# REGULATING DYNAMIC ATTENTION ALLOCATION IN DIFFUSION MULTIMODAL LARGE LANGUAGE MODELS

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

## ABSTRACT

Diffusion multimodal large language models (dMLLMs) generate text by iteratively denoising a masked response, and at every step, image and output tokens compete for the same softmax attention budget. We find that a small set of output tokens, which we term *output sinks*, absorbs a disproportionate share of this budget. These sinks shift across positions between steps and claim a growing share of attention as denoising proceeds, while the share allocated to the image declines. Reduced visual attention can limit the visual evidence available for generation and contribute to hallucination. Most sinks are low-semantic tokens (masks, punctuation, and function words), but some are content words. Motivated by these findings, we propose **Elbow Proportional Suppression (EPS)**, a training-free attention intervention. For each head at each step, EPS detects sinks with a parameter-free elbow criterion and attenuates only the low-semantic ones relative to an adaptively selected anchor token via an additive logit shift. We prove that EPS redistributes the suppressed attention proportionally to all remaining keys; empirically, most of it is reallocated to the image. Controlled interventions show that both visual reallocation and the choice of suppressed tokens reduce hallucination. Across three dMLLMs, EPS reduces object hallucination (by up to 8.0 points in $\text{CHAIR}_S$) and improves description quality and multimodal reasoning. On LLaDA-V and LaViDA, EPS achieves the best results on 24 of 26 metrics against five training-free baselines. Code and analysis tools will be released.

## 1 INTRODUCTION

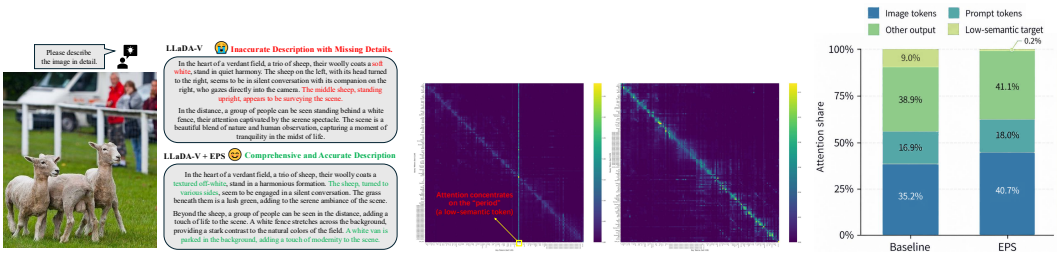
Diffusion language models (dLLMs) generate text by iteratively denoising a masked sequence with bidirectional attention (Nie et al., 2025; Ye et al., 2025), and diffusion multimodal large language models (dMLLMs) extend them to multimodal inputs by incorporating a visual encoder (You et al., 2025; Yang et al., 2025; Li et al., 2025). At each denoising step, a dMLLM processes the entire response in parallel, and every output position, whether masked or filled, attends to the image. Since attention is normalized by a single softmax over all keys, image and output tokens compete for a shared attention budget. The allocation of this budget thus directly shapes how much visual evidence each denoising step can exploit.

A key factor in this allocation is the *attention sink*, a token that attracts disproportionate attention from many queries. In autoregressive (AR) models, sinks typically form at the initial token or at recurring delimiters (Xiao et al., 2024; Sun et al., 2024), and in AR multimodal models they have been linked to weak visual grounding and hallucination (Kang et al., 2025; Tang et al., 2025b; Huang et al., 2024). Existing mitigation methods accordingly exploit these fixed positions or the causal decoding order. In text-only dLLMs, sinks instead shift across positions as denoising proceeds (Rulli et al., 2026; Zhang et al., 2026c; Dai et al., 2026), but these models lack visual inputs, leaving the effect of sinks on visual grounding unexamined. Sinks in dMLLMs therefore remain unexplored, even though dMLLMs are precisely where sinks and the image compete for the same budget: attention absorbed by output sinks leaves less probability mass available to other regions, including the image. This raises two questions: *Do output-side sinks divert attention from visual evidence during denoising? If so, does returning this attention to the image reduce hallucination?*

To answer the first question, we analyze 8.2M head–layer–step attention records from LLaDA-V over 500 images and all 128 denoising steps (Section 2), which yields three findings. **(i) Output**



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.




(a) Image and descriptions (b) Baseline and EPS attention (c) Regional allocation

Figure 1: **Attention regulation and generation in LLaDA-V.** (a) EPS revises the sheep-orientation claim and adds the background white van. (b) Baseline (left) and EPS (right) attention maps use separate color scales. EPS reduces vertical column concentration. (c) EPS reduces the share of low-semantic targets $\mathcal{S}$ and increases image share. "Other output" excludes $\mathcal{S}$.

**sinks concentrate attention and drift.** At nearly every step, a small set of output tokens, which we term *output sinks*, absorbs a disproportionate share of the attention to the response, with the top one receiving 9–12$\times$ the uniform share. Unlike AR sinks, they are not fixed: the top position changes in 13.5–23.7% of consecutive steps. **(ii) Output sinks gain attention as image share declines.** As denoising proceeds, their share of the total budget rises from 15.5% to 20.3%, while the image share falls from 40.8% to 33.4%. This decline in image attention can limit access to visual evidence, particularly in late denoising, when most tokens are committed. They persist even with blank or noise images, indicating that they arise independently of visual content. **(iii) Output sinks are mostly, but not only, low-semantic.** Masks, punctuation, and function words constitute about two-thirds of output sinks, yet content words receive up to half of the sink attention in some layers and phases.

Answering the second question requires an intervention that reallocates attention from output sinks toward the image without disrupting generation. Guided by the three findings, we design **Elbow Proportional Suppression (EPS)**, a training-free method with three components. Since sinks drift (i), EPS detects them anew for every head and step with a parameter-free elbow criterion on the received attention. Since some sinks are content words (iii), it suppresses only the low-semantic ones. Since sinks compete with the image for a shared budget (ii), it attenuates them via an additive logit shift calibrated to an anchor, the first token below the elbow. We prove that this shift redistributes the suppressed attention proportionally to all remaining keys, including every image token. EPS applies to masked and filled positions alike and leaves the decoding schedule unchanged.

With EPS, we answer the second question affirmatively: reallocating attention from output sinks to the image reduces hallucination. In LLaDA-V, image tokens receive 59.7–67.2% of the released attention, and phase-averaged image share increases by 6.7–6.9 percentage points over the baseline (Figure 4). Controlled interventions attribute the hallucination reduction to both where the released attention goes and which tokens release it: restricting immediate visual reallocation forfeits about half of the $\text{CHAIR}_S$ reduction and most of the recall gain, and under a common reduced removal budget, suppressing EPS-selected tokens outperforms suppressing random or attention-matched tokens. Across three dMLLMs, EPS reduces object hallucination by up to 8.0 $\text{CHAIR}_S$ points and improves description quality and multimodal reasoning. On LLaDA-V and LaViDA, EPS achieves the best results on 24 of 26 metrics against five training-free baselines.

Our contributions are threefold:

- **Analysis.** To our knowledge, we present the first systematic analysis of attention sinks in dMLLMs, revealing that output sinks concentrate attention, drift across denoising steps, claim an increasing share of attention as image share declines, and are mostly, but not only, low-semantic.
- **Method.** We propose EPS, a training-free, step-adaptive intervention that suppresses low-semantic output sinks and provably redistributes their attention proportionally to all remaining keys.
- **Evidence.** Controlled interventions show that both visual reallocation and the choice of suppressed tokens reduce hallucination, and EPS yields gains on seven benchmarks across three dMLLMs.[1]

[1]Appendix A discusses related work.

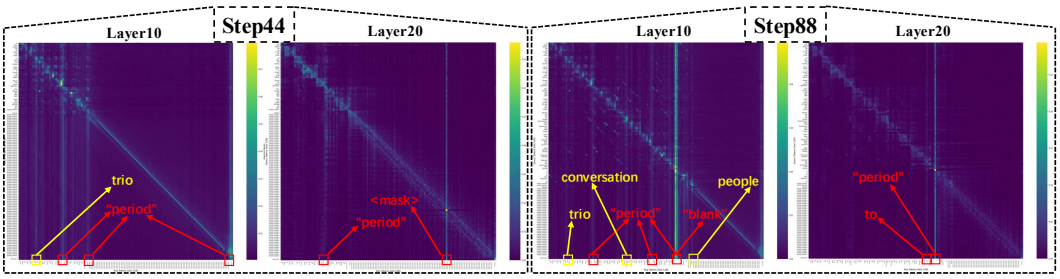



Figure 2: **Output concentration across steps and layers in LLaDA-V.** Red and yellow annotations mark low-semantic and content tokens, respectively, in one example.

