including ai2d (Kembhavi et al., 2016), ScienceQA (Lu et al., 2022), MMMU (Yue et al., 2024), MMStar (Chen et al., 2024b), SEED-Bench (Li et al., 2023a), and MMB (Yuan et al., 2023).

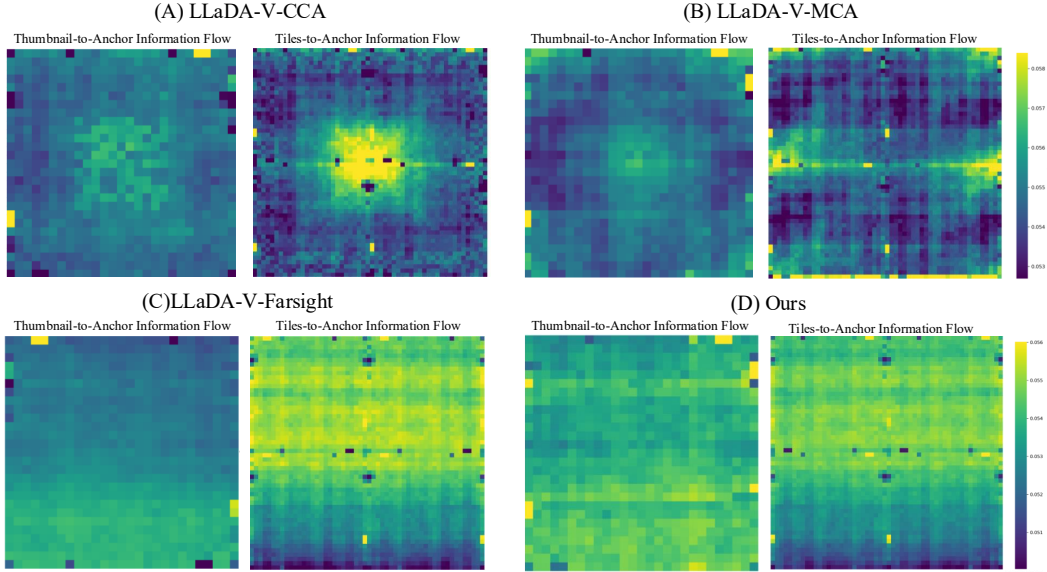


Figure 4: Bidirectional visual-semantic interaction between *image_newline* and standard image content tokens.

Method	Random		Popular		Adversarial		Average	
	Acc \uparrow	F1-score \uparrow	Acc \uparrow	F1-score \uparrow	Acc \uparrow	F1-score \uparrow	Acc \uparrow	F1-score \uparrow
LLaDA-V	88.67	87.44	86.93	85.79	85.27	84.26	86.96	85.83
+CCA	91.27	91.11	87.57	87.80	82.23	83.43	87.02	87.45
+MCA	91.13	90.73	87.33	87.62	82.83	84.30	87.10	87.55
+Farsight	88.87	88.76	87.43	87.72	83.13	83.76	86.48	86.75
+Ours	89.31	89.97	87.60	87.81	84.77	84.36	87.23	87.38

Table 1: POPE Evaluation Results on MSCOCO. Higher is better (\uparrow).

Method	long			short		
	$C_S\downarrow$	$C_I\downarrow$	len	$C_S\downarrow$	$C_I\downarrow$	len
LLaDA-V	30.0	7.6	123.4	17.4	6.8	61.8
+ CCA	23.0	5.9	123.7	17.4	5.4	61.3
+ MCA	25.6	5.8	123.6	15.0	5.2	61.2
+ Farsight	28.0	7.7	123.0	13.6	5.5	61.8
+ Ours	25.4	6.7	124.1	11.2	4.3	61.5

Table 2: Evaluation on CHAIR Hallucination Benchmark. Lower C_S and C_I indicate fewer hallucinations (\downarrow).

5.2 Evaluation Results on Hallucination Benchmarks

In the POPE (Li et al., 2023b) benchmark (Table 1), our method demonstrates highly competitive performance. Compared to the baseline model LLaDA-V, EDAR achieves significant improvements across all three subsets (Random, Popular and Adversar-

ial). Notably, on the more challenging Popular and Adversarial subsets, our method attains the highest scores in both Accuracy and F1-score. This indicates that by enforcing a more uniform attention distribution, EDAR effectively addresses the "semantic blind spot" issue, enabling the model to perceive and confirm objects present in the image more reliably.

