# WHEN ANCHORS COMMIT, COHERENCE BREAKS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Diffusion multimodal large language models (dMLLMs) generate responses through iterative denoising, with predictive confidence determining which masked positions are committed first. Under full-sequence confidence-based decoding, we identify a three-phase commitment pattern in the evaluated trajectories: opening sentences first, tail anchors next, and middle content last. This timing imbalance leaves extended masked spans in the middle of a response whose surrounding content is already fixed, and is associated with semantic discontinuities. We investigate local attention concentration as an intervention target: sequence-wide detection can prioritize dominant peaks at already committed positions while overlooking concentrations in unresolved regions. We propose **GADS** (Gap-Aware Dynamic Suppression), a training-free intervention that detects concentrated attention columns within a local window and applies a reference-preserving logit bias, while retaining full-sequence token selection. GADS shifts commitment toward previously delayed middle content and reduces the tail-before-middle gap from 31.6 to 6.7 global denoising steps on LLaVA-Bench. Across LLaDA-V, MMaDA, and LaViDa, it improves LLaVA-Bench scores by 12.0, 21.3, and 2.1 points, respectively, while reducing object hallucination rates on all three backbones. These results support local attention intervention as an effective approach to improving commitment trajectories and long-form multimodal generation.

## 1 INTRODUCTION

Autoregressive multimodal large language models (MLLMs) generate tokens left-to-right, conditioning each prediction on a previously generated prefix. Diffusion MLLMs (dMLLMs) Nie et al. (2025); Bie et al. (2025); You et al. (2025); Li et al. (2025c); Yang et al. (2025b); Yu et al. (2025); Swerdlow et al. (2025) instead generate through iterative denoising under bidirectional attention. We study *full-sequence confidence-based decoding*: at each denoising step, all remaining masked positions in the generation span are eligible for commitment, and the sampler selects positions according to predictive confidence. Unlike semi-autoregressive decoding, this setting does not restrict commitment to sequentially activated blocks. All commitment times in our trajectory analysis are *global denoising steps*.

Figures 1 and 2 illustrate a three-phase commitment pattern: opening sentences are committed first, tail anchors next, and mid-response content last. We use *structural tokens* to refer to `<eos>`, periods, and newlines; frequent function words such as "of" and "to" are not included in this category. Early commitment of a structural token is not inherently harmful. The concern arises when a response's opening and tail become fixed while substantial intervening content remains unresolved. In the illustrated trajectories, this timing imbalance accompanies topic drift and inconsistencies between earlier and later content.

One possible explanation for the imbalance is that structural predictions are easier than image-conditioned content predictions. Multiple descriptions may be compatible with an image, whereas some formatting tokens are more predictable. Instruction-tuning templates may further favor these tokens. We treat these explanations as hypotheses: the present comparisons do not independently identify the contributions of modality, task ambiguity, and training data.

Our analysis examines *local attention concentration* as an intervention target. Attention sinks have been studied in autoregressive LLMs Xiao et al. (2023); Sun et al. (2024); Zhang et al. (2024; 2025d) and text-only dLLMs Rulli et al. (2025); Huang et al. (2026); Ni et al. (2026a); Frkovic et al.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

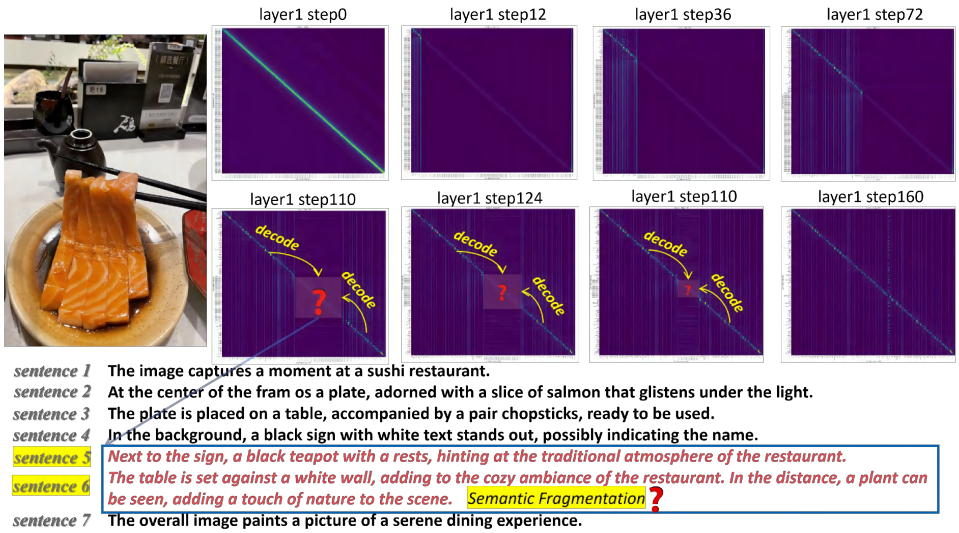


Figure 1: **Confidence-driven decoding in dMLLMs produces a three-phase order.** Commit-step profile of a LLaDA-V response: opening sentences commit first, tail anchors (<eos>, final period) commit next, and middle paragraphs commit last. The frozen opening and closing force mid-response content to be filled into a fixed frame, which produces semantic discontinuity at block boundaries.

(2026); Zhang et al. (2026e;c); Dai et al. (2026); Ni et al. (2026b). We investigate their relationship to structural-token commitment in multimodal full-sequence decoding. Concentrated attention and high predictive confidence are distinct quantities: a column receiving substantial attention does not, by itself, establish that the prediction at that position becomes overconfident. Our intervention experiments examine how changing local attention affects predictions and the resulting commitment trajectory.

A detector that aggregates attention over the entire generation span can prioritize a large peak at an already committed position and overlook a smaller concentration in an unresolved region (Figure 5). This motivates GADS (Gap-Aware Dynamic Suppression), which detects and reweights concentrated attention within a *local intervention window*. This window defines where attention is measured and modified; it does not restrict which masked positions the full-sequence sampler may commit. Our contributions are three-fold:

1. We characterize a three-phase commitment pattern in full-sequence multimodal diffusion decoding and quantify its timing imbalance on LLaVA-Bench using global commit steps and paired bootstrap intervals.
2. We propose GADS, a training-free local attention intervention based on column-energy gaps and reference-preserving logit biases. We examine changes in both commitment trajectories and response quality.
3. Experiments on LLaDA-V, MMaDA, and LaViDa show improvements in long-form evaluation scores and reductions in hallucination metrics. The trajectory and boundary analyses provide complementary evidence about continuity, rather than treating general benchmark scores as direct measures of coherence.

## 2 Motivation: Confidence-Driven Decoding Order and Its Pathology

In our full-sequence setting, the sampler selects commitments from all remaining masked generation positions at each global step. Attention interventions can change token predictions and, consequently, the confidence ranking used by the sampler. We first characterize the observed commitment

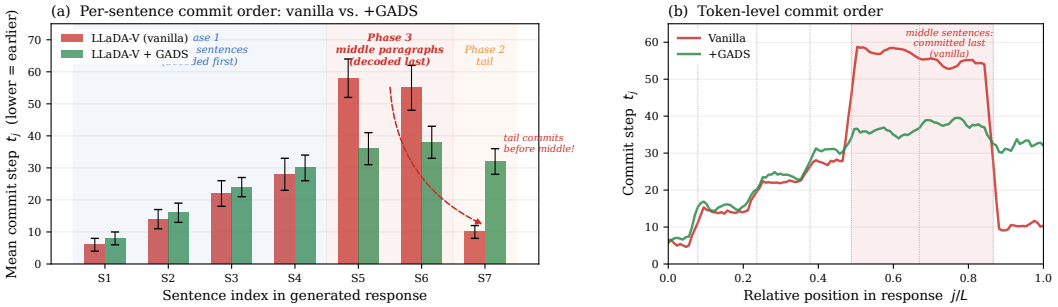


Figure 2: **Illustrative global commitment profiles under full-sequence decoding.** **(a)** Per-sentence mean global commit step for the displayed responses. Vanilla LLaDA-V (red) commits the tail before some middle sentences; GADS (green) reduces this timing imbalance. **(b)** Token-level global commit steps across normalized response positions. The mid-response peak is reduced with GADS. Sentence indices and normalized positions describe each generated response and do not guarantee semantic alignment between different outputs.