## 2 Dynamic Concentration and Multimodal Allocation

To characterize output sinks, we use an elbow criterion to identify a set $\mathcal{E}$ of highly attended output columns. We analyze their positions, token identities, and coverage across queries. This set includes both low-semantic and content tokens; EPS intervenes only on its low-semantic subset $\mathcal{S}$ (Section 3).

### 2.1 Attention across Denoising Steps and Layers

Figure 2 shows one LLaDA-V example at steps 44 and 88 and layers 10 and 20. Concentrated columns occur at different positions and include both low-semantic tokens (red: full stops, labeled "period", `<mask>`, and `to`) and content words (yellow). These views motivate separate population measurements of concentration, mobility, and token composition in Sections 2.4–2.5.

### 2.2 Region-Level Attention Allocation

Consider one attention head at a fixed layer and denoising step. Let $P \in \mathbb{R}^{M \times L}$ contain attention probabilities from $M$ output queries to $L$ visible keys. Output slots include masked and filled positions. Partition the keys into image $\mathcal{O}_{\text{img}}$, prompt/system $\mathcal{O}_{\text{prm}}$, and output $\mathcal{O}_{\text{gen}}$ regions. For any region $\mathcal{R}$, define

$$A_i(\mathcal{R}) = \sum_{j \in \mathcal{R}} P_{ij}, \qquad A_i(\mathcal{O}_{\text{img}}) + A_i(\mathcal{O}_{\text{prm}}) + A_i(\mathcal{O}_{\text{gen}}) = 1. \tag{1}$$

For the elbow set $\mathcal{E} \subset \mathcal{O}_{\text{gen}}$ defined in Section 3,

$$A_i(\mathcal{O}_{\text{img}}) \leq 1 - A_i(\mathcal{E}). \tag{2}$$

We call $A_i(\mathcal{E})$ the *elbow-set share*, measured against the full visible-key budget. Composition, coverage, and matching statistics describe $\mathcal{E}$ before eligibility filtering. Equation 2 bounds image allocation; Section 4.4 tests the effects of suppressing its low-semantic subset.

### 2.3 Concentration and Mobility

The received mass of output key $j$ and its fraction within the output region are

$$v_j = M^{-1} \sum_{i=1}^{M} P_{ij}, \qquad \widetilde{v}_j = \frac{v_j}{\sum_{r \in \mathcal{O}_{\text{gen}}} v_r}. \tag{3}$$

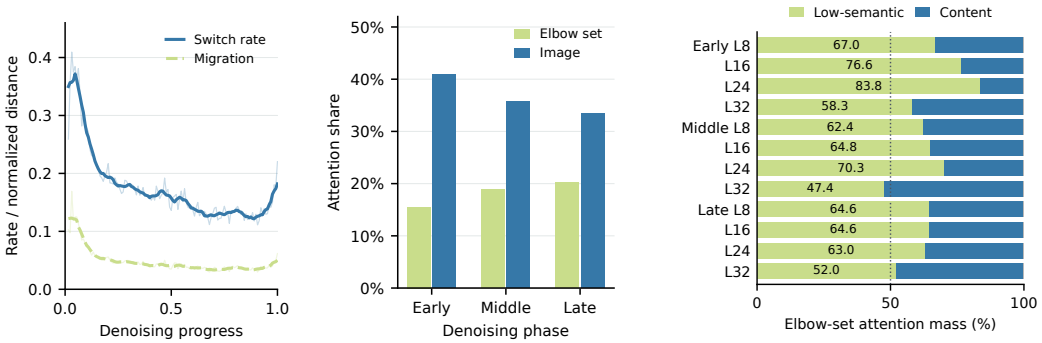
Top-1 and Top-5 fractions of $\widetilde{v}$, normalized entropy, and a median/MAD indicator measure within-output concentration. Elbow-set share instead uses the multimodal budget. Event frequency averages the indicator within each image, then across images; prevalence counts images with at least one event (Appendix C).

Let $j_t^*$ be the highest-mass output slot at step $t$, and $N = |\mathcal{O}_{\text{gen}}|$. We measure positional mobility by adjacent-step switching and distance $|j_t^* - j_{t-1}^*|/(N-1)$, separately from token identity.

For token composition, we distinguish *low-semantic* pieces from *content* pieces. Masks, end-of-sequence markers, punctuation, and function words belong to the first group. Other alphanumeric

Table 1: **Population concentration in LLaDA-V.** Sample-first means for the population study in Section 2.4: records are averaged within each image and phase, then across images. Measurements cover 128 output positions, 128 denoising invocations, and layers 8, 16, 24, and 32. Top-$k$ and entropy are normalized within output keys. Elbow-set share is $A(\mathcal{E})$ before eligibility filtering; elbow-set and image shares use the full attention budget.

| Phase (steps) | Event freq. | Top-1 | Top-5 | Entropy | Elbow-set share | Image share |
|---|---|---|---|---|---|---|
| Early (1–42) | 0.962 | 0.092 | 0.235 | 0.895 | 0.155 | 0.408 |
| Middle (43–85) | 0.943 | 0.073 | 0.208 | 0.905 | 0.189 | 0.358 |
| Late (86–128) | 0.946 | 0.073 | 0.197 | 0.908 | 0.203 | 0.334 |




(a) Position mobility (b) Regional allocation (c) Elbow-set composition

Figure 3: **Position dynamics, allocation, and composition.** (a) Dominant-slot switching and migration; light curves are step-level means and heavy curves use a centered seven-step moving average. (b) Phase-averaged elbow-set share $A(\mathcal{E})$ and image share of total attention. (c) Low-semantic and content shares of attention mass within $\mathcal{E}$ across phases and layers.

pieces belong to the second. The same rule applies at every output position, including the first and last slots. Unrecognized pieces remain unclassified. Appendix C specifies the lexical rules.

### 2.4 Population Findings

LLaDA-V generates 128 output slots over 128 denoising steps for 500 images. We record all 32 heads at layers 8, 16, 24, and 32. We use three fixed phases: early (steps 1–42), middle (43–85), and late (86–128). Statistics average records within each image and phase, then average images. Table 1 summarizes all three phases. The Top-1 share is 0.073–0.092, against a uniform share of 0.0078. Top-5 is 0.197–0.235, against 0.0391. Every image has concentration events in each phase. Their within-image frequencies average 0.962, 0.943, and 0.946. A small set of positions therefore receives much of the output attention throughout denoising.

**Mobility decreases while concentration persists.** Figure 3(a) shows declining mobility: phase-averaged switching falls from 0.237 to 0.135, and migration from 0.068 to 0.037. The step-level curves rebound near the end, but later phases are less mobile on average. Concentration therefore persists even as dominant positions become more stable. Appendix D reports the dominant-position trajectory.

**Regional allocation follows a different trajectory.** Figure 3(b) shows elbow-set share rising from 0.155 to 0.203. Image share falls from 0.408 to 0.334. Meanwhile, within-output Top-1 and Top-5 concentration decline modestly. The elbow set can therefore take more of the total budget as its largest columns become less dominant within output attention. Regional shares capture this change in visual access. Appendix D gives the full position summaries.

### 2.5 Elbow-Set Composition and Query Coverage

We examine $\mathcal{E}$ on a 200-image sample (Appendix E). Measurements cover layers 8, 16, 24, and 32 over all 128 steps. Figure 3(c) shows its token composition across the 12 phase/layer strata.

**Composition and enrichment differ.** Low-semantic pieces form 51.9–85.0% of $\mathcal{E}$. They receive 47.4–83.8% of elbow-set mass, exceeding half in 11 of 12 strata. Across strata, they average 65.7% of members and 64.6% of mass. They also occupy 75.8% of available slots. Their majority in $\mathcal{E}$ therefore differs from enrichment relative to availability. Punctuation is strongly enriched in several strata. Content pieces receive up to 52.4% of elbow-set mass. EPS applies the low-semantic filter after detection. Table 5 gives the full ranges.