In the CHAIR (Rohrbach et al., 2018) benchmark (Table 2), our method performs equally well. For long descriptions, our method significantly outperforms the baseline and Farsight. For short descriptions, our method achieves state-of-the-art performance among all methods on both the CHAIRs and CHAIRi metrics. This proves EDAR can effectively reduce the model’s tendency to invent objects during text generation. Some cases can be seen in Table 6 and Table 7.

Method	General Multimodal Benchmarks							Hallucination Benchmark			
	ai2d \uparrow	SQA \uparrow	MMMU \uparrow	MMStar \uparrow	SEED $_{img}$ \uparrow	SEED $_{vid}$ \uparrow	MMB \uparrow	CHAIR $_S$ \downarrow	CHAIR $_I$ \downarrow	POPE-A \uparrow	POPE-F1 \uparrow
LLaDA-V	77.8	88.89	48.6	60.1	74.8	53.7	82.9	17.4	6.8	86.96	85.83
+ CCA	61.3	79.90	39.33	53.3	75.3	52.9	75.4	17.4	5.4	87.02	87.45
+ MCA	62.7	77.05	40.7	45.98	65.8	49.2	76.2	15.0	5.2	87.10	87.55
+ DoPE	78.5	89.12	51.2	60.5	74.9	53.9	85.5	16.4	6.6	86.85	85.98
+ Farsight	78.1	89.2	48.8	61.2	73.6	53.2	83.2	13.6	5.5	86.48	86.75
+ EDAR	80.6	89.74	49.8	62.2	75.8	54.9	84.5	11.2	4.3	87.23	87.38
LaViDa-D	71.5	83.6	42.6	54.8	69.3	47.5	73.8	18.2	7.1	85.2	84.5
+ EDAR	73.8	85.1	43.5	56.9	70.1	49.6	74.9	14.4	5.8	87.0	86.2
LaViDa-L	73.2	85.4	43.3	56.5	71.2	49.8	70.5	17.5	6.2	85.8	85.2
+ EDAR	75.9	87.2	46.9	58.7	74.8	52.1	75.4	13.5	5.1	87.5	87.1

Table 3: Aggregated Evaluation Results on various multimodal discrete diffusion model variants.

5.3 Evaluation Results of EDAR on General Vision-Language Tasks and Benchmarks

As shown in Table 3, to verify whether our method mitigates hallucinations at the cost of general capabilities, we conducted evaluations on multiple general-purpose vision-language benchmarks.

Our method achieves the best performance across multiple general benchmarks, outperforming the baseline and all other compared methods. This result is highly significant. Unlike methods such as CCA (Xing et al., 2024a) and MCA (Zhao et al., 2025), which reset positional IDs to mitigate hallucinations, these methods lead to noticeable performance degradation in general benchmarks because this drastic alteration of positional IDs damages the image’s structural integrity, despite being effective in specific hallucination metrics. In contrast, EDAR not only excels in hallucination tasks but also comprehensively enhances the model’s general multimodal capabilities.

5.4 Ablation Study

We conduct ablation studies to verify the effectiveness of our semantic blind spot elimination and analyze hyperparameter sensitivity.

Elimination of Semantic Blind Spots. The Original baseline (Fig. 1c) exhibits clear semantic blind spots in the thumbnail. Fig. 4 compares the information flows across different methods. While CCA (Xing et al., 2024a) and MCA (Zhao et al., 2025) partially mitigate the global semantic blind spots, they cause significant disturbance to the Localized tiles. In contrast, both Farsight (Tang et al., 2025b) and our method maintain the stability of

the localized tiles. Crucially, our method achieves a more uniform thumbnail information flow than Farsight, making it the most effective method for eliminating semantic blind spots and preserving the information flow of localized tiles.

Columns (M)	α	β	CHAIR $_S$ \downarrow	CHAIR $_I$ \downarrow
N	0.01	5.0	12.8	5.3
N	0.01	10.0	11.2	4.3
N	0.01	15.0	12.5	4.6
N/2	0.05	5.0	13.4	5.4
N/2	0.05	10.0	13.8	5.3
N/2	0.05	15.0	13.6	5.4

Table 4: Ablation on hyperparameters: columns M and decay bounds α, β . N denotes the sequence token length. The default setting is highlighted.