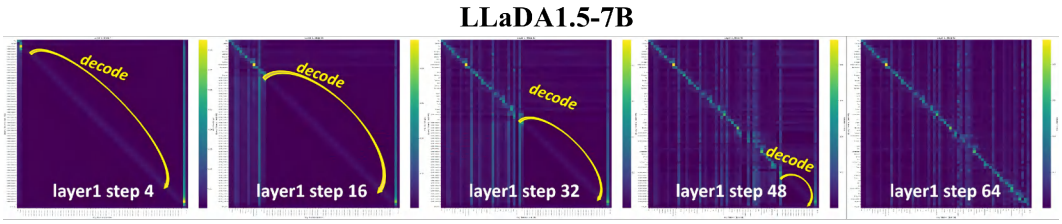


Figure 3: **A text-only comparison example.** Layer-1 attention maps of LLaDA-1.5-7B across the displayed denoising steps. The example illustrates a different attention pattern from the multimodal cases. Attention maps alone do not establish commitment order or rule out the phenomenon in text-only dLLMs; that comparison requires matched trajectory measurements.

trajectories, then examine local attention concentration as an intervention target. We distinguish observations and intervention effects from the stronger claim that commitment order uniquely mediates changes in response quality.

**Takeaway 1**

Under full-sequence confidence-based decoding, tail anchors can be committed before intervening semantic content. The resulting global timing imbalance is associated with discontinuities in the evaluated responses. Early structural commitment alone is not sufficient to diagnose an error.

### 2.1 The Three-Phase Decoding Order

We instrument LLaDA-V to record, for each generated position $j \in \{0, \ldots, L-1\}$, the global denoising step $t_j$ at which it is first committed (Figures 2 and 4). Here $L$ denotes the allocated generation length. Because commitment candidates are selected over the full sequence, a tail position can be committed while earlier positions remain masked. The displayed trajectories exhibit three phases: opening sentences first, tail anchors next, and middle content last. These are descriptive phases of the observed trajectories, not constraints imposed by the sampler.

**Quantifying the pattern at dataset scale.** On the 60-prompt LLaVA-Bench In-the-Wild set, we report per-prompt measurements and aggregate them across prompts. These include global commit times; structural-token confidence; a tail-before-middle timing gap; boundary-similarity dip; and topic-switch rate. For prompts without a valid <eos>, the EOS-time statistic uses $S + 1$ as a missing-termination penalty, where $S$ is the total denoising budget. This value is not an observed commitment time, and abnormal termination should also be reported separately.

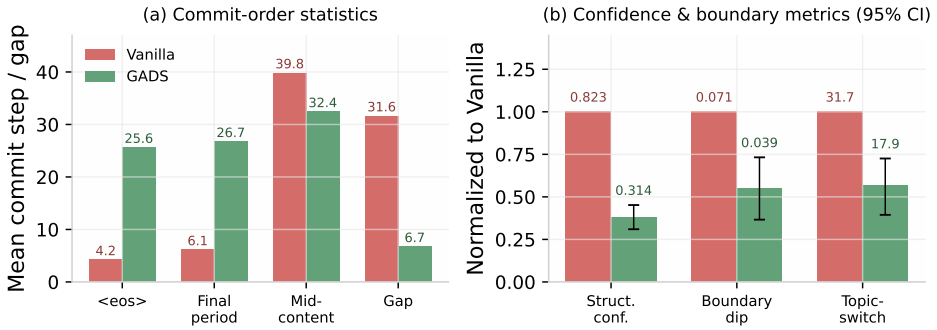


Figure 4: **Dataset-level commitment and continuity measurements** (LLaVA-Bench, 60 prompts; 10,000 paired bootstrap resamples). (a) GADS delays mean global commitment of <eos> from 4.2 to 25.6 and the final period from 6.1 to 26.7; the reported tail-before-middle gap decreases from 31.6 to 6.7 steps. (b) Reported structural-token confidence decreases from 0.823 to 0.314, boundary-similarity dip from 0.071 to 0.039, and topic-switch rate from 31.7% to 17.9%. Relative mean output length changes from 1.000 to 1.012. Similar mean lengths do not establish distributional equivalence or exclude truncation in individual responses.

Let $\mathcal{M}_i$ and $\mathcal{T}_i$ denote the middle-content and tail-anchor positions for prompt $i$, respectively. For prompts with valid position sets, define

$$\bar{t}^{(i)}_{\text{mid}} = \frac{1}{|\mathcal{M}_i|} \sum_{j \in \mathcal{M}_i} t_j, \qquad \bar{t}^{(i)}_{\text{tail}} = \frac{1}{|\mathcal{T}_i|} \sum_{j \in \mathcal{T}_i} t_j, \qquad \Delta_{\text{gap}} = \frac{1}{N} \sum_{i=1}^{N} \left( \bar{t}^{(i)}_{\text{mid}} - \bar{t}^{(i)}_{\text{tail}} \right), \tag{1}$$

where $N$ is the number of evaluated prompts included under the stated missing-data rule. A positive gap means that the selected tail anchors are committed earlier, on average, than the selected middle content.

The reported mean EOS commit time increases from 4.2 to 25.6 with GADS, while mean middle-content commit time decreases from 39.8 to 32.4. The reported gap decreases from 31.6 to 6.7 steps, a reduction of approximately 79%. These changes accompany a smaller boundary-similarity dip (0.071 to 0.039) and a lower topic-switch rate (31.7% to 17.9%). They establish joint changes in timing and continuity measurements, but do not isolate the contribution of timing from other effects of the attention intervention. The reported length-change interval includes zero; this is not an equivalence test and does not establish an absence of truncation.

**Local attention concentration as an intervention target.** The inspected attention maps show concentrations at structural-token positions. GADS changes these local attention patterns and is accompanied by changes in confidence and commitment timing. However, incoming attention to a key position is not the same as the output confidence at that position: the former directly affects the representations of attending queries. Establishing the link requires measurements at matched intermediate states, rather than inferring it from co-location alone.

**Why early commitments may constrain later content.** Masked diffusion training includes configurations with visible tokens on both sides of a masked span Nie et al. (2025); the mask shape itself is therefore not evidence of an unseen training condition. A relevant distinction is that visible training tokens come from a coherent reference sequence, whereas visible inference tokens are model predictions that may be mutually incompatible. An opening and a tail committed before intervening content may constrain subsequent completion. This provides a hypothesis for the observed cross-sentence topic drift and conclusion–premise inconsistencies, without assuming that the two ends were predicted independently or that the mask configuration is out of distribution.

**Scope of the multimodal observation.** Our findings concern the evaluated multimodal models and full-sequence decoding configurations. Figure 3 supplies a text-only comparison example, but does not establish the absence of similar behavior in text-only dLLMs. Task ambiguity, output

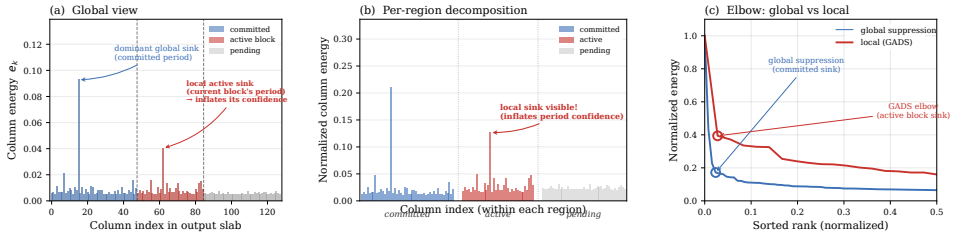


Figure 5: **Global and local views of attention concentration.** global suppression refers to directly reducing the attention applied to the columns. (a) The displayed sequence-wide energy profile contains a dominant peak at a committed period and a smaller peak in the local intervention region. (b) Region-wise normalization makes the latter concentration more visible. (c) Global and local energy profiles can yield different gap selections. Local selection does not imply that positions outside the window are ineligible for commitment, and a committed position is not necessarily irrelevant to subsequent predictions.

length, instruction tuning, and sampling settings may all influence the gap between structural- and content-token confidence. Controlled comparisons are needed to attribute the observed differences specifically to multimodality.

### 2.2 Local Detection Within Full-Sequence Decoding

A large attention peak at an already committed position can dominate a sequence-wide selection criterion. Restricting detection to a local region changes which concentrations are selected (Figure 5). This motivates a distinction between the *commitment domain*, which includes all remaining masked generation positions, and the *intervention window*, which defines the query–key submatrix used by GADS.

At global step $t$, let $\mathcal{C}_t$ and $\mathcal{U}_t$ be the committed and unresolved generation positions. They partition the generation span but need not be contiguous. Let $\mathcal{W}_t$ be the local intervention window. Window membership and commitment status are different attributes. We describe the attention operator conditional on a local window $\mathcal{W}_t$, independently of the full-sequence commitment policy. Concentrations outside the window may still influence predictions; the local detector targets a particular subset instead of assuming that all other positions are irrelevant.

### 2.3 A Local-Suppression Comparison

Figure 6 compares vanilla decoding, global suppression (refers to directly reducing the attention applied to the columns), and GADS using continuity measurements. The displayed results favor GADS over the global variant. This comparison motivates local detection, but does not by itself isolate locality if selection rules, reference levels, or bias strengths also differ. The comparison therefore evaluates the complete interventions; attributing the difference exclusively to locality additionally requires matching the other design choices.