**Concentration extends beyond detection queries.** We detect $\mathcal{E}$ on one query fold and evaluate its columns on the other, excluding self-attention. Columns in $\mathcal{E}$ exceed uniform non-self output attention on 35.5–77.8% of held-out queries. They rank in the query's Top-5 on 13.6–31.5%. These are means over columns within each phase/layer/fold stratum. Appendix E reports matching controls outside $\mathcal{E}$ and their coverage. Section 4.4 tests suppression and redistribution using $\mathcal{S}$.

### 2.6 From Allocation Patterns to Quality Interventions

We test whether suppressing $\mathcal{S} = \mathcal{E} \cap \mathcal{L}_t$, where $\mathcal{L}_t$ contains currently low-semantic positions, improves image-conditioned generation. Removed mass and redistribution are measured on $\mathcal{S}$. Image-constrained EPS tests visual reallocation, common-budget controls test target selection, and category ablations compare the complete elbow set with its lexical subsets (Section 4.4).

## 3 Elbow Proportional Suppression

EPS detects an elbow in received attention mass and uses the cutoff token as an anchor. It then selects low-semantic columns for suppression. Logits at content positions and the anchor remain unchanged.

### 3.1 Probability Aggregation and Elbow Detection

For output queries $Q \in \mathbb{R}^{M\times d}$ and visible keys $K \in \mathbb{R}^{L\times d}$, let $Z = QK^\top/\sqrt{d} + B$. Here $B$ contains the attention mask and additive bias. Compute $P = \mathrm{softmax}(Z)$ over visible keys, then $v_j = M^{-1}\sum_i P_{ij}$ for output keys. Detection compares output columns while retaining the original cross-region normalization.

Sort the masses as $\hat{v}_1 \geq \cdots \geq \hat{v}_N$, with $\sigma(k)$ mapping rank to key index. Normalize both axes and measure distance to the line $x + y = 1$:

$$x_k = \frac{k-1}{N-1}, \quad y_k = \frac{\hat{v}_k - \hat{v}_N}{\hat{v}_1 - \hat{v}_N + \epsilon}, \quad d_k = \frac{|x_k + y_k - 1|}{\sqrt{2}}. \tag{4}$$

The cutoff is $k^* = \max\{2, \min\{N, \arg\max_k d_k\}\}$. The **Elbow Anchor** is $a = \sigma(k^*)$, and $\mathcal{E} = \{\sigma(k) : 1 \leq k < k^*\}$ is the unfiltered elbow set. For $N \geq 2$, this leaves the anchor outside $\mathcal{E}$. The elbow rule also operates on flat distributions and is distinct from the concentration-event test. For $N < 2$, EPS is the identity.

### 3.2 Low-Semantic Eligibility

Let $\mathcal{L}_t$ contain confirmed low-semantic output positions at the current denoising step. Mask and end-of-sequence tokens are identified by their tokenizer IDs. For ordinary pieces, tokenizer whitespace markers are removed before matching punctuation or the function-word list in Appendix C. Content, unknown pieces, and other special tokens are excluded. Eligibility depends on token identity, not output position. The suppressed set is

$$\mathcal{S} = \mathcal{E} \cap \mathcal{L}_t. \tag{5}$$

The anchor $a$ is unchanged. Filtering covers all of $\mathcal{E}$, including eligible positions ranked after content tokens. An empty $\mathcal{S}$ leaves attention unchanged. Setting $\mathcal{L}_t = \mathcal{O}_{\mathrm{gen}}$ gives the complete-set variant used in the category ablation.

**Algorithm 1** Elbow Proportional Suppression for one head and invocation

**Input:** Output-query logits $Z$; output keys $\mathcal{O}_{\text{gen}}$; eligible positions $\mathcal{L}_t$; $\alpha \in (0, 1]$; $\epsilon = \delta = 10^{-12}$
**Output:** Adjusted logits $Z'$ for the standard attention value aggregation

$N \leftarrow |\mathcal{O}_{\text{gen}}|$; $Z' \leftarrow Z$
**if** $N < 2$ **then**
  **return** $Z'$
**end if**
$P \leftarrow \mathrm{softmax}(Z)$ over all visible keys, computed in float32
$v_j \leftarrow M^{-1} \sum_i P_{ij}$ for every $j \in \mathcal{O}_{\text{gen}}$
$(\hat{v}, \sigma) \leftarrow \mathrm{sort}(v, \text{descending})$
For each rank $k$, compute $x_k$, $y_k$, and $d_k$ as defined above
$k^* \leftarrow \mathrm{clip}(\arg\max_k d_k, 2, N)$; $a \leftarrow \sigma(k^*)$
$\mathcal{E} \leftarrow \{\sigma(k) : 1 \le k < k^*\}$
$\mathcal{S} \leftarrow \mathcal{E} \cap \mathcal{L}_t$
**if** $\mathcal{S} = \varnothing$ **then**
  **return** $Z'$
**end if**
**for** $i = 1, \ldots, M$ **do**
  $m_i \leftarrow \max_{j \in \mathcal{S}} P_{ij}$
  $\gamma_i \leftarrow \min\{1, \alpha P_{ia}/(m_i + \epsilon)\}$
  $\bar{\gamma}_i \leftarrow \max\{\delta, \gamma_i\}$
  $Z'_{ij} \leftarrow Z_{ij} + \log \bar{\gamma}_i$ for every $j \in \mathcal{S}$, in logit dtype
**end for**
**return** $Z'$ for the attention update $\mathrm{softmax}(Z')V$

### 3.3 Row-Wise Proportional Suppression

For a nonempty $\mathcal{S}$ and each output query, define

$$m_i = \max_{j \in \mathcal{S}} P_{ij}, \qquad \gamma_i = \min\left\{1, \frac{\alpha P_{ia}}{m_i + \epsilon}\right\}, \qquad 0 < \alpha \le 1. \tag{6}$$

The maximum and suppression factor are recomputed on the filtered set. EPS adds $\log \gamma_i$ to its logits in row $i$ and leaves other logits unchanged. The shift is non-positive and preserves relative probabilities within the suppressed set. Smaller $\alpha$ gives stronger suppression.

### 3.4 Probability Guarantees

Assume finite logits on the relevant visible keys and write $P' = \mathrm{softmax}(Z')$. The result applies to either eligibility configuration. For an empty suppressed set, the update is the identity.

**Proposition 3.1** (EPS allocation guarantees)**.** *For nonempty $\mathcal{S}$, in exact arithmetic without the numerical floor, every $j \in \mathcal{S}$ satisfies*

$$P'_{ij} \le \alpha P'_{ia}. \tag{7}$$

*Let $s_i = \sum_{j \in \mathcal{S}} P_{ij}$. For every non-suppressed key $r \notin \mathcal{S}$,*

$$P'_{ir} = \frac{P_{ir}}{1 - (1 - \gamma_i) s_i} \ge P_{ir}. \tag{8}$$

*The increase is strict when $P_{ir} > 0$, $s_i > 0$, and $\gamma_i < 1$.*

Appendix H gives the proofs and numerical diagnostics. The floor $\bar{\gamma}_i = \max(\delta, \gamma_i)$ preserves non-suppressed monotonicity in exact arithmetic but can loosen the anchor bound.

**Visual redistribution.** Image keys are always non-suppressed. Summing Equation 8 gives, at the same attention state,

$$A'_i(\mathcal{O}_{\text{img}}) = \frac{A_i(\mathcal{O}_{\text{img}})}{1 - (1 - \gamma_i) s_i} \ge A_i(\mathcal{O}_{\text{img}}). \tag{9}$$

Image, prompt, and other non-suppressed keys all gain probability. The image-constrained control tests the quality benefit of allowing visual reallocation.