Ablation Study on Hyperparameters. As shown in Table 4, we have conducted an ablation study on the hyperparameters of columns M and decay bounds α, β across CHAIR (Rohrbach et al., 2018) benchmarks. The analysis reveals two key findings. First, setting the number of columns M to the full sequence length N yields significantly better performance than using a shorter length. Second, the model is sensitive to the decay bounds, with performance peaking at $\beta = 10.0$ and degrading for both lower and higher values. This result suggests that a moderate, well-balanced decay bound is optimal, avoiding values that are either too restrictive or too permissive. Based on these observations, we selected the optimal configuration ($M = N, \alpha = 0.01, \beta = 10.0$), highlighted in the table, as the default for all our experiments.

6 Conclusion

In this paper, we analyzed the internal mechanism in diffusion-based Multimodal Large Language Models. Through our analysis of the internal information flow, we identified attention concentration and semantic blind spots, which contribute to hallucinations. To address this issue, this work introduces Extrinsic Distance-Aware Regularization (EDAR), a novel training-free strategy that effectively reduces hallucinations in dMLLMs by calibrating attention to ensure a uniform information flow. By alleviating key issues like attention concentration and semantic blind spots, the method enhances factual consistency and ensures comprehensive information aggregation from the entire visual canvas. Our method not only significantly reduces hallucinations but also enhances performance on general multimodal benchmarks.

7 Limitations

While EDAR is an effective training-free strategy, its primary limitation is the introduction of computational and memory overhead during inference. This overhead stems from the need to construct an extrinsic regularization matrix and to concatenate it to the key space for each attention computation. This process enlarges the dimensionality of the softmax operation, resulting in increased inference latency and memory consumption. A detailed quantitative analysis of this overhead is provided in Appendix B. Consequently, this efficiency trade-off may hinder its practical application, suggesting a direction for future work in optimizing the regularization process.

References

- Kyungho Bae, Jinhyung Kim, Sihaeng Lee, Soonyoung Lee, Gunhee Lee, and Jinwoo Choi. 2025. Mash-llm: Mitigating action-scene hallucination in video-llms through disentangled spatial-temporal representations. *arXiv preprint arXiv:2503.15871*.
- Jiaqi Bai, Hongcheng Guo, Zhongyuan Peng, Jian Yang, Zhoujun Li, Mohan Li, and Zhihong Tian. 2025a. Mitigating hallucinations in large vision-language models by adaptively constraining information flow. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23442–23450.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Ming-sheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibao Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 2025b. [Qwen3-vl technical report](#). *Preprint*, arXiv:2511.21631.
- Ivana Balažević, Yuge Shi, Pinelopi Papalampidi, Rahma Chaabouni, Skanda Koppula, and Olivier J Hénaff. 2024. Memory consolidation enables long-context video understanding. *arXiv preprint arXiv:2402.05861*.
- Liwei Che, Tony Qingze Liu, Jing Jia, Weiyi Qin, Ruixiang Tang, and Vladimir Pavlovic. 2025. Eazy: Eliminating hallucinations in llms by zeroing out hallucinatory image tokens. *arXiv preprint arXiv:2503.07772*.
- Beitao Chen, Xinyu Lyu, Lianli Gao, Jingkuan Song, and Heng Tao Shen. 2024a. Alleviating hallucinations in large vision-language models through hallucination-induced optimization. *arXiv preprint arXiv:2405.15356*.
- Beitao Chen, Xinyu Lyu, Lianli Gao, Jingkuan Song, and Heng Tao Shen. 2025a. Attention hijackers: Detect and disentangle attention hijacking in llms for hallucination mitigation. *arXiv preprint arXiv:2503.08216*.
- Cong Chen, Mingyu Liu, Chenchen Jing, Yizhou Zhou, Fengyun Rao, Hao Chen, Bo Zhang, and Chunhua Shen. 2025b. Perturbollava: Reducing multimodal hallucinations with perturbative visual training. *arXiv preprint arXiv:2503.06486*.
- Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024b. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.
- Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhui Wang. 2025c. [Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling](#). *Preprint*, arXiv:2412.05271.

- DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, et al. 2024. [Deepseek llm: Scaling open-source language models with longtermism](#). *Preprint*, arXiv:2401.02954.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Jinhao Duan, Fei Kong, Hao Cheng, James Diffenderfer, Bhavya Kaikhura, Lichao Sun, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. 2025. Truthprint: Mitigating lvlm object hallucination via latent truthful-guided pre-intervention. *arXiv preprint arXiv:2503.10602*.
- Shansan Gong, Shivam Agarwal, Yizhe Zhang, Jiacheng Ye, Lin Zheng, Mukai Li, Chenxin An, Peilin Zhao, Wei Bi, Jiawei Han, Hao Peng, and Lingpeng Kong. 2025. Scaling diffusion language models via adaptation from autoregressive models. In *The Thirteenth International Conference on Learning Representations*.
- Xuan Gong, Tianshi Ming, Xinpeng Wang, and Zhihua Wei. 2024. Damro: Dive into the attention mechanism of lvlm to reduce object hallucination. *arXiv preprint arXiv:2410.04514*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Jihao Gu, Yingyao Wang, Meng Cao, Pi Bu, Jun Song, Yancheng He, Shilong Li, and Bo Zheng. 2024. Token preference optimization with self-calibrated visual-anchored rewards for hallucination mitigation. *arXiv preprint arXiv:2412.14487*.
- Jinghan He, Kuan Zhu, Haiyun Guo, Junfeng Fang, Zhenglin Hua, Yuheng Jia, Ming Tang, Tat-Seng Chua, and Jinqiao Wang. 2024. Cracking the code of hallucination in lvlms with vision-aware head divergence. *arXiv preprint arXiv:2412.13949*.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.
- Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.
- Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaying Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046.
- Zhehan Kan, Ce Zhang, Zihan Liao, Yapeng Tian, Wenming Yang, Junyuan Xiao, Xu Li, Dongmei Jiang, Yaowei Wang, and Qingmin Liao. 2024. Catch: Complementary adaptive token-level contrastive decoding to mitigate hallucinations in lvlms. *arXiv preprint arXiv:2411.12713*.
- Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. [A diagram is worth a dozen images](#). *Preprint*, arXiv:1603.07396.
- Junho Kim, Hyunjun Kim, Yeonju Kim, and Yong Man Ro. 2024. Code: Contrasting self-generated description to combat hallucination in large multi-modal models. *arXiv preprint arXiv:2406.01920*.
- Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. [Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback](#). *Preprint*, arXiv:2309.00267.
- Bin Li, Dehong Gao, Yeyuan Wang, Linbo Jin, Shangling Yu, Xiaoyan Cai, and Libin Yang. 2025a. Instruction-aligned visual attention for mitigating hallucinations in large vision-language models. *arXiv preprint arXiv:2503.18556*.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. [Llava-onevision: Easy visual task transfer](#). *Preprint*, arXiv:2408.03326.
- Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.
- Shawn Li, Jiashu Qu, Yuxiao Zhou, Yuehan Qin, Tiankai Yang, and Yue Zhao. 2025b. Treble counterfactual vlms: A causal approach to hallucination. *arXiv preprint arXiv:2503.06169*.
- Shufan Li, Konstantinos Kallidromitis, Hritik Bansal, Akash Gokul, Yusuke Kato, Kazuki Kozuka, Jason Kuen, Zhe Lin, Kai-Wei Chang, and Aditya Grover. 2025c. Lavidia: A large diffusion model for vision-language understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Shuo Li, Jiajun Sun, Guodong Zheng, Xiaoran Fan, Yujiong Shen, Yi Lu, Zhiheng Xi, Yuming Yang, Wenming Tan, Tao Ji, et al. 2025d. Mitigating object hallucinations in mlms via multi-frequency perturbations. *arXiv preprint arXiv:2503.14895*.

- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. [Microsoft coco: Common objects in context](#). *Preprint*, arXiv:1405.0312.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26286–26296.
- Shi Liu, Weijie Su, Xizhou Zhu, Wenhai Wang, and Jifeng Dai. 2025a. Comemo: Lvlms need image context with image memory. *arXiv preprint arXiv:2506.06279*.
- Shi Liu, Kecheng Zheng, and Wei Chen. 2024. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. *arXiv preprint arXiv:2407.21771*.
- Zikang Liu, Longteng Guo, Yepeng Tang, Tongtian Yue, Junxian Cai, Kai Ma, Qingbin Liu, Xi Chen, and Jing Liu. 2025b. Vrope: Rotary position embedding for video large language models. *arXiv preprint arXiv:2502.11664*.
- Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. [Learn to explain: Multimodal reasoning via thought chains for science question answering](#). *Preprint*, arXiv:2209.09513.
- Maria Lymperaiou, Giorgos Ffllandrianos, Angeliki Dimitriou, Athanasios Voulodimos, and Giorgos Stamou. 2025. Halcece: A framework for explainable hallucination detection through conceptual counterfactuals in image captioning. *arXiv preprint arXiv:2503.00436*.
- Fan Ma, Xiaojie Jin, Heng Wang, Yuchen Xian, Jiashi Feng, and Yi Yang. 2024. Vista-llama: Reducing hallucination in video language models via equal distance to visual tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13151–13160.
- Shunqi Mao, Chaoyi Zhang, and Weidong Cai. 2025. Through the magnifying glass: Adaptive perception magnification for hallucination-free vlm decoding. *arXiv preprint arXiv:2503.10183*.
- 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*.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. [Training language models to follow instructions with human feedback](#). *Preprint*, arXiv:2203.02155.
- Ofir Press, Noah A. Smith, and Mike Lewis. 2022. [Train short, test long: Attention with linear biases enables input length extrapolation](#). *Preprint*, arXiv:2108.12409.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.
- Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. 2024. Mitigating object hallucination via data augmented contrastive tuning. *arXiv preprint arXiv:2405.18654*.
- Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. 2025. Octopus: Alleviating hallucination via dynamic contrastive decoding. *arXiv preprint arXiv:2503.00361*.
- Feilong Tang, Zile Huang, Chengzhi Liu, Qiang Sun, Harry Yang, and Ser-Nam Lim. 2025a. [Intervening anchor token: Decoding strategy in alleviating hallucinations for MLLMs](#). In *The Thirteenth International Conference on Learning Representations*.
- Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zile Huang, Haochen Xue, Ziyang Chen, Zelin Peng, Zhiwei Yang, Sijin Zhou, et al. 2025b. Seeing far and clearly: Mitigating hallucinations in mllms with attention causal decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26147–26159.
- Qwen Team. 2024. [Qwen2.5: A party of foundation models](#).
- Chongjun Tu, Peng Ye, Dongzhan Zhou, Lei Bai, Gang Yu, Tao Chen, and Wanli Ouyang. 2025. Attention reallocation: Towards zero-cost and controllable hallucination mitigation of mllms. *arXiv preprint arXiv:2503.08342*.
- Chao Wang, Weiwei Fu, and Yang Zhou. 2025. Tpc: Cross-temporal prediction connection for vision-language model hallucination reduction. *arXiv preprint arXiv:2503.04457*.
- Jiaqi Wang, Yifei Gao, and Jitao Sang. 2024a. Valid: Mitigating the hallucination of large vision language models by visual layer fusion contrastive decoding. *arXiv preprint arXiv:2411.15839*.
- Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. 2023. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*.

- Zehao Wang, Xinpeng Liu, Xiaoqian Wu, Yudonglin Zhang, Zhou Fang, Yifan Fang, Junfu Pu, Cewu Lu, and Yong-Lu Li. 2024b. Verb mirage: Unveiling and assessing verb concept hallucinations in multimodal large language models. *arXiv preprint arXiv:2412.04939*.
- Jinfeng Wei and Xiaofeng Zhang. 2024. Dopra: Decoding over-accumulation penalization and re-allocation in specific weighting layer. *Proceedings of the 32nd ACM International Conference on Multimedia*.
- Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Hao Jiang, Fei Wu, and Linchao Zhu. 2024. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. *arXiv preprint arXiv:2404.14233*.
- Chunzhao Xie, Tongxuan Liu, Lei Jiang, Yuting Zeng, Yunheng Shen, Weizhe Huang, Jing Li, Xiaohua Xu, et al. 2025a. Tarac: Mitigating hallucination in lvlms via temporal attention real-time accumulative connection. *arXiv preprint arXiv:2504.04099*.
- Mingrui Xie, Tianxiang Xu, Qianhai Tang, Shanming Yao, Xiaofeng Zhang, and Junliang Du. 2025b. Dapbr: Distance-aware positional encoding for mitigating object hallucination in lvlms. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8638–8649.
- Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. 2024a. Mitigating object hallucination via concentric causal attention. *arXiv preprint arXiv:2410.15926*.
- Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. 2024b. Mitigating object hallucination via concentric causal attention. *arXiv preprint arXiv:2410.15926*.
- Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. 2024. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. *arXiv preprint arXiv:2412.13817*.
- Ling Yang, Ye Tian, Bowen Li, Xinchun Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. 2025. MMaDA: Multimodal large diffusion language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025. *Dream 7b*.
- Hao Yin, Guangzong Si, and Zilei Wang. 2025. Clear-sight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2503.13107*.
- Koon-Ting Yip, Qiyan Zhao, Wenhao Yu, Liangyu Yuan, Mingkai Li, Xiaofeng Zhang, Jianmin Ji, Yanyong Zhang, Qing Jiang, and Ka-Veng Yuen. 2026a. Sope: Spherical coordinate-based positional embedding for enhancing spatial perception of 3d lvlms. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR' 26)*.
- Koon-Ting Yip, Qiyan Zhao, Wenhao Yu, Xiaofeng Zhang, Jianming Ji, Yanyong Zhang, and Ka-Veng Yuen. 2026b. C²rope: Causal continuous rotary positional encoding for 3d large multimodal-models reasoning. In *2026 IEEE International Conference on Robotics and Automation (ICRA' 26)*.
- Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. 2025. Llada-v: Large language diffusion models with visual instruction tuning. *arXiv preprint arXiv:2505.16933*.
- Runpeng Yu, Xinyin Ma, and Xinchao Wang. 2025. *Dimple: Discrete diffusion multimodal large language model with parallel decoding*. *Preprint, arXiv:2505.16990*.
- Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024. Rlh-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.
- Liu Yuan, Duan Haodong, Zhang Yuanhan, Li Bo, Zhang Songyang, and Zhao Wangbo. 2023. Mmbench: is your multi-modal model an all-around player. *arXiv:2307.06281*.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhua Chen. 2024. *Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi*. *Preprint, arXiv:2311.16502*.
- Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. 2023. Halls-switch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption. *arXiv preprint arXiv:2310.01779*.
- Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. 2025. Shallow focus, deep fixes: Enhancing shallow layers vision attention sinks to alleviate hallucination in lvlms. pages 3512–3534.
- Xiaofeng Zhang, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. 2024. From redundancy to relevance: Enhancing explainability in multimodal large language models. *arXiv preprint arXiv:2406.06579*.
- Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Jiawei Cao, Hao Cheng, and Kaijie Wu. 2026a. What drives attention sinks? a study of massive activations and rotational positional encoding in large vision-language models. *Information Processing & Management*, 63(2):104431.