## 3 Method: Gap-Aware Dynamic Suppression

The observations in §2 motivate intervening on locally concentrated attention during full-sequence denoising. GADS modifies a local attention submatrix while retaining the sampler's full-sequence commitment domain. The intervention can affect both token representations and their confidence ranking; improved response quality need not be mediated exclusively by changes in commitment order.

### 3.1 Setup

Consider one attention head at one layer and global denoising step. Let $A \in \mathbb{R}^{n\times n}$ be its pre-softmax logit matrix over $n$ positions, including conditioning and generation positions, and let $\mathfrak{S}$

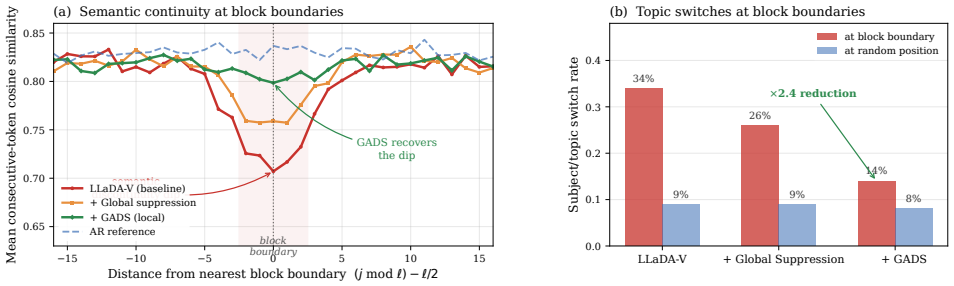


Figure 6: **Continuity measurements under global and local intervention.** global suppression refers to directly reducing the attention applied to the columns. (a) Consecutive-token cosine similarity around the evaluated boundaries for vanilla decoding (red), global suppression (orange), and GADS (green). Similarity is a continuity proxy, not a direct measure of discourse consistency. (b) Reported topic-switch rates at evaluated boundaries on the CHAIR 500-image set decrease from 34% with vanilla decoding to 14% with GADS. These measurements use a different evaluation set from Figure 4.

denote row-wise softmax. Let $\mathcal{W} \subseteq \{1, \ldots, n\}$ be a local intervention window contained in the generation span, with $m = |\mathcal{W}|$. We omit the step index for brevity. The window is an input to the operator defined below; the same set indexes its query rows and key columns. This formulation separates attention reweighting from the choice of window. The window is not a commitment mask: unresolved positions outside $\mathcal{W}$ remain eligible for selection by the sampler.

### 3.2 Construction

**Step 1: Measure local column energy.** Within the window, we compute the mean attention received by each column,

$$e_k = \frac{1}{m} \sum_{i \in \mathcal{W}} \left[\mathfrak{S}(A_{\mathcal{W}})\right]_{i,k}, \quad k \in \mathcal{W}, \tag{2}$$

where $A_{\mathcal{W}}$ is the $m \times m$ sub-block of $A$ indexed by $\mathcal{W}$ and the softmax is renormalized within the window, so the energies sum to one within the window. Their baseline still depends on the window size. Uniform attention yields $e_k = 1/m$; an over-attended column (an *attention sink* in the terminology of prior work Xiao et al. (2023); Dai et al. (2026); Ni et al. (2026b)) yields $e_k \gg 1/m$.

**Step 2: Locate the concentration gap.** Sort the energies in non-increasing order, $\tilde{e}_1 \geq \cdots \geq \tilde{e}_m$, and let $\pi(r)$ denote the original position at rank $r$. Ties are resolved by position index. Define the candidate split indices

$$\mathcal{I} = \{r \in \{1, \ldots, m-1\} : \tilde{e}_r \geq 2/m\}. \tag{3}$$

If $m < 2$ or $\mathcal{I}$ is empty, GADS returns $A$ unchanged. Otherwise,

$$i^\star = \arg\max_{r \in \mathcal{I}} \frac{\tilde{e}_r}{\tilde{e}_{r+1} + \varepsilon}, \qquad \varepsilon = 10^{-8}, \tag{4}$$

with the smallest rank chosen in a tie. This candidate-set form specifies the elevated-head search, including its empty-set case. The selected gap is a heuristic for separating larger from smaller column energies; it does not identify semantic versus non-semantic columns by itself. The column at rank $i^\star$ is the *reference column*. We reserve "structural anchor" for token categories and use "reference" for this energy-based algorithmic choice.

**Step 3: Decide whether to intervene.** Let $M = \sum_{r=1}^{i^\star} \tilde{e}_r$ be the cumulative attention mass of the top $i^\star$ columns. The mass condition is

$$T = \mathbb{1}[M > 2i^\star/m] \in \{0, 1\}. \tag{5}$$

It is evaluated separately for each head and layer. Because every selected top-ranked energy is already at least $2/m$, the candidate filter implies $M \geq 2i^\star/m$. Thus, the mass condition mainly resolves the equality case of the preceding filter rather than providing an independent measure of harmful attention.

**Step 4: Suppress above the reference.** For a key position $k \in \mathcal{W}$, let $\rho(k)$ denote its sorted rank. Define

$$\gamma_k = \begin{cases} \min(1, \tilde{e}_{i^\star}/e_k), & T = 1 \text{ and } \rho(k) < i^\star, \\ 1, & \text{otherwise.} \end{cases} \tag{6}$$

We set $\gamma_k = 1$ for all positions in a no-op case and extend it by one to positions outside $\mathcal{W}$. The intervention is

$$[\mathcal{G}(A)]_{i,j} = A_{i,j} + \mathbb{1}[(i,j) \in \mathcal{W} \times \mathcal{W}] \log \gamma_j. \tag{7}$$

The attention output is $\mathfrak{S}(\mathcal{G}(QK^\top/\sqrt{d}))V$, where $Q, K, V$ are the query, key, and value matrices and $d$ is the query/key dimension. Selected columns receive non-positive logit biases; the reference receives zero direct bias. This construction does not guarantee that normalized column energies equal the reference energy, because softmax renormalization is row dependent.

**Scope of the suppression rule.** The reference is preserved by construction. Consequently, when $i^\star = 1$, no column ranks above it and the intervention is the identity even if the mass condition is satisfied. The present rule therefore does not suppress every isolated concentration. Selecting a reference below the largest gap would address this case but defines a different intervention, requiring separate evaluation.

### 3.3 Structural Properties

The following properties characterize the intervention algebraically. They do not guarantee preservation of semantic information or improved generation quality.

**Proposition 1** (Logit-support locality)**.** $\big[\mathcal{G}(A)\big]_{i,j} = A_{i,j}$ *whenever* $(i,j) \notin \mathcal{W} \times \mathcal{W}$.

*Proof.* Immediate from the indicator in equation 7. □

This locality statement concerns logits. For a query inside $\mathcal{W}$, changing logits assigned to keys inside the window changes the full-row softmax denominator and can therefore alter attention probabilities assigned to visual or prompt keys outside the window. Query rows outside $\mathcal{W}$ are unchanged at this attention operation, although changes to hidden states may propagate through subsequent layers.

**Proposition 2** (Reference-logit preservation)**.** *When a reference is defined, its column receives* $\gamma = 1$ *independently of the trigger value* $T$.

*Proof.* The first branch of equation 6 applies only to columns ranked strictly above $i^\star$, and the reference has rank exactly $i^\star$. □

Proposition 2 states that the reference receives no direct logit penalty. Its normalized attention probability can nevertheless change. For finite logits, all multiplicative factors remain positive, so the operation reweights attention rather than assigning selected entries exactly zero probability.

**Proposition 3** (Log-bias equivalence)**.** *For any row* $\mathbf{a} \in \mathbb{R}^n$ *and* $\boldsymbol{\gamma} \in (0,1]^n$*, with* $p_k = \mathfrak{S}(\mathbf{a})_k$*,*

$$\mathfrak{S}(\mathbf{a} + \log \boldsymbol{\gamma})_k \;=\; \frac{\gamma_k\, p_k}{\sum_j \gamma_j\, p_j}.$$

*Proof.* $\exp(a_k + \log \gamma_k) = \gamma_k \exp(a_k)$; the claim follows from the definition of $\mathfrak{S}$. □

Proposition 3 establishes the equivalence between additive logit bias and multiplicative attention reweighting followed by row-wise renormalization. For a query inside $\mathcal{W}$, the factors outside the window are one; for a query outside $\mathcal{W}$, all factors are one. This algebraic equivalence does not establish the computational cost of collecting the attention statistics or compatibility with a particular fused-attention implementation.