Table 2: **Downstream evaluation on LLaDA-V and LaViDA.** Bold and underline indicate the best and second-best listed scores within each backbone, including ties. LLaVA-Bench All is an independently scored dimension.

| Method | CHAIR | | | LLaVA-Bench | | | | POPE | | MathVista | SQA | MMMU | MMStar |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $C_S \downarrow$ | $C_I \downarrow$ | Recall↑ | Conv↑ | Details↑ | Complex↑ | All↑ | Acc↑ | F1↑ | Score↑ | Score↑ | Score↑ | Score↑ |
| LLaDA-V | 30.0 | 7.6 | 66.5 | 72.2 | 62.4 | 71.8 | 70.8 | 85.62 | 85.70 | 52.5 | 85.9 | 48.6 | 57.1 |
| + OPERA | 32.5 | 7.8 | 58.5 | 67.5 | 60.1 | 70.5 | 68.2 | 83.82 | 84.95 | 53.5 | 86.8 | 45.4 | 56.0 |
| + Farsight | 28.0 | 7.7 | 54.6 | 66.7 | 63.3 | 71.6 | 68.0 | 86.48 | 86.75 | 52.1 | 86.2 | 48.8 | 55.7 |
| + MCA | 25.5 | **5.8** | 67.6 | 69.9 | 59.8 | 73.1 | 69.6 | 87.10 | 87.55 | 53.8 | 87.1 | 48.7 | 58.2 |
| + AdaBlock | 24.0 | 6.5 | 66.8 | 69.9 | 61.7 | 72.1 | 70.7 | 87.02 | 87.45 | 54.3 | 89.9 | 49.7 | 60.1 |
| + Wavefront | 25.0 | 6.4 | 67.1 | 73.0 | 63.3 | **74.8** | 71.4 | 85.33 | 84.53 | 54.7 | 90.1 | 49.2 | 58.3 |
| **+ EPS (Ours)** | **23.0** | 6.0 | **69.2** | **75.2** | **67.2** | 74.7 | **72.1** | **87.35** | **87.76** | **57.5** | **91.7** | **50.8** | **61.5** |
| LaViDA | 33.0 | 9.3 | 57.6 | 73.0 | 63.5 | 75.5 | 71.6 | 83.80 | 82.86 | 44.8 | 72.3 | 40.8 | 61.8 |
| + OPERA | 35.0 | 9.7 | 51.2 | 69.8 | 61.2 | 72.3 | 69.5 | 82.36 | 81.71 | 43.5 | 70.4 | 41.2 | 59.9 |
| + Farsight | 31.5 | 9.0 | 50.3 | 70.6 | 64.6 | 73.8 | 70.9 | 84.62 | 83.94 | 42.7 | 72.8 | 38.1 | 56.8 |
| + MCA | 28.0 | 8.9 | 57.8 | 71.4 | 62.7 | 74.9 | 70.7 | 85.41 | 85.08 | 43.9 | 73.2 | 42.1 | 60.6 |
| + AdaBlock | 26.5 | 8.2 | 58.5 | 72.1 | 65.2 | 74.1 | 72.9 | 85.26 | 84.79 | 44.6 | 74.1 | 43.5 | 62.2 |
| + Wavefront | 27.0 | 7.8 | 59.1 | 75.3 | 63.0 | 76.2 | 71.0 | 84.11 | 83.25 | 45.2 | 73.0 | 44.1 | 62.1 |
| **+ EPS (Ours)** | **25.0** | **7.4** | **59.8** | **76.5** | **68.1** | **77.5** | **73.6** | **85.77** | **85.51** | **45.6** | **74.7** | **45.3** | **63.5** |

# 4 EXPERIMENTS

## 4.1 EXPERIMENTAL SETUP

We evaluate LLaDA-V (You et al., 2025), LaViDA (Li et al., 2025), and LaViDA-Dream on seven benchmarks: CHAIR (Rohrbach et al., 2018) and POPE (Li et al., 2023) for hallucination, LLaVA-Bench (Liu et al., 2023) for description quality, MathVista (Lu et al., 2024) and ScienceQA (SQA) (Lu et al., 2022) for reasoning, and MMMU (Yue et al., 2024) and MMStar (Chen et al., 2024) for multimodal understanding. We compare EPS with five training-free methods: OPERA (Huang et al., 2024), Farsight (Tang et al., 2025b), MCA (Zhao et al., 2025), AdaBlock (Lu et al., 2026), and Wavefront (Yang et al., 2026). All benchmarks use a maximum output length of 128 tokens. EPS preserves the underlying decoding schedule and uses $\alpha = 0.8$ unless stated otherwise. Appendix B provides detailed evaluation settings.

## 4.2 MAIN RESULTS

On LLaDA-V, EPS leads on eleven of thirteen metrics (Table 2). CHAIR $C_S$ falls from 30.0 to 23.0, and Recall rises from 66.5 to 69.2. The exceptions are $C_I$ (6.0 versus MCA's 5.8) and LLaVA-Bench Complex (74.7 versus Wavefront's 74.8). All four LLaVA-Bench scores improve over the backbone. POPE F1 gains 2.06 points; MathVista, SQA, MMMU, and MMStar gain 5.0, 5.8, 2.2, and 4.4 points.

On LaViDA, EPS leads on all thirteen metrics. Relative to baseline, $C_S$ falls from 33.0 to 25.0, $C_I$ from 9.3 to 7.4, and Recall rises from 57.6 to 59.8. EPS also outperforms the strongest comparators by 1.5 $C_S$ points and 0.4 $C_I$ points. Gains span description quality, POPE, and all four reasoning and understanding benchmarks; Table 2 gives every score. LLaVA-Bench Conv and Details gain 3.5 and 4.6 points, while POPE F1 gains 2.65 points.

## 4.3 MEASURED ATTENTION REDISTRIBUTION

EPS reduces suppressed-set share $A(\mathcal{S})$ from 11.1–12.5% to 0.9–1.0% across phases (Figure 4). Image share rises by 6.9, 6.7, and 6.9 percentage points. In the same-state intervention, image keys receive 67.2%, 62.7%, and 59.7% of the mass released from $\mathcal{S}$. Other output receives 19.8%, 24.7%, and 27.9%, while prompt keys receive 13.0%, 12.6%, and 12.4%. Most released attention therefore reaches the image region in every phase. This visual destination is absent in text-only dLLMs. The same-state comparison isolates immediate redistribution from changes in the generated sequence. Appendix G reports the complete allocation measurements.

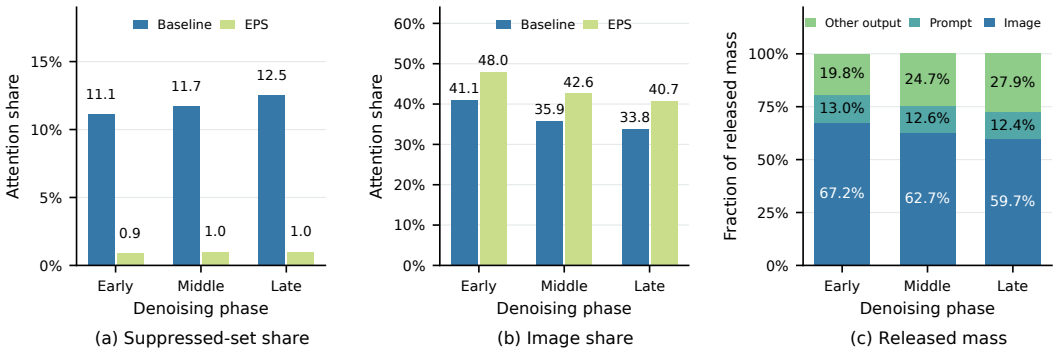



Figure 4: **EPS redirects attention toward image keys.** (a) Suppressed-set share denotes the attention mass assigned to the low-semantic eligible subset $\mathcal{S} = \mathcal{E} \cap \mathcal{L}_t$ identified before each EPS update. (b) Image share in baseline and EPS rollouts. (c) Destinations of mass released from $\mathcal{S}$ by the same-state intervention. Table 10 gives all regional changes.

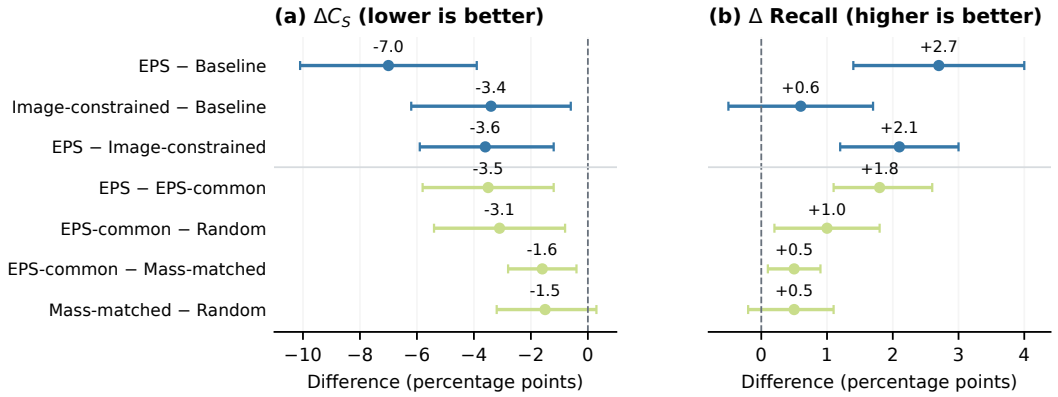



Figure 5: **Quality effects on CHAIR.** Bars show 95% confidence intervals. Each contrast subtracts the second strategy from the first. The upper group tests visual reallocation; the lower group tests selection and dose. EPS denotes full-dose suppression. EPS-common uses the same selection with a common reduced budget. Tables 13 and 14 give absolute scores and confidence intervals.