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Xiaosong Yuan, Qiyao Zhao, Jiawei Cao, Feilong Tang, Sinan Fan, Yaomin Shen, Chen Shen, et al. 2026b. Hallucination begins where saliency drops. In *The Fourteenth International Conference on Learning Representations*.

Haozhe Zhao, Shuzheng Si, Liang Chen, Yichi Zhang, Maosong Sun, Mingjia Zhang, and Baobao Chang. 2024. Looking beyond text: Reducing language bias in large vision-language models via multimodal dual-attention and soft-image guidance. *arXiv preprint arXiv:2411.14279*.

Qiyao Zhao, Xiaofeng Zhang, Shuochen Chang, Qianyu Chen, Xiaosong Yuan, Xuhang Chen, Luoqi Liu, Jiajun Zhang, Xu-Yao Zhang, and Da-Han Wang. 2026. Context tokens are anchors: Understanding the repetition curse in dmlms from an information flow perspective. In *The Fourteenth International Conference on Learning Representations*.

Qiyao Zhao, Xiaofeng Zhang, yiheng Li, Yun Xing, Yuan Xiaosong, Feilong Tang, Sinan Fan, Xuhang Chen, Xuyao Zhang, and Dahan Wang. 2025. Mca-llava: Manhattan causal attention for reducing hallucination in large vision-language models. *The 33rd ACM International Conference on Multimedia*.

Haojie Zheng, Tianyang Xu, Hanchi Sun, Shu Pu, Ruoxi Chen, and Lichao Sun. 2024. Thinking before looking: Improving multimodal llm reasoning via mitigating visual hallucination. *arXiv preprint arXiv:2411.12591*.

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. 2025. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. In *The Thirteenth International Conference on Learning Representations*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

Fengqi Zhu, Rongzhen Wang, Shen Nie, Xiaolu Zhang, Chunwei Wu, Jun Hu, Jun Zhou, Jianfei Chen, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Llada 1.5: Variance-reduced preference optimization for large language diffusion models. *Preprint*, arXiv:2505.19223.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2024. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. *arXiv preprint arXiv:2402.18476*.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2020. Fine-tuning language models from human preferences. *Preprint*, arXiv:1909.08593.

A Information Flow Analysis on LaViDa

To verify that the two bottlenecks we identified in Section 3 (attention concentration on *image_newline* anchors and RoPE-induced semantic blind spots) are not specific to LLaDA-V, we replicate the same information-flow analysis on LaViDa. Following the protocol in Section 3.2, we randomly sample 1000 examples from MSCOCO and track the attention dynamics across all layers during inference.