Table 1: **Consolidated results across benchmarks.** Three dMLLM backbones (LLaDA-V, MMaDA, LaViDa) against four methods: two scheduler-side methods (AdaBlock, Wavefront) and two confidence-guided decoding methods (CoTA, CACD). **Green** = best per metric within each backbone; underline = second best.

| | | LLaVA-Bench | | | | CHAIR | | | MathVista | ScienceQA | MME | | GQA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Method** | all↑ | conv↑ | detail↑ | complex↑ | $C_s$↓ | $C_i$↓ | recall↑ | acc↑ | acc↑ | cog.↑ | perc.↑ | EM↑ |
| **LLaDA-V** | Baseline | 57.8 | 34.5 | 64.2 | 69.7 | 6.4 | 4.8 | 36.2 | 52.9 | 89.74 | 343.9 | 1411.9 | 52.14 |
| | AdaBlock | 68.0 | 57.9 | 65.4 | 76.9 | 6.2 | 4.8 | 37.9 | 53.8 | 89.74 | 343.9 | 1411.9 | 52.14 |
| | Wavefront | 51.3 | 43.5 | 54.5 | 54.9 | 6.2 | 4.6 | 33.0 | 53.1 | **89.79** | 395.4 | 1370.6 | 52.11 |
| | CoTA | 55.0 | 33.4 | 60.8 | 67.7 | 7.5 | 17.6 | **63.1** | **54.8** | 89.50 | 490.5 | 1492.7 | 51.8 |
| | CACD | – | – | – | – | **5.38** | **3.8** | 37.3 | 33.5 | 79.52 | – | – | – |
| | **GADS (ours)** | **69.8** | **61.0** | **66.3** | **77.0** | 5.4 | 4.0 | 38.0 | 54.1 | **89.79** | **508.2** | **1508.2** | **53.94** |
| **MMaDA** | Baseline | 23.4 | 23.0 | 28.5 | 20.9 | 16.3 | 7.6 | 39.2 | 29.8 | **57.11** | 229.3 | 842.6 | 39.45 |
| | AdaBlock | 23.4 | 27.2 | 22.4 | 17.5 | 17.4 | 7.5 | 43.0 | 30.3 | 56.47 | 229.3 | **843.0** | 42.17 |
| | Wavefront | 20.9 | 17.8 | 28.7 | 17.8 | 18.6 | 9.4 | 38.9 | 29.9 | 56.47 | 239.3 | 744.5 | 38.93 |
| | CoTA | 21.2 | 17.8 | 28.6 | 18.4 | 18.9 | 8.8 | **58.4** | 29.7 | 56.9 | 225.5 | 832.0 | 38.3 |
| | CACD | – | – | – | – | 19.76 | 10.28 | 40.63 | 28.0 | 51.71 | – | – | – |
| | **GADS (ours)** | **44.7** | **28.1** | **45.7** | **54.8** | **15.4** | **6.9** | 40.8 | **30.5** | 56.92 | **268.9** | 836.3 | **43.14** |
| **LaViDa** | Baseline | 57.1 | **38.5** | 64.1 | 65.3 | 10.2 | 6.9 | 34.7 | 41.0 | 71.54 | 357.9 | 1355.4 | 56.58 |
| | AdaBlock | 50.0 | 32.3 | 52.6 | 60.2 | 6.6 | 5.0 | 31.6 | 41.3 | 71.59 | 351.8 | 1339.3 | 56.03 |
| | Wavefront | 46.2 | 37.3 | 47.8 | 51.0 | 5.8 | **4.3** | 31.7 | 41.1 | 72.73 | 337.1 | 1338.5 | 54.25 |
| | CoTA | 53.8 | 34.0 | 55.7 | 60.6 | 16.3 | 7.1 | **58.1** | 41.3 | 71.2 | 340.7 | 1348.2 | 54.7 |
| | CACD | – | – | – | – | **5.08** | 7.58 | 35.31 | **45.70** | **73.03** | – | – | – |
| | **GADS (ours)** | **59.2** | 35.4 | **70.5** | **68.9** | 5.4 | 4.4 | 35.4 | 41.1 | 71.54 | **361.8** | **1359.2** | **56.89** |

# 4 EXPERIMENTS

## 4.1 SETUP AND DATASETS

We evaluate three dMLLMs with distinct visual integration strategies: LLaDA-V, MMaDA-8B-Base, and LaViDa. We compare our approach with four other methods: AdaBlockLu et al. (2025) (adaptive block decoding with semantic boundary alignment) and WavefrontYang et al. (2025a) (dynamic boundary expansion starting from the final position). We also compare COTAZhao et al. (2026a) to alleviate duplicate tokens in dMLLM by intervening in attention and entroy, thus enhancing the accuracy of the full-text description, and CACD Zhao et al. (2026b) to improve the ill-conditioned decoding of dMLLM that relies solely on confidence output through cluster sampling.

Benchmarks fall into three groups: (i) long-form and reasoning tasks, namely CHAIR Rohrbach et al. (2018) and LLaVA-Bench (In-the-Wild) Liu et al. (2023) for image description and MathVista Lu et al. (2024) for mathematical reasoning; (ii) general VQA (ScienceQA Lu et al. (2022)); and (iii) discriminative generation (MME Yin et al. (2023), GQA Hudson & Manning (2019)).

## 4.2 EVALUATION RESULTS AND ANALYSIS

**GADS improves long-form semantic coherence.** The largest gains appear on benchmarks requiring sustained multi-sentence coherence, where the three-phase pathology is most active. On LLaDA-V, GADS reaches an LLaVA-Bench *all* score of 69.8, outperforming AdaBlock (68.0), CoTA (55.0), and Wavefront (51.3). The gap is largest on *conv* (61.0 vs. 57.9 for AdaBlock), which requires mid-response content to stay consistent with the opening and closing. On LaViDa, GADS improves *all* from 57.1 to 59.2 with the largest gain on *detail* (64.1 to 70.5). On MMaDA the effect is larger still: GADS raises *all* from 23.4 to 44.7 and *complex* from 20.9 to 54.8, while every baseline that reports LLaVA-Bench remains within 3 points of the vanilla model. One possible explanation is that MMaDA's smaller capacity makes it more susceptible to confidence-landscape distortion, since its denoiser has fewer parameters to compensate.

**GADS achieves the best hallucination–recall balance.** On CHAIR, no single method wins every cell, and the trade-offs are informative. CoTA reports the highest recall (up to 63.1) but inflates $\text{CHAIR}_i$ to 17.6 on LLaDA-V, consistent with over-generating object mentions rather than grounding them. CACD edges out GADS on LLaDA-V ($\text{CHAIR}_s$ 5.38 vs. 5.4, $\text{CHAIR}_i$ 3.81 vs. 4.0), but degrades sharply on MMaDA ($\text{CHAIR}_s$ 19.76 vs. our 15.4) and worsens $\text{CHAIR}_i$ on LaViDa (7.58

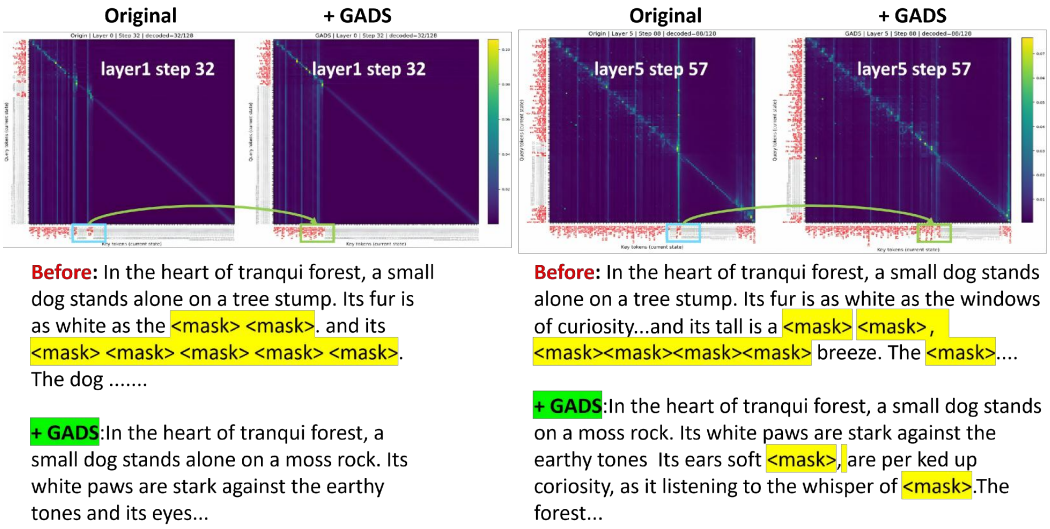


Figure 7: **Attention map from a GADS case study**. It can be observed that GADS alters the original decoding order, optimizing semantic coherence, red indicates tokens that have already been decoded.

vs. 4.4). GADS is the only method that stays within 0.1 of the best $\text{CHAIR}_s$ on every backbone while keeping $\text{CHAIR}_i$ at or near the minimum, and it is the strongest method that also reports long-form results, where CACD does not produce entries.