### 4.4 Mechanism Analysis and Ablations

**Visual reallocation.** Full-dose EPS lowers $C_S$ from 30.0 to 23.0 and raises Recall from 66.5 to 69.2 (Table 13). Image-constrained EPS reaches $C_S = 26.6$ and Recall$= 67.1$. Allowing visual reallocation thus lowers $C_S$ by another 3.6 points and raises Recall by 2.1 points. The paired $C_S$ interval excludes zero (Figure 5). Full-dose EPS also has the lowest $C_I$, at 6.0. Non-image redistribution already helps; visual reallocation further improves all three metrics.

**Suppressed-set selection and dose.** At a common feasible budget, suppressing the EPS set $\mathcal{S}$ improves both metrics over random and mass-matched control selections. Relative to mass matching, $C_S$ falls by 1.6 points ($[-2.8, -0.4]$) and Recall rises by 0.5 points ($[0.1, 0.9]$). The corresponding gains over random selection are 3.1 and 1.0 points. Thus, the choice of suppressed positions helps beyond matching received attention mass. Full-dose EPS further lowers $C_S$ by 3.5 points and raises Recall by 1.8 points over EPS-common. The paired $C_S$ interval excludes zero. Full-dose EPS leads on all three quality metrics. Figure 5 shows all contrasts. Appendix J reports removal diagnostics.

**Category-selective interventions.** The category ablation compares suppression targets within $\mathcal{E}$ (Table 8). Unfiltered EPS suppresses all of $\mathcal{E}$ and lowers $C_S$ from 33.5 to 29.0. Low-semantic EPS leads on all three metrics. It reaches $C_S = 27.0$, $C_I = 7.23$, and $\text{Recall}_{\text{cat}} = 68.35$. Relative to unfiltered EPS, the gains are 2.0, 0.75, and 1.17 points. Content-only EPS worsens all three metrics

from baseline. These results support selecting low-semantic positions before suppression. Each selection uses its own realized removal dose. Appendix E gives the full setup.

**Suppression strength.** Appendix K reports the suppression-strength trade-off. Reducing $\alpha$ to 0.5 further lowers $C_S$, while the default 0.8 leads on the remaining metrics.

### 4.5 Generalization to LaViDA-Dream

LaViDA-Dream replaces LaViDA's diffusion-native LLaDA-8B backbone (Li et al., 2025; Nie et al., 2025) with autoregressive-initialized Dream-7B (Ye et al., 2025). Both use full-sequence masked diffusion, but Dream-7B predicts the next token at each position. This tests transfer across pretraining origins and prediction alignments.

Table 3: **Generalization to LaViDA-Dream.** Bold denotes the better score. The benchmark suite matches Table 2.

| Method | CHAIR | | | LLaVA-Bench | | | | POPE | | MathVista | SQA | MMMU | MMStar |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $C_S \downarrow$ | $C_I \downarrow$ | Recall↑ | Conv↑ | Details↑ | Complex↑ | All↑ | Acc↑ | F1↑ | Score↑ | Score↑ | Score↑ | Score↑ |
| LaViDA-Dream | 27.5 | 6.7 | 63.7 | 75.1 | 63.8 | 76.8 | 73.1 | 85.12 | 84.58 | 46.2 | 73.0 | 44.6 | 62.7 |
| + EPS | **25.0** | **5.4** | **66.0** | **79.6** | **67.5** | **79.5** | **78.0** | **87.64** | **87.31** | **50.0** | **77.7** | **47.2** | **65.3** |

EPS improves all thirteen metrics (Table 3): $C_S$ falls from 27.5 to 25.0, $C_I$ from 6.7 to 5.4, and Recall rises from 63.7 to 66.0. LLaVA-Bench gains range from 2.7 to 4.9 points. POPE accuracy and F1 gain 2.52 and 2.73 points, MathVista and SQA gain 3.8 and 4.7 points, and MMMU and MMStar each gain 2.6 points. These results support transfer to an autoregressive-initialized diffusion backbone.

### 4.6 Computational Efficiency

EPS preserves the number of denoising steps and reuses the attention logits, without additional model forward passes or attention matrix products. It adds one pre-intervention softmax, column reductions, sorting, and elementwise logit updates. Per head and step, these operations cost $O(ML + N \log N)$, compared with $O(MLd)$ for attention matrix products with head dimension $d$ (Appendix N). The extra dense operations lack the factor $d$, leaving the dominant attention FLOP scaling unchanged. Attention materialization and kernel choices can nevertheless affect runtime.

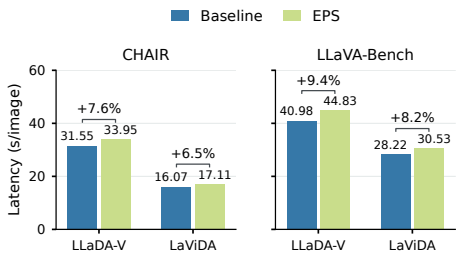



Figure 6: Per-image latency.

Figure 6 reports measured latency in seconds per image. On CHAIR and LLaVA-Bench, respectively, EPS adds 7.6% and 9.4% for LLaDA-V, and 6.5% and 8.2% for LaViDA. These measurements indicate a modest latency cost for the quality improvements reported above. The absolute increase is 1.04–3.85 seconds per image across the four settings.

## 5 Conclusion

Output attention remains concentrated during denoising in dMLLMs as dominant positions stabilize. The elbow set includes both low-semantic and content tokens. EPS suppresses selected low-semantic positions through anchor-calibrated logit shifts. Most released attention reaches image keys. Controlled comparisons support both visual reallocation and the choice of suppression targets. Low-semantic selection also outperforms content-only suppression. Across three dMLLMs, EPS improves hallucination, description quality, and multimodal reasoning. Attention allocation is thus a useful intervention point in dMLLMs.

## AI USE STATEMENT

Generative AI tools assisted with manuscript editing, code inspection, and preparation of figures and tables from existing experimental outputs. They were not used to generate or fabricate experimental observations.

## REFERENCES

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are We on the Right Way for Evaluating Large Vision-Language Models? *arXiv preprint arXiv:2403.20330*, 2024. URL https://arxiv.org/abs/2403.20330.

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 Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 8266–8286, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.404. URL https://aclanthology.org/2026.findings-acl.404/.

Jia Deng, Junyi Li, Wayne Xin Zhao, Jinpeng Wang, Hongyu Lu, and Ji-Rong Wen. Beyond fully random masking: Attention-guided denoising and optimization for diffusion language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 44502–44514, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.2060. URL https://aclanthology.org/2026.acl-long.2060/.

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. URL https://arxiv.org/abs/2606.12232.

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. URL https://openreview.net/forum?id=78Nn4QJTEN.

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 Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6853–6876, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.311. URL https://aclanthology.org/2026.acl-long.311/.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 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*, pp. 13418–13427, 2024.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025.

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 model for vision-language understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. *arXiv preprint arXiv:2304.08485*, 2023. URL https://arxiv.org/abs/2304.08485.

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. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2509.26432.

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. *arXiv preprint arXiv:2209.09513*, 2022. URL https://arxiv.org/abs/2209.09513.

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 *ICLR*, 2024.

Zanlin Ni, Shenzhi Wang, Yang Yue, Tianyu Yu, Weilin Zhao, Yeguo Hua, Tianyi Chen, Jun Song, Cheng Yu, Bo Zheng, and Gao Huang. The flexibility trap: Rethinking the value of arbitrary order in diffusion language models. *arXiv preprint arXiv:2601.15165*, 2026. URL https://arxiv.org/abs/2601.15165.

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. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *EMNLP*, 2018.

Maximo Eduardo Rulli, Simone Petruzzi, Edoardo Michielon, Fabrizio Silvestri, Simone Scardapane, and Alessio Devoto. Attention sinks in diffusion language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 43244–43256, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.2145. URL https://aclanthology.org/2026.findings-acl.2145/.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling*, 2024.