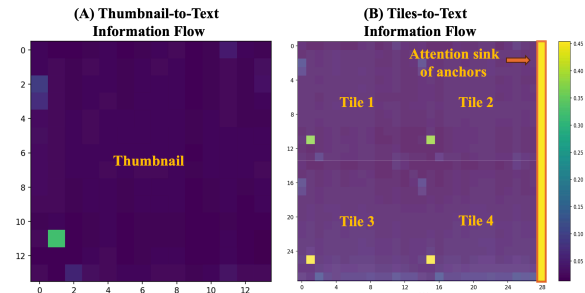


Figure 5: Visual-to-text information flow on LaViDa. The *image_newline* tokens act as semantic anchors, mirroring the phenomenon observed on LLaDA-V.

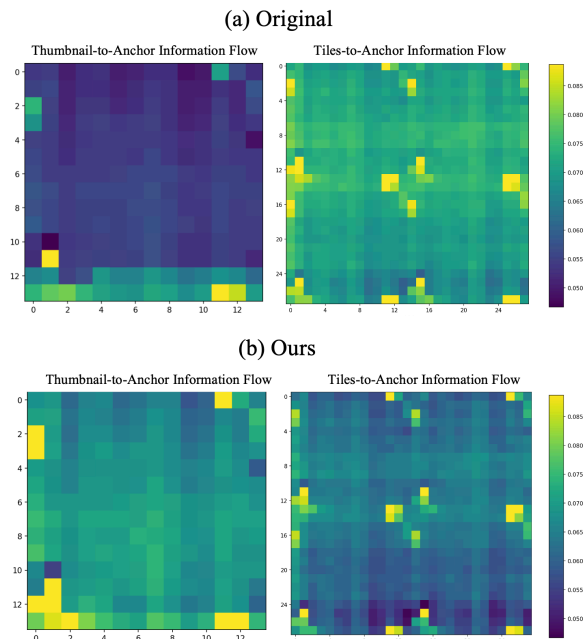


Figure 6: Bidirectional information flow between *image_newline* anchors and visual tokens on LaViDa. (a) The original model suffers from the same semantic blind spots on the thumbnail as LLaDA-V. (b) After applying EDAR, the information flow becomes notably more uniform, effectively eliminating the blind spots.

Figure 5 visualizes the visual-to-text information flow on LaViDa. Consistent with the observation on LLaDA-V, text tokens attend only

sparse to standard visual content tokens, while the *image_newline* anchors absorb a disproportionately large share of the attention budget, exhibiting a pronounced attention sink.

Figure 6 further illustrates the bidirectional information flow between the *image_newline* anchors and the visual tokens. The thumbnail region again exhibits a clear long-range decay pattern: tokens appearing earlier in the flattened sequence receive markedly lower bidirectional attention than later tokens, forming the same "semantic blind spots" as in LLaDA-V, while the localized tiles remain relatively uniform.

Moreover, Figure 6(b) shows that after applying EDAR on LaViDa, the bidirectional information flow between the *image_newline* anchors and the thumbnail becomes substantially more uniform: the previously under-attended tokens at the beginning of the flattened thumbnail sequence are now activated with comparable intensity to the later ones, while the localized tiles remain stable. This mirrors the improvement we observe on LLaDA-V (Figure 4), indicating that EDAR consistently eliminates the semantic blind spots across different dMLLMs.

These results confirm that the attention-sink and semantic-blind-spot issues are not idiosyncratic to LLaDA-V but appear to be general bottlenecks in dMLLMs, motivating the need for a generic remedy such as EDAR.

B Computational Overhead of EDAR

As noted in our limitations, although EDAR is training-free and requires no additional parameters, the explicit attention re-distribution introduces extra computation at every decoding step, leading to moderate increases in inference latency and GPU memory consumption. To quantify this overhead, we benchmark the original LLaDA-V and LLaDA-V+EDAR on a single NVIDIA RTX PRO 6000 GPU under two input lengths: $N = 1674$ and $N = 3888$. In the experiment, we use the original inference method without `fast_dllm`.

Table 5 reports the per-token latency and peak GPU memory usage. EDAR incurs a $1.16\times$ latency overhead and a $1.03\times$ memory overhead at $N = 1674$, and a $1.39\times$ latency overhead with a $1.11\times$ memory overhead at $N = 3888$. The overhead grows mildly with sequence length. Nevertheless, the absolute memory increase remains small, and the latency cost is a reasonable trade-off

given the consistent accuracy gains reported in the main paper. We leave further optimization (e.g., fused kernels) to future work.