**Discriminative benchmarks.** On short-form tasks the three-phase pathology has less room to develop, and no method dominates every cell. GADS obtains the best MME cognition score on all three backbones (508.2 / 268.9 / 361.8) and the best GQA exact match on all three (53.94 / 43.14 / 56.89), and it never falls more than 0.2 points below the vanilla baseline on any discriminative metric. CoTA and CACD win isolated cells (MathVista on LLaDA-V and LaViDa respectively) but regress elsewhere (CACD drops ScienceQA on LLaDA-V from 89.74 to 79.52).

### 4.3 Case Study: GADS Improves Commitment Order

Figure 7 compares intermediate decoding states for the same prompt. Vanilla decoding leaves long `<mask>` spans in the middle despite having committed much of the opening and tail. With GADS, the middle region is more fully resolved, with shorter masked gaps and more continuous, coherent text. The accompanying attention maps show reduced local concentration. This example supports the trajectory analysis in Section 2: GADS shifts commitment toward previously delayed mid-response content, reducing the fragmented pattern of unresolved tokens.

## 5 Conclusion and Limitation

We identified a critical pathology in dMLLM decoding: local attention sinks artificially inflate structural-token confidence, forcing a rigid three-phase commit order that fragments mid-response semantics. To resolve this, we proposed **GADS**, a training-free mechanism that dynamically detects and suppresses sinks within the active window, flattening the local confidence landscape. By delaying premature structural commits, GADS restores a content-driven decoding trajectory, significantly improving long-form coherence and reducing hallucinations across multiple backbones. **Limitation**:While GADS is training-free and effective across multiple dMLLM backbones, several limitations remain: (1) **Short-text tasks**: The three-phase decoding pathology requires sufficient sequence length to manifest; gains on very short answers (e.g., single-word VQA) are marginal. (2) **Ultra-long contexts**: For multi-turn dialogues exceeding 10K tokens, cumulative sink dynamics may demand hierarchical windowing, which we leave for future work.

## 6 Declaration of the Use of Generative AI (Large Language Models)

Generative AI tools, including Grammarly and ChatGPT, were used solely for grammar checking and language polishing. All technical content, experimental design, data analysis, and conclusions were generated and verified exclusively by the human authors. The use of AI tools does not affect the originality or authorship of this work.

## 7 Ethics Statement

This work studies the decoding behavior of publicly available multimodal diffusion language models and proposes a training-free inference-time intervention. All models (LLaDA-V, MMaDA, LaViDa) and benchmarks (LLaVA-Bench, CHAIR, MathVista, ScienceQA, MME, GQA) are publicly released research artifacts used within their intended evaluation scope and licenses. We introduce no new datasets and collect no human-subject data beyond the small preference study described in the appendix, in which annotators evaluated model outputs that contain no personal or sensitive information. Our method aims to reduce semantic incoherence and hallucination in long-form generation; we do not identify negative societal impacts beyond those inherent to general-purpose generative models, such as the residual possibility of fluent but factually incorrect output, which our hallucination evaluations are designed to measure rather than eliminate entirely.

## 8 Reproducibility Statement

We take the following measures to support reproduction of our results. The full GADS pipeline is specified mathematically in Section 3. All backbones and benchmarks are publicly available; baseline methods (AdaBlock, Wavefront, CoTA, CACD) were run with their official implementations and recommended settings under identical decoding budgets. Commit-step instrumentation, the dataset-scale statistics of Section 2.1, and the paired bootstrap confidence intervals (10,000 resamples) are computed with fixed random seeds. Latency measurements are reported on a single NVIDIA A800 GPU. Anonymous source code, including the GADS implementation, evaluation scripts, and commit-step analysis tools, is provided to the area chair and will be made publicly available upon publication.

## References

Shaik Aman. Logicdiff: Logic-guided denoising improves reasoning in masked diffusion language models. *arXiv preprint arXiv:2603.26771*, 2026.

Tiwei Bie, Maosong Cao, Kun Chen, Lun Du, Mingliang Gong, Zhuochen Gong, Yanmei Gu, Jiaqi Hu, Zenan Huang, Zhenzhong Lan, et al. Llada2. 0: Scaling up diffusion language models to 100b. *arXiv preprint arXiv:2512.15745*, 2025.

Haolin Chen, Shiyu Wang, Can Qin, Bo Pang, Zuxin Liu, Jielin Qiu, Jianguo Zhang, Yingbo Zhou, Zeyuan Chen, Ran Xu, et al. Coda: Coding lm via diffusion adaptation. *arXiv preprint arXiv:2510.03270*, 2025a.

Tianlang Chen, Minkai Xu, Jure Leskovec, and Stefano Ermon. Rfg: Test-time scaling for diffusion large language model reasoning with reward-free guidance. *arXiv preprint arXiv:2509.25604*, 2025b.

Tianqi Chen, Shujian Zhang, and Mingyuan Zhou. Dlm-one: Diffusion language models for one-step sequence generation. *arXiv preprint arXiv:2506.00290*, 2025c.

Shuang Cheng, Yihan Bian, Dawei Liu, Linfeng Zhang, Qian Yao, Zhongbo Tian, Wenhai Wang, Qipeng Guo, Kai Chen, Biqing Qi, et al. Sdar: A synergistic diffusion-autoregression paradigm for scalable sequence generation. *arXiv preprint arXiv:2510.06303*, 2025.

Xin Dai, Pengcheng Huang, Zhenghao Liu, Shuo Wang, Yukun Yan, Chaojun Xiao, Yu Gu, Ge Yu, and Maosong Sun. Revealing the attention floating mechanism in masked diffusion models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 8266–8286, 2026.

Stipe Frkovic, Metod Jazbec, Dan Zhang, Christian A Naesseth, Ilija Bogunovic, and Eric Nalisnick. Re-evaluating confidence remasking in masked diffusion language models. *arXiv preprint arXiv:2606.12232*, 2026.

Yifeng Gao, Ziang Ji, Yuxuan Wang, Biqing Qi, Hanlin Xu, and Linfeng Zhang. Self speculative decoding for diffusion large language models, 2025. URL https://arxiv.org/abs/2510.04147.

Shansan Gong, Ruixiang Zhang, Huangjie Zheng, Jiatao Gu, Navdeep Jaitly, Lingpeng Kong, and Yizhe Zhang. Diffucoder: Understanding and improving masked diffusion models for code generation. *arXiv preprint arXiv:2506.20639*, 2025.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *The Thirteenth International Conference on Learning Representations*, 2025.

Xiaochuang Han, Sachin Kumar, and Yulia Tsvetkov. Ssd-lm: Semi-autoregressive simplex-based diffusion language model for text generation and modular control. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11575–11596, 2023.

Guangxin He, Shen Nie, Fengqi Zhu, Yuankang Zhao, Tianyi Bai, Ran Yan, Jie Fu, Chongxuan Li, and Binhang Yuan. Ultrallada: Scaling the context length to 128k for diffusion large language models. *arXiv preprint arXiv:2510.10481*, 2025.

Pengcheng Huang, Tianming Liu, Zhenghao Liu, Yukun Yan, Shuo Wang, Tong Xiao, Zulong Chen, and Maosong Sun. Empirical analysis of decoding biases in masked diffusion models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6853–6876, 2026.

Qidong Huang, Xiao wen Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Neng H. Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13418–13427, 2023.

Zemin Huang, Yuhang Wang, Zhiyang Chen, and Guo-Jun Qi. Don't settle too early: Self-reflective remasking for diffusion language models. *arXiv preprint arXiv:2509.23653*, 2025.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Xiangqi Jin, Yuxuan Wang, Yifeng Gao, Zichen Wen, Biqing Qi, Dongrui Liu, and Linfeng Zhang. Thinking inside the mask: In-place prompting in diffusion llms. *arXiv preprint arXiv:2508.10736*, 2025.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. 2025.

Samar Khanna, Siddhant Kharbanda, Shufan Li, Harshit Varma, Eric Wang, Sawyer Birnbaum, Ziyang Luo, Yanis Miraoui, Akash Palrecha, Stefano Ermon, et al. Mercury: Ultra-fast language models based on diffusion. *arXiv e-prints*, pp. arXiv–2506, 2025.

Younjoo Lee, Junghoo Lee, Seungkyun Dan, Jaiyoung Park, and Jung Ho Ahn. Dyllm: Efficient diffusion llm inference via saliency-based token selection and partial attention. *arXiv preprint arXiv:2603.08026*, 2026.

Jia-Nan Li, Jian Guan, Wei Wu, and Chongxuan Li. Refusion: A diffusion large language model with parallel autoregressive decoding. *arXiv preprint arXiv:2512.13586*, 2025a.

Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jiaqi Wang, and Dahua Lin. Beyond fixed: Training-free variable-length denoising for diffusion large language models. *arXiv preprint arXiv:2508.00819*, 2025b.

Shufan Li, Konstantinos Kallidromitis, Hritik Bansal, Akash Gokul, Yusuke Kato, Kazuki Kozuka, Jason Kuen, Zhe Lin, Kai-Wei Chang, and Aditya Grover. Lavida: A large diffusion language model for multimodal understanding, 2025c.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

Jingyu Liu, Xin Dong, Zhifan Ye, Rishabh Mehta, Yonggan Fu, Vartika Singh, Jan Kautz, Ce Zhang, and Pavlo Molchanov. Tidar: Think in diffusion, talk in autoregression. *arXiv preprint arXiv:2511.08923*, 2025.

Guanxi Lu, Hao Mark Chen, Yuto Karashima, Zhican Wang, Daichi Fujiki, and Hongxiang Fan. Adablock-dllm: Semantic-aware diffusion llm inference via adaptive block size. *arXiv preprint arXiv:2509.26432*, 2025.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024.

Zanlin Ni, Shenzhi Wang, Yang Yue, Tianyu Yu, Weilin Zhao, Yeguo Hua, Tianyi Chen, Jun Song, Cheng Yu, Bo Zheng, et al. The flexibility trap: Rethinking the value of arbitrary order in diffusion language models. *arXiv preprint arXiv:2601.15165*, 2026a.

Zanlin Ni, Shenzhi Wang, Yang Yue, Tianyu Yu, Weilin Zhao, Yeguo Hua, Tianyi Chen, Jun Song, Cheng Yu, Bo Zheng, et al. The flexibility trap: Rethinking the value of arbitrary order in diffusion language models. *arXiv preprint arXiv:2601.15165*, 2026b.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018.

Kevin Rojas, Jiahe Lin, Kashif Rasul, Anderson Schneider, Yuriy Nevmyvaka, Molei Tao, and Wei Deng. Improving reasoning for diffusion language models via group diffusion policy optimization. *arXiv preprint arXiv:2510.08554*, 2025.

Maximo Eduardo Rulli, Simone Petruzzi, Edoardo Michielon, Fabrizio Silvestri, Simone Scardapane, and Alessio Devoto. Attention sinks in diffusion language models. *arXiv preprint arXiv:2510.15731*, 2025.

Jucheng Shen and Yeonju Ro. Beyond static cutoffs: One-shot dynamic thresholding for diffusion language models. *arXiv preprint arXiv:2511.02077*, 2025.

Adi Shnaidman, Erin Feiglin, Osher Yaari, Efrat Mentel, Amit Levi, and Raz Lapid. Activation steering for masked diffusion language models. *arXiv preprint arXiv:2512.24143*, 2025.

Yingte Shu, Yuchuan Tian, Chao Xu, Yunhe Wang, and Hanting Chen. Deferred commitment decoding for diffusion language models with confidence-aware sliding windows. *arXiv preprint arXiv:2601.02076*, 2026a.

Yingte Shu, Yuchuan Tian, Chao Xu, Yunhe Wang, and Hanting Chen. Deferred commitment decoding for diffusion language models with confidence-aware sliding windows. *arXiv preprint arXiv:2601.02076*, 2026b.

Haocheng Sun, Cynthia Xin Wen, and Edward Hong Wang. Why mask diffusion does not work. *arXiv preprint arXiv:2510.03289*, 2025.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*, 2024.

Alexander Swerdlow, Mihir Prabhudesai, Siddharth Gandhi, Deepak Pathak, and Katerina Fragkiadaki. Unified multimodal discrete diffusion. *arXiv preprint arXiv:2503.20853*, 2025.

Feilong Tang, Zile Huang, Chengzhi Liu, Qiang Sun, Harry Yang, and Ser-Nam Lim. Intervening anchor token: Decoding strategy in alleviating hallucinations for MLLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=zGb4WgCW5i.

Lipeng Wan, Junjie Ma, Jianhui Gu, Zeyang Liu, Xuyang Lu, and Xuguang Lan. Geoblock: Inferring block granularity from dependency geometry in diffusion language models. *arXiv preprint arXiv:2603.26675*, 2026.

Guanghan Wang, Yair Schiff, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Remasking discrete diffusion models with inference-time scaling (2025). *arXiv preprint arXiv:2503.00307*.

Kangyu Wang, Zhiyun Jiang, Haibo Feng, Weijia Zhao, Lin Liu, Jianguo Li, Zhenzhong Lan, and Weiyao Lin. Creditdecoding: Accelerating parallel decoding in diffusion large language models with trace credits. *arXiv preprint arXiv:2510.06133*, 2025a.

Wen Wang, Bozhen Fang, Chenchen Jing, Yongliang Shen, Yangyi Shen, Qiuyu Wang, Hao Ouyang, Hao Chen, and Chunhua Shen. Time is a feature: Exploiting temporal dynamics in diffusion language models. *arXiv preprint arXiv:2508.09138*, 2025b.

Zeqing Wang, Gongfan Fang, Xinyin Ma, Xingyi Yang, and Xinchao Wang. Sparsed: Sparse attention for diffusion language models. *arXiv preprint arXiv:2509.24014*, 2025c.

Jinfeng Wei and Xiaofeng Zhang. Dopra: Decoding over-accumulation penalization and reallocation in specific weighting layer. *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024.

Yue Wu and Jian Huang. Stability-weighted decoding for diffusion language models. *arXiv preprint arXiv:2604.17068*, 2026.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. Mitigating object hallucination via concentric causal attention. *arXiv preprint arXiv:2410.15926*, 2024.

Zhen Xiong, Yujun Cai, Zhecheng Li, and Yiwei Wang. Unveiling the potential of diffusion large language model in controllable generation. *arXiv preprint arXiv:2507.04504*, 2025.

Haojin Yang, Rui Hu, Zequn Sun, Rui Zhou, Yujun Cai, and Yiwei Wang. Wavefrontdiffusion: Dynamic decoding schedule for improved reasoning. *arXiv preprint arXiv:2511.19473*, 2025a.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*, 2025b.

Lin Yao. Remask, don't replace: Token-to-mask refinement in masked diffusion language models. *arXiv preprint arXiv:2604.18738*, 2026.

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

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. Llada-v: Large language diffusion models with visual instruction tuning. *arXiv preprint arXiv:2505.16933*, 2025.

Runpeng Yu, Xinyin Ma, and Xinchao Wang. Dimple: Discrete diffusion multimodal large language model with parallel decoding. *arXiv preprint arXiv:2505.16990*, 2025.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. *arXiv preprint arXiv:2406.15765*, 2024.

Ruqian Zhang, Yijiao Zhang, Juan Shen, Zhongyi Zhu, and Annie Qu. Syntax-guided diffusion language models with user-integrated personalization. *arXiv preprint arXiv:2510.01028*, 2025a.

Siyue Zhang, Yilun Zhao, Liyuan Geng, Arman Cohan, Luu Anh Tuan, and Chen Zhao. Diffusion vs. autoregressive language models: A text embedding perspective. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4273–4303, 2025b.

Stephen Zhang, Mustafa Khan, and Vardan Papyan. Attention sinks: A 'catch, tag, release' mechanism for embeddings. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025c.

Xiaofeng Zhang, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. From redundancy to relevance: Enhancing explainability in multimodal large language models. *arXiv preprint arXiv:2406.06579*, 2024.

Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. Shallow focus, deep fixes: Enhancing shallow layers vision attention sinks to alleviate hallucination in lvlms. pp. 3512–3534, 2025d.

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Jiawei Cao, Hao Cheng, and Kaijie Wu. What drives attention sinks? a study of massive activations and rotational positional encoding in large vision–language models. *Information Processing & Management*, 63(2):104431, 2026a.

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

Zihou Zhang, Zheyong Xie, Li Zhong, Haifeng Liu, and Shaosheng Cao. One token is enough: Improving diffusion language models with a sink token. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 6479–6490, 2026c.

Zihou Zhang, Zheyong Xie, Li Zhong, Haifeng Liu, Yao Hu, and Shaosheng Cao. One token is enough: Improving diffusion language models with a sink token. *arXiv preprint arXiv:2601.19657*, 2026d.

Zihou Zhang, Zheyong Xie, Li Zhong, Haifeng Liu, Yao Hu, and Shaosheng Cao. One token is enough: Improving diffusion language models with a sink token. *arXiv preprint arXiv:2601.19657*, 2026e.

Hanyang Zhao, Dawen Liang, Wenpin Tang, David Yao, and Nathan Kallus. Diffpo: Training diffusion llms to reason fast and furious via reinforcement learning. *arXiv preprint arXiv:2510.02212*, 2025a.