Shangwen Sun, Alfredo Canziani, Yann LeCun, and Jiachen Zhu. The spike, the sparse and the sink: Anatomy of massive activations and attention sinks. *arXiv preprint arXiv:2603.05498*, 2026.

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 (ICLR)*, 2025a.

Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zile Huang, Haochen Xue, Ziyang Chen, Zelin Peng, Zhiwei Yang, Sijin Zhou, et al. Seeing far and clearly: Mitigating hallucinations in mllms with attention causal decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26147–26159, 2025b.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim M. Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier H'enaff, Jeremiah Harmsen, Andreas Steiner, and Xiao-Qi Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *ArXiv*, abs/2502.14786, 2025.

Jinfeng Wei and Xiaofeng Zhang. Dopra: Decoding over-accumulation penalization and re-allocation in specific weighting layer. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7065–7074, 2024.

Fangzheng Wu and Brian Summa. Attention sinks in diffusion transformers: A causal analysis. *arXiv preprint arXiv:2605.09313*, 2026. URL https://arxiv.org/abs/2605.09313.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=NG7sS51zVF.

Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, et al. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308*, 2025.

Haojin Yang, Rui Zhou, Rui Hu, Zequn Sun, Yujun Cai, and Yiwei Wang. WavefrontDiffusion: Dynamic decoding schedule for improved reasoning. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2511.19473.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. MMaDA: Multimodal large diffusion language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b, 2025. URL https://hkunlp.github.io/blog/2025/dream.

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, 2025.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Lin. Unveiling and harnessing hidden attention sinks: enhancing large language models without training through attention calibration. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 57659–57677, 2024.

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 Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*, 2024.

Lunbin Zeng, Jingfeng Yao, Bencheng Liao, Hongyuan Tao, Wenyu Liu, and Xinggang Wang. Diffusionvl: Translating any autoregressive models into diffusion vision language models. *arXiv preprint arXiv:2512.15713*, 2025.

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*, 2025.

Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. Seeing clearly by layer two: Enhancing attention heads to alleviate hallucination in lvlms. *arXiv preprint arXiv:2411.09968*, 2024.

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 (ICLR' 26 Oral)*, 2026a.

Xiaofeng Zhang, Yuanchao Zhu, Qiyan Zhao, Xiaosong Yuan, Jiawei Cao, and Xuhang Chen. Tokenpenalty: Alleviating attention sinks and positional decay in lvlms. In *The 64th Annual Meeting of the Association for Computational Linguistics (ACL' 2026 Findings)*, 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 Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 6479–6490, San Diego, California, United States, July 2026c. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.323. URL https://aclanthology.org/2026.findings-acl.323/.

Qiyan Zhao, Xiaofeng Zhang, Yiheng Li, Yun Xing, Xiaosong Yuan, Feilong Tang, Sinan Fan, Xuhang Chen, Da-Han Wang, and Xu-Yao Zhang. Mca-llava: Manhattan causal attention for reducing hallucination in large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3981–3990, 2025.

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*, 2026. URL https://arxiv.org/abs/2608.22367.

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

Xianwei Zhuang, Zhihong Zhu, Yuxin Xie, Liming Liang, and Yuexian Zou. Vasparse: Towards efficient visual hallucination mitigation via visual-aware token sparsification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4189–4199, 2025.

# A Related Work

## A.1 dLLMs and dMLLMs

dLLMs generate through iterative denoising with bidirectional attention (Nie et al., 2025; Zhu et al., 2025; Ye et al., 2025). dMLLMs add visual representations and multimodal instruction tuning (You et al., 2025; Yang et al., 2025; Li et al., 2025; Yu et al., 2025; Xin et al., 2025; Zeng et al., 2025). Our study examines how output concentration shares the attention budget with image and prompt keys during denoising.

## A.2 Decoding Order and Remasking

Decoding policies shape which positions are filled and when. AdaBlock-dLLM adapts block size (Lu et al., 2026), while WavefrontDiffusion changes the active token frontier (Yang et al., 2026). Huang et al. (2026) identify boundary and trivial-token biases in uncertainty-based decoding and introduce UNCODE to calibrate token selection. Ni et al. (2026) show that arbitrary-order decoding can bypass uncertain tokens needed for reasoning exploration. Frkovic et al. (2026) find that confidence-remasking gains depend on block length and sampling settings. In dMLLMs, CACD combines token confidence with decoded-neighbor proximity and uses block-free decoding to improve contextual readiness (Zhao et al., 2026). These methods motivate careful control of the decoding schedule. EPS regulates attention within each update while keeping that schedule fixed.

## A.3 Attention Sinks across Model Families

Attention sinks receive disproportionate attention across queries (Xiao et al., 2024; Yu et al., 2024; Zhang et al., 2025; Sun et al., 2026). Initial-token sinks can support stable long-context computation in autoregressive LLMs (Xiao et al., 2024; Sun et al., 2024). Gu et al. (2025) trace sink emergence during pretraining and connect it to softmax normalization and key-like attention biases. In dLLMs, sink positions move during denoising. Rulli et al. (2026) find only minor performance degradation when these positions are masked. Dai et al. (2026) identify floating anchors with different structural and semantic roles across layers. In text-to-image diffusion, Wu & Summa (2026) use score- and value-path interventions to separate perceptual changes from semantic alignment. Our analysis focuses on image-conditioned text generation and measures both output concentration and visual allocation.

## A.4 Attention-Guided and Attention-Regulated Diffusion Inference

Attention can guide denoising order or serve as an intervention target. AGDO derives denoising order from attention to revealed context and uses attention-critical tokens in supervised fine-tuning and reinforcement learning (Deng et al., 2026). Zhang et al. (2026c) stabilize moving sinks with an added token that attends only to itself and remains visible to other tokens. EPS instead selects low-semantic columns among existing output positions. An anchor-calibrated logit shift reduces their attention and redistributes probability across the remaining keys. Image-constrained and common-budget controls test the destination and source of this redistribution.

## A.5 Attention Intervention in MLLMs

Visual attention concentration has been linked to weak grounding and hallucination (Kang et al., 2025; Zhang et al., 2026a;b). Existing methods apply decoding penalties (Huang et al., 2024; Wei & Zhang, 2024), redirect attention (Tang et al., 2025b;a), target selected layers (Zhang et al., 2024), or sparsify visual tokens (Zhuang et al., 2025). These studies motivate attention regulation for visual grounding. EPS adapts target selection to the changing output sequence in dMLLMs and tests how released attention reaches image keys.

# B Evaluation Configurations

**Attention studies.** LLaDA-V combines LLaDA-8B-Instruct (Nie et al., 2025) with a SigLIP2 visual encoder (Tschannen et al., 2025) and a vision–language projector. We use `GSAI-ML/LLaDA-V`

with the `llava_llada` conversation template. The prompt is "Please describe the image in detail." COCO val2014 filenames are sorted before model execution. The population study uses the first 500 images. Visual ablations, flow analysis, and common-budget diagnostics use the first 200. The quality comparisons in Appendix J use the full CHAIR evaluation sample. Output length, block length, and denoising steps are all 128. Decoding uses one output block and low-confidence remasking. Temperature, classifier-free guidance scale, and seed are zero; batch size is one. We use eager attention and analyze probabilities in float32. Model precision follows the loading default. Each invocation records all 32 heads at layers 8, 16, 24, and 32. EPS acts with $\alpha = 0.8$ on both filled and masked output-query rows.

**Formal task configurations.** All EPS experiments use $\alpha = 0.8$ unless stated otherwise. This default applies to every backbone, benchmark, and mechanism comparison. The suppression-strength ablation explicitly varies $\alpha$ (Table 16). Common-budget variants also limit removed mass. Full-dose EPS applies the standard update without that limit. EPS preserves each baseline's decoding schedule. LaViDA-Dream uses full-sequence denoising.

## C Population Metrics and Aggregation

For each image, layer, head, and invocation, we average over output-query rows to obtain $v_j$. Top-$k$ sums the $\min(k, N)$ largest output-normalized masses $\widetilde{v}_j$. Normalized entropy is

$$H_{\text{norm}} = -\frac{\sum_{j \in \mathcal{O}_{\text{gen}}} \widetilde{v}_j \log(\widetilde{v}_j + \epsilon)}{\log N}, \qquad \epsilon = 10^{-12}. \tag{10}$$

The concentration-event criterion is

$$C_{\text{robust}} = \frac{\max_j v_j - \text{median}(v)}{1.4826\,\text{MAD}(v) + \epsilon} > 3.5, \tag{11}$$

where $\text{MAD}(v) = \text{median}_j\,|v_j - \text{median}(v)|$. The implementation uses the lower median for an even number of values. This event threshold is separate from geometric elbow detection, which can return a nonempty $\mathcal{E}$ below the threshold.