Seq Len (N)	Method	Latency (ms/tok)	GPU Mem (GB)
1674	Original	148.4	26.1
	+EDAR	173.2 (1.162 \times)	27.0 (1.034 \times)
3888	Original	267.2	38.2
	+EDAR	371.7 (1.391 \times)	42.2 (1.105 \times)

Table 5: Computational overhead of Extrinsic Distance-Aware Regularization on LLaDA-V.


User	Please describe the image in detail.
Image	
LLaDA-V	<p>In the heart of a bustling city, a daring skateboarder is caught in a moment of pure exhilaration. The skateboarder, clad in a white t-shirt and gray pants, is suspended in mid-air, arms outstretched in a display of balance and control. His skateboard, a testament to his skill, is poised above an orange and white barrier, its vibrant colors contrasting against the muted tones of the background buildings.</p>
+CCA	<p>The image captures a dynamic moment of a skateboarder performing a trick over a striped barrier. The skateboarder is in mid-air, with the skateboard visible beneath them. The background is a street scene with a clear blue sky and a few clouds. The building in the background has large windows and a sign with a green and white logo. The skateboarder is wearing a black shirt and gray pants, and the skateboard is orange and white striped.</p>
+MCA	<p>In the heart of a bustling city, the athlete leaps above a yellow and white barrier, wearing a black t-shirt and gray pants. His helmet shines under sunlight, with spectators and tall beige buildings forming the lively urban background. The scene is a symphony of motion and energy, set against the backdrop of the urban landscape.</p>
+Farsight	<p>In the heart of a bustling city, a skateboarder is caught in a moment of pure exhilaration. The skateboarder, clad in a black t-shirt and gray pants, is suspended in mid-air, his arms outstretched in a display of balance and control. His skateboard, a testament to his skill, is directly beneath him, hovering above the orange and white striped construction barrier he's jumping over. The skateboard is seen on the ground, and a red brick building glows behind under a blue sky.</p>
+Ours	<p>In the heart of a bustling city, a daring skateboarder is caught in mid-air, defying gravity with an impressive trick. The skateboarder, clad in a black t-shirt and gray pants, is suspended above a red and white striped concrete barrier. His skateboard, a loyal companion in this daring feat, is visible beneath him. The backdrop to this thrilling scene is a beige building, its windows reflecting the clear blue sky. A crowd of onlookers watches intently, their attention riveted on the skateboarder's airborne performance.</p>

Table 6: Case Study (1)
19373


User	Please describe the image in detail.	
Image		
LLaDA-V	<p>In the heart of the action, a baseball player in a blue and white uniform is captured mid-swing, his body twisted in a powerful arc as he attempts to hit the ball. The ball, a pristine white, is frozen in mid-air, positioned slightly above and to the right of the player. Behind him, a catcher in a green and black uniform is crouched in anticipation, ready to catch the ball. The field itself is enclosed by a chain link fence, providing a stark contrast to the lush green of the field.</p>	
+CCA	<p>The image captures a moment during a baseball game. The batter is in the process of swinging the bat, while the catcher is crouched behind him, ready to catch the ball. The batter is wearing a green and white uniform, while the umpire is in a red and black uniform. The scene is set on a baseball field, with spectators visible in the background. The lighting is dim, and the atmosphere is focused on the intense moment of the game.</p>	
+MCA	<p>The batter, wearing blue and white, swings forcefully as the catcher crouches in the dirt. Several players running across the grass under bright daylight add an energetic mood, though inaccurate. Other players and spectators can be seen in the background, adding to the lively atmosphere of the game.</p>	
+Farsight	<p>This image captures a dynamic moment in a baseball game. The central figure is a batter, dressed in a blue and white uniform, who is in the midst of swinging his bat. The ball is frozen in mid-air, indicating the force of the swing. The catcher stands upright, ready to catch the ball. The ball has hit the ground, raising a small cloud of dust. The image beautifully encapsulates the intensity and dynamism of a baseball game.</p>	
+Ours	<p>In the heart of the action, a young baseball player, clad in a vibrant blue and white uniform, is captured mid-swing. His body is coiled with energy, his bat poised to strike the ball. The ball, frozen in mid-air above the batter, mirrors the intensity of the moment. The catcher, dressed in a contrasting red uniform, crouches behind the batter, ready to catch the ball. The field itself is enclosed by a chain-link fence, providing a stark contrast against the green of the field.</p>	

Table 7: Case Study (2)