Qiyan Zhao, Xiaofeng Zhang, yiheng Li, Yun Xing, Yuan Xiaosong, Feilong Tang, Sinan Fan, Xuhang Chen, Xuyao Zhang, and Dahan Wang. Mca-llava: Manhattan causal attention for reducing hallucination in large vision-language models. *Proceedings of the 32nd ACM International Conference on Multimedia*, 2025b.

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

Siyan Zhao, Devaansh Gupta, Qinqing Zheng, and Aditya Grover. d1: Scaling reasoning in diffusion large language models via reinforcement learning. *arXiv preprint arXiv:2504.12216*, 2025c.

Yikai Zhao, Qiyan Zhao, Jiaquan Zhang, Xiaofeng Zhang, Xiaosong Yuan, and Pengzhou Cheng. Context-aware cluster decoding: Semantic anchor-driven coherence in dmllms. *arXiv preprint arXiv:2608.22367*, 2026b.

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

Yuchen Zhu, Wei Guo, Jaemoo Choi, Petr Molodyk, Bo Yuan, Molei Tao, and Yongxin Chen. Enhancing reasoning for diffusion llms via distribution matching policy optimization. *arXiv preprint arXiv:2510.08233*, 2025b.

# A Appendix

Algorithm 1 is applied independently to each attention head and layer during a denoising forward pass. The head computes $\mathrm{Softmax}_{\mathrm{row}}(A')V$ over the full attention rows. Token commitments are then selected from all remaining masked generation positions according to predictive confidence and the current step's commitment budget. Sorting ties are resolved by position index, and gap ties by the smallest rank.

# B More experiment

Additional dataset-scale evidence is reported in the appendix: Fig. 8 quantifies the pathological commit order over the full evaluation set with paired bootstrap confidence intervals; Fig. 9 presents counterfactual controls isolating the local attention sink as the causal driver; Fig. 10 shows that the decoding overhead remains below 5% at $L{=}512$, confirming that improvements concentrate exactly where the pathology is active.

# C Related Work

**Diffusion Language Models and Their Multimodal Extensions.** Masked diffusion language models have rapidly scaled to competitive performance with autoregressive counterparts. LLaDA Nie et al. (2025) first demonstrated 8B-parameter masked diffusion LMs; Dream Ye et al. (2025) and Mercury Khanna et al. (2025) pushed inference speed and quality further; LLaDA-MoE Zhu et al. (2025a) introduced sparse mixture-of-experts into the diffusion backbone; and UltraLLaDA He et al. (2025) extended context length to 128K tokens. Beyond pure text, several works have integrated vision encoders into diffusion LMs to form dMLLMs: LLaDA-V You et al. (2025) applies visual instruction tuning, LaViDa Li et al. (2025c) couples a diffusion decoder with a large vision encoder, and MMaDA Yang et al. (2025b) unifies multimodal generation under a single diffusion framework. Hybrid paradigms that blend diffusion with autoregression have also emerged, including SDAR Cheng et al. (2025), TiDAR Liu et al. (2025), ReFusion Li et al. (2025a), and CoDA Chen et al. (2025a). Other extensions target controllable generation Zhang et al. (2025a); Xiong et al. (2025) and code synthesis Gong et al. (2025). Our work is orthogonal to all of the above: we do not modify the model architecture or training, but intervene at *inference time* on the attention map to correct a pathological decoding order that is common to all confidence-driven dMLLMs.

**Decoding Strategies for Diffusion LMs.** The standard block-diffusion sampler of LLaDA commits tokens via a fixed schedule within fixed-length blocks. A growing body of work has recognized the limitations of this rigid scheme and proposed improved decoding strategies along several

**Algorithm 1** Gap-Aware Dynamic Suppression (GADS)

**Require:** Attention logits $A \in \mathbb{R}^{n\times n}$; local window $\mathcal{W}$; $\varepsilon = 10^{-8}$
**Ensure:** Modified attention logits $A'$

$A' \leftarrow A$, $m \leftarrow |\mathcal{W}|$
**if** $m < 2$ **then**
  **return** $A'$
**end if**
*// Measure local attention concentration*
$P \leftarrow \mathrm{Softmax}_{\mathrm{row}}(A_{\mathcal{W},\mathcal{W}})$
**for** each $k \in \mathcal{W}$ **do**
  $e_k \leftarrow \frac{1}{m}\sum_{q\in\mathcal{W}} P_{q,k}$
**end for**
Sort energies: $\tilde{e}_1 \geq \cdots \geq \tilde{e}_m$
Let $\pi(r)$ be the position at rank $r$
$\mathcal{I} \leftarrow \{r \in \{1,\ldots,m-1\} : \tilde{e}_r \geq 2/m\}$
**if** $\mathcal{I} = \varnothing$ **then**
  **return** $A'$
**end if**
*// Locate the gap and check the mass condition*
$i^\star \leftarrow \arg\max_{r\in\mathcal{I}} \dfrac{\tilde{e}_r}{\tilde{e}_{r+1}+\varepsilon}$
$M \leftarrow \sum_{r=1}^{i^\star} \tilde{e}_r$
**if** $M \leq 2i^\star/m$ **or** $i^\star = 1$ **then**
  **return** $A'$
**end if**
*// Suppress columns above the reference*
$e_{\mathrm{ref}} \leftarrow \tilde{e}_{i^\star}$
**for** $r = 1,\ldots,i^\star - 1$ **do**
  $k \leftarrow \pi(r)$
  $\gamma_k \leftarrow \min(1, e_{\mathrm{ref}}/e_k)$
  **for** each $q \in \mathcal{W}$ **do**
    $A'_{q,k} \leftarrow A_{q,k} + \log\gamma_k$
  **end for**
**end for**
**return** $A'$

axes. *Adaptive block sizing*: AdaBlock Lu et al. (2025) uses delimiter look-ahead to cut blocks at semantic boundaries; GeoBlock Wan et al. (2026) infers block granularity from dependency geometry; and variable-length denoising Li et al. (2025b) removes the fixed-block constraint entirely. *Confidence-aware scheduling*: WavefrontDiffusion Yang et al. (2025a) assigns more denoising steps to uncertain regions; dynamic thresholding Shen & Ro (2025) replaces static top-$k$ with one-shot adaptive cutoffs; deferred commitment decoding Shu et al. (2026b;a) delays token commit using confidence-aware sliding windows; and CreditDecoding Wang et al. (2025a) allocates compute via trace credits. *Remasking and refinement*: self-reflective remasking Huang et al. (2025) allows the sampler to undo premature commits; token-to-mask refinement Yao (2026) replaces committed tokens with masks for a second denoising pass; and stability-weighted decoding Wu & Huang (2026) scores tokens by temporal stability across steps. *Scaling and reasoning*: inference-time scaling via remasking Wang et al., reward-free guidance Chen et al. (2025b), logic-guided denoising Aman (2026), and reinforcement-learning-based methods Zhu et al. (2025b); Rojas et al. (2025); Zhao et al. (2025a;c) improve reasoning quality by spending more compute on hard positions. Other directions include temporal feature exploitation Wang et al. (2025b), activation steering Shnaidman et al. (2025), saliency-based token selection Lee et al. (2026), in-place prompting Jin et al. (2025), and sink-token augmentation Zhang et al. (2026d). All of the above operate on the *scheduler axis* — deciding *when* and *how many* tokens to commit. GADS operates on the complementary *attention axis*: it changes *how* the denoiser attends at each step, suppressing the local sinks that inflate

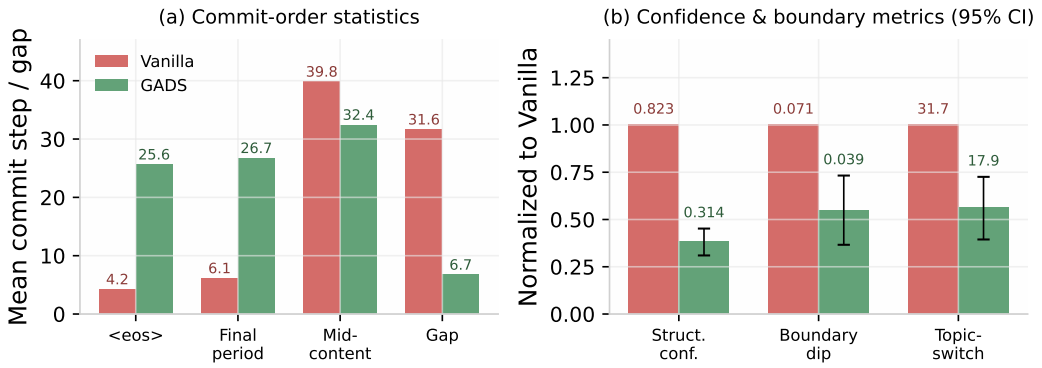


Figure 8: **Dataset-scale statistics of the pathological commit order** (LLaVA-Bench-60, global commit steps, paired 95% bootstrap CIs). (a) Commit-step locations: GADS delays the end-of-sequence anchor ($4.2 \rightarrow 25.6$) and the final period ($6.1 \rightarrow 26.7$) to the end of generation, closing the mid-content gap from 31.6 to 6.7 steps (CI $[-29.7, -20.4]$). (b) Normalized pathology metrics: mid-generation confidence dip ($0.071 \rightarrow 0.039$), commit confidence ($0.823 \rightarrow 0.314$, CI of the difference $[-0.568, -0.451]$), and topic-switch rate ($31.7\% \rightarrow 17.9\%$) all drop markedly, while the relative response length is unchanged ($1.000 \rightarrow 1.012$), confirming that the pathology is a population-level phenomenon rather than an isolated case, and that GADS removes it without truncating responses.