**Repeated measurements and missing values.** For each metric and phase, we average valid records within an image, then average images with at least one valid record. Event frequency uses the same procedure. Sample prevalence is the fraction of images with any event in the phase. For invocations $t = 1, \ldots, 128$, progress is $t/128$. Phase boundaries are $1/3$ and $2/3$. The 500-image population study has 2,688,000, 2,752,000, and 2,752,000 head–layer–invocation records across the three phases, totaling 8,192,000. Each 200-image flow, visual-ablation, or common-budget arm has 1,075,200, 1,100,800, and 1,100,800 records, totaling 3,276,800. Missing values are excluded, not replaced by zero.

**Mobility.** For each image, layer, and head, adjacent invocations give switching $\mathbb{I}[j_t^* \neq j_{t-1}^*]$ and distance $|j_t^* - j_{t-1}^*|/(N-1)$. The first invocation has no predecessor and contributes neither metric. Early-phase population mobility thus uses 2,624,000 transitions. For output slots $j = 1, \ldots, N$, normalized position is $(j_t^* - 1)/(N-1)$.

**Token identity.** We group tokens by identity at all output positions. Low-semantic pieces include mask IDs, end-of-sequence IDs, punctuation, and function words. The mask ID is 126336 for the evaluated LLaDA-V tokenizer. Special IDs are resolved before lexical matching. After tokenizer whitespace markers are removed, a nonempty punctuation-only piece is low-semantic. A function-word match is also low-semantic. Other alphanumeric pieces are content. Empty or unrecognized pieces remain unclassified unless their IDs identify a mask or end-of-sequence token. The first and last output slots follow the same rules as interior slots.

The eligibility filter uses the function-word list *a, an, the, to, of, in, on, at, for, and, or, is, are, was, were, be, with, that, this, it, as, by, from*. Observational composition uses the expanded lexicon identified as `expanded_v1_20260920` in the data export. Both apply their lexical rules at every output position.

## D Complete Position Dynamics

Figure 2 gives a qualitative example at steps 44 and 88 and layers 10 and 20. Here we report complete phase summaries and additional trajectories for the population study, which measures layers 8, 16, 24, and 32 over all 128 steps. Table 4 gives phase summaries. Figure 7 shows the dominant-position trajectory. Mobility declines, but mean position is non-monotonic. The composition analysis below measures token identity separately from location.

Table 4: **Phase-level population dynamics.** Sample-first means over the same images and four layers as Table 1. Position and migration distance are normalized by 127. The first invocation has no adjacent-step transition and is excluded from switch and migration means.

| Metric | Early | Middle | Late |
|---|---|---|---|
| Dominant position | 0.447 | 0.437 | 0.526 |
| Adjacent-step switch rate | 0.237 | 0.155 | 0.135 |
| Migration distance | 0.068 | 0.040 | 0.037 |

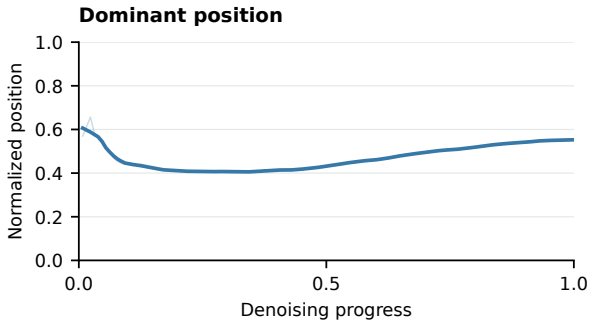



Figure 7: **Dominant-position dynamics.** Mean normalized position of the most-attended output slot. The light curve shows invocation-level sample-first means. The heavy curve is a centered seven-invocation moving average.

## E Elbow-Set Composition, Query Coverage, and Selective Interventions

**Sample and aggregation.** We select 200 COCO val2014 images before model execution with seed 20260920. Each has at least one annotated object category. All four arms use these images. LLaDA-V uses 128 output slots and 128 steps. Inference uses greedy decoding, seed 0, and float16 precision. We record layers 8, 16, 24, and 32. Records are averaged within each image, phase, and layer before averaging images. Composition ranges span the 12 stratum means.

**Lexical composition.** Elbow-set count enrichment divides the count share within $\mathcal{E}$ by class availability within each record. We average valid records within images, then average images. Masks are identified by token ID. Other special IDs form a separate class. Punctuation contains no alphanumeric characters, and function words follow an explicit lexicon. Remaining alphanumeric pieces form the content group. The composition lexicon contains more function words than the eligibility list. This sample has no non-mask special tokens. Other/unknown pieces account for at most 0.25% of elbow-set mass. Their count enrichment uses 5–18 images per stratum. Figure 8 gives the full distributions.

**Combined low-semantic composition.** The composition analysis describes $\mathcal{E}$ and combines masks, punctuation, and function words into the low-semantic group. Non-mask special tokens are absent in this sample. We sum shares within each image/phase/layer stratum, then average images. Count

share exceeds 50% in all 12 strata; mass share does so in 11. Equal-stratum means are 65.7% for count, 64.6% for mass, and 75.8% for availability. In the early phase, this group occupies 92.3% of output slots but forms 65.4–85.0% of $\mathcal{E}$. Majority composition and enrichment are therefore distinct. Figure 8 shows the lexical breakdown of the two semantic groups.

Table 5: **Elbow-set composition by token identity.** Ranges of sample-first means over 12 phase/layer strata; all values are percentages. Counts and attention mass describe $\mathcal{E}$ before filtering. Low-semantic combines masks, punctuation, and function words across all output slots. Non-mask special tokens are absent here. Unclassified pieces remain in the denominator and account for at most 0.25% of elbow-set mass. Figure 8 gives the lexical breakdown.

| Class | Slot share | Elbow-set count | Elbow-set mass |
|---|---|---|---|
| Low-semantic | 59.3–92.3 | 51.9–85.0 | 47.4–83.8 |
| Content | 7.7–40.6 | 15.0–48.0 | 16.2–52.4 |

The filled, non-boundary Top-1 analysis uses records whose dominant slot is filled and lies inside the output boundaries. Membership and availability use the same records and the expanded lexicon. Content pieces account for 28.5–70.3% of dominant slots (Table 6). Their availability is 45.0–49.0%. Content tokens thus retain a substantial share of dominant positions.

**Held-out query coverage.** Query folds are balanced within masked and filled rows using split seed 20260920. We detect $\mathcal{E}$ on one fold and evaluate it on the other, in both directions. For output key $j$ and query $i \neq j$, we normalize $P_{ij}$ by the row's total non-self output mass. Coverage is the fraction of valid queries where this share exceeds $1/(N-1)$. We average over columns in $\mathcal{E}$, then within and across images. Row Top-5 incidence ranks non-self output keys. The all-query variant detects $\mathcal{E}$ using all rows. Both variants exclude self-attention during evaluation.

Query-matching controls pair each column in $\mathcal{E}$ with a distinct column outside $\mathcal{E}$ having the same mask state. Normalized position distance is at most 0.25, and the received-mass ratio is between 0.5 and 2. The content control also requires a content piece. Selection and matching use only the detection fold. Table 7 gives match rates and residual mismatch. Elbow-set summaries cover all columns in $\mathcal{E}$; control summaries cover available matches. The two summaries therefore use different subsets.

**Selective intervention and quality.** Unfiltered EPS sets $\mathcal{S} = \mathcal{E}$. Content-only EPS retains content positions within $\mathcal{E}$. Low-semantic EPS uses $\mathcal{S} = \mathcal{E} \cap \mathcal{L}_t$ with the eligibility filter and anchor-calibrated update in Section 3. A restricted set can be empty and changes the realized dose. Table 8 compares these three interventions with the baseline.

$\text{Recall}_{\text{cat}}$ divides correctly mentioned distinct COCO categories by annotated categories, summed across images. $C_I$ uses object-mention counts. A caption with no object mentions contributes zero to the corpus denominator.