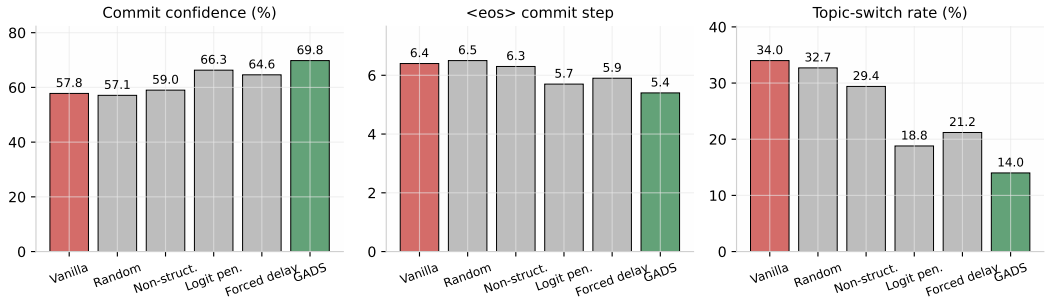


Figure 9: **Counterfactual controls with matched bias budget.** Suppressing randomly chosen or non-structural columns leaves all three metrics at the Vanilla level, excluding a generic perturbation effect. Explicitly penalizing end-of-sequence logits (Logit pen.) or forcing delayed terminator commits (Forced delay) reduces the topic-switch rate but does not recover commit quality, and both fall short of GADS (69.8 / 5.4 / 14.0%). The local attention-sink itself is therefore the causal driver of the pathological commit order.

structural-token confidence and thereby reshaping the confidence landscape that all schedulers consume. GADS is therefore orthogonal to, and composable with, any of the above strategies.

**Attention Sinks in AR-LLMs and MLLMs.** Attention sinks — tokens that absorb disproportionate softmax mass — were first characterized in autoregressive LLMs by StreamingLLM Xiao et al. (2023), which showed that the initial token acts as a "no-op" repository for surplus attention. Subsequent work extended the analysis to punctuation tokens Yu et al. (2024), image tokens in multimodal LLMs Kang et al. (2025), and the interplay between sinks and massive activations Sun et al. (2024). Mitigation strategies for AR models include penalty-based methods such as OPERA Huang et al. (2023) and DOPRA Wei & Zhang (2024), attention redirection approaches like COTA Zhang et al. (2025d; 2026a); Zhao et al. (2026a); Zhang et al. (2026b) and Intervening Anchor Token Tang et al. (2025), and causal attention modifications such as CCA Xing et al. (2024) and MCA Zhao et al. (2025b). However, as the concurrent study of Zhang et al. (2026d) have shown, the sink phenomenon in diffusion LMs is fundamentally different: sinks are *dynamic* (migrating from end-of-sequence to period tokens as decoding progresses) rather than static, and the bidirectional attention of dMLLMs invalidates the causal assumptions underlying all AR-based mitigations. In this work we show that

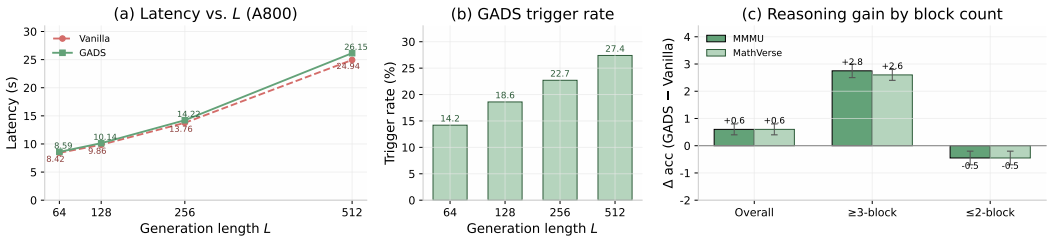


Figure 10: **Efficiency and stratified reasoning gains.** (a) Wall-clock latency of Vanilla and GADS across generation lengths on a single A800; the overhead stays below 5% at $L$=512 ($+2.1\%/+2.9\%/+3.4\%/+4.9\%$ for $L$=64/128/256/512). (b) The intervention is triggered on only 14.2–27.4% of steps, since GADS acts exclusively within the detected local sink window. (c) Accuracy change on MMMU and MathVerse, stratified by the number of decoded blocks (error bars span the observed range across the two backbones). The overall change is small ($+0.4$ to $+0.8$ points) and decomposes cleanly: gains concentrate on multi-block responses ($\geq$3 blocks, $+2.4$ to $+3.0$), where the pathological commit order is active, while short responses ($\leq$2 blocks) show a negligible negative shift ($-0.2$ to $-0.7$), consistent with our scope analysis that the pathology does not arise in short-form answers.

the right intervention scope is the *local decoding window*, and that suppressing local sinks not only improves attention quality but directly *reorders the confidence-driven decoding trajectory*, addressing the root cause of mid-response semantic incoherence.

**Attention Mechanisms in Diffusion-based LLMs.** Diffusion models have recently emerged as a powerful paradigm for text generation, offering a compelling alternative to autoregressive approaches by employing iterative refinement processes Wang et al. (2025c); Chen et al. (2025c); Zhang et al. (2025b). Diffusion models and semi-autoregressive hybrids Sun et al. (2025); Han et al. (2023) successfully scale these processes to handle complex reasoning tasks by utilizing bidirectional attention mechanisms that model the entire sequence simultaneously Gao et al. (2025). Despite these architectural shifts, recent empirical studies indicate that the attention sink persists in diffusion-based LLMs, mirroring the semantic distortion seen in autoregressive models Rulli et al. (2025). However, the standard mitigation strategies designed for autoregressive models are structurally incompatible with this paradigm. These existing methods rely heavily on the causal masking assumption, where tokens only attend to predecessors, allowing for the use of static sinks as fixed anchors for KV caching Zhang et al. (2025c); Gu et al. (2025). In contrast, diffusion models employ a bidirectional decoding mechanism that allows every token to attend to both past and future contexts simultaneously at every denoising step Sun et al. (2025); Zhang et al. (2025b). This fundamental difference in visibility, combined with the fact that sink positions in diffusion models are dynamic rather than static Rulli et al. (2025), renders causal-based sink retention strategies ineffective. Consequently, mitigating attention sinks in diffusion models requires a novel approach that reconciles the presence of sinks with the global visibility and dynamic nature of bidirectional attention.

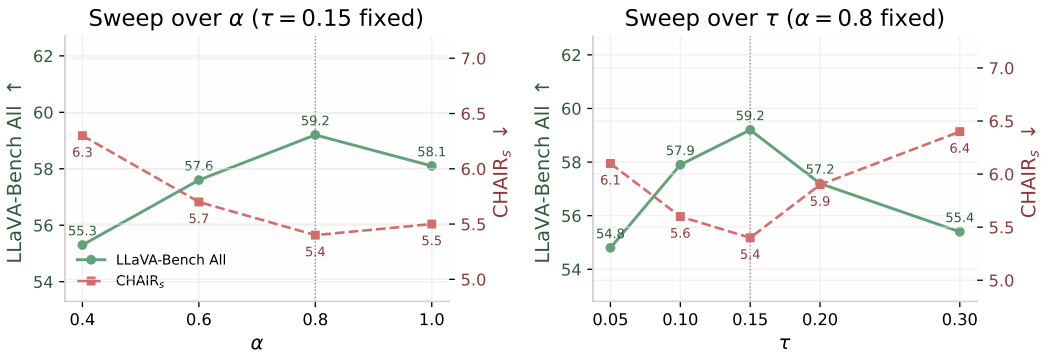


Figure 11: **Robustness to the fixed constants.** LLaVA-Bench All (green, left axis) and CHAIR$_s$ (red, right axis) when perturbing $\alpha$ and $\tau$ around their fixed values (dotted line). Performance is stable within $\alpha \in [0.6, 1.0]$ and $\tau \in [0.10, 0.20]$.