Low-semantic EPS reduces $C_S$ by 6.5 points and $C_I$ by 1.78 points relative to baseline, while $\text{Recall}_{\text{cat}}$ rises by 1.85 points. It also outperforms unfiltered suppression on all three metrics. Content-only suppression raises $C_S$ by 1.5 points and $C_I$ by 0.33 points while lowering $\text{Recall}_{\text{cat}}$ by 0.68 points. The category study is separate from the full-sample comparisons in Section 4.4.

Table 6: **Composition after conditioning on a filled interior dominant slot.** Expanded-lexicon ranges across 12 phase/layer strata, in percent. Both Top-1 membership and category availability use the same records whose dominant slot is filled and non-boundary.

| Class | Availability | Top-1 membership |
|---|---|---|
| Punctuation | 12.5–18.8 | 7.4–53.1 |
| Function word | 36.2–38.4 | 10.9–37.1 |
| Content piece | 45.0–49.0 | 28.5–70.3 |

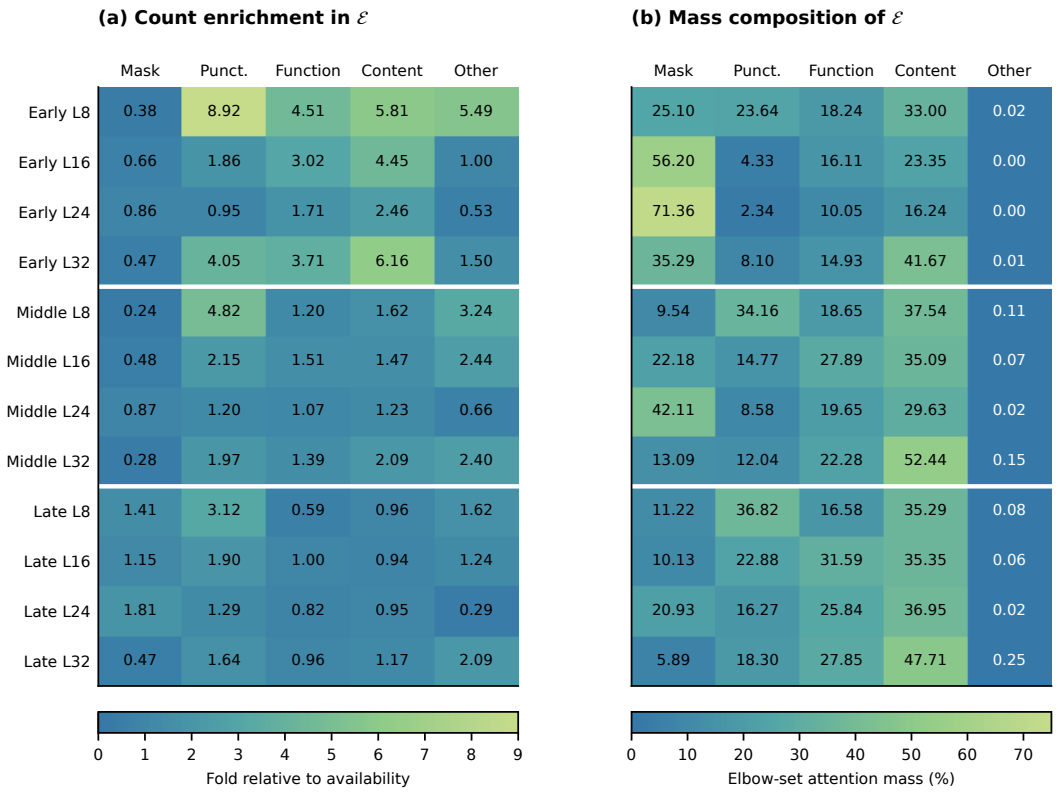



Figure 8: **Elbow-set enrichment and composition across phases and layers.** Expanded-lexicon, all-output-key analysis on the 200-image set. Left: mean count enrichment within $\mathcal{E}$ relative to class availability. Right: class shares of attention mass within $\mathcal{E}$. Both panels describe the detected set before filtering. Cells show sample-first stratum means. High enrichment can coexist with a small mass share when a class is rare.

Table 7: **Query coverage and matching support for the elbow set $\mathcal{E}$.** Ranges across phase/layer strata (and both folds for held-out evaluation). Coverage, row Top-5 incidence, and match rates are percentages. Coverage is per column in $\mathcal{E}$, above the uniform non-self output share. Mass mismatch is the absolute $\log_2$ received-mass ratio of matched pairs. Elbow-set summaries include all detected columns; control summaries include available matches outside $\mathcal{E}$.

| Selection | Coverage | Row Top-5 | Match rate | Mass mismatch |
|---|---|---|---|---|
| *All queries* | | | | |
| Elbow set $\mathcal{E}$ | 37.0–77.8 | 14.5–31.4 | — | — |
| Matched non-candidates | 25.3–47.1 | 3.0–7.0 | 28.4–64.3 | 0.571–0.669 |
| Matched content | 19.9–42.6 | 2.8–6.9 | 5.8–42.7 | 0.709–0.789 |
| *Held-out queries* | | | | |
| Elbow set $\mathcal{E}$ | 35.5–77.8 | 13.6–31.5 | — | — |
| Matched non-candidates | 25.3–47.9 | 3.0–6.9 | 35.1–68.3 | 0.590–0.686 |
| Matched content | 20.2–45.2 | 3.1–7.5 | 5.8–34.8 | 0.756–0.795 |

Table 8: **Category-selective interventions.** Quality scores are percentages. Unfiltered EPS suppresses the complete elbow set $\mathcal{E}$; Low-semantic EPS uses $\mathcal{S} = \mathcal{E} \cap \mathcal{L}_t$. Content-only EPS targets content positions within $\mathcal{E}$. Bold marks the best quality score.

| Method | $C_S \downarrow$ | $C_I \downarrow$ | $\text{Recall}_{\text{cat}} \uparrow$ |
|---|---|---|---|
| Baseline | 33.5 | 9.01 | 66.50 |
| Unfiltered EPS | 29.0 | 7.98 | 67.18 |
| Content-only EPS | 35.0 | 9.34 | 65.82 |
| **Low-semantic EPS (Ours)** | **27.0** | **7.23** | **68.35** |

## F VISUAL-INPUT ABLATIONS

On the first 200 images, we replace visual input with a blank image or seeded noise. The model, description prompt, and decoding settings stay fixed. Concentration persists (Table 9). Early/late event frequencies are 0.952/0.940 for blank images and 0.952/0.935 for noise. Regional allocation changes. Late image-position share is 0.338 with matched images, 0.254 with blank images, and 0.315 with noise (Figure 9).

Blank inputs use a uniform RGB value of 127 at the original image size; noise uses seeded uint8 pixels and the ordinary image processor. Both conditions retain the visual encoder and multimodal checkpoint. We compare their attention allocation with the matched-image condition.

Table 9: **Visual ablations.** Sample-first means over 200 COCO images per condition, 128 invocations, and the four measured layers. Event frequency, Top-$k$, entropy, and region shares follow Table 1. Cyclic image shift reuses the identical image set with a common generic prompt, so its aggregate attention statistics match the matched-image condition; it does not test semantic matching.

| Visual condition | Phase | Event freq. | Top-1 | Top-5 | Entropy | Elbow set | Image |
|---|---|---|---|---|---|---|---|
| Matched image | Early | 0.961 | 0.092 | 0.235 | 0.895 | 0.155 | 0.411 |
| | Middle | 0.945 | 0.072 | 0.208 | 0.905 | 0.191 | 0.359 |
| | Late | 0.948 | 0.071 | 0.196 | 0.908 | 0.204 | 0.338 |
| Blank image | Early | 0.952 | 0.076 | 0.211 | 0.912 | 0.167 | 0.322 |
| | Middle | 0.939 | 0.061 | 0.194 | 0.916 | 0.197 | 0.281 |
| | Late | 0.940 | 0.057 | 0.182 | 0.917 | 0.218 | 0.254 |
| Noise image | Early | 0.952 | 0.098 | 0.249 | 0.886 | 0.163 | 0.392 |
| | Middle | 0.947 | 0.070 | 0.209 | 0.903 | 0.194 | 0.342 |
| | Late | 0.935 | 0.055 | 0.181 | 0.916 | 0.215 | 0.315 |
| Cyclic image shift | Early | 0.961 | 0.092 | 0.235 | 0.895 | 0.155 | 0.411 |
| | Middle | 0.945 | 0.072 | 0.208 | 0.905 | 0.191 | 0.359 |
| | Late | 0.948 | 0.071 | 0.196 | 0.908 | 0.204 | 0.338 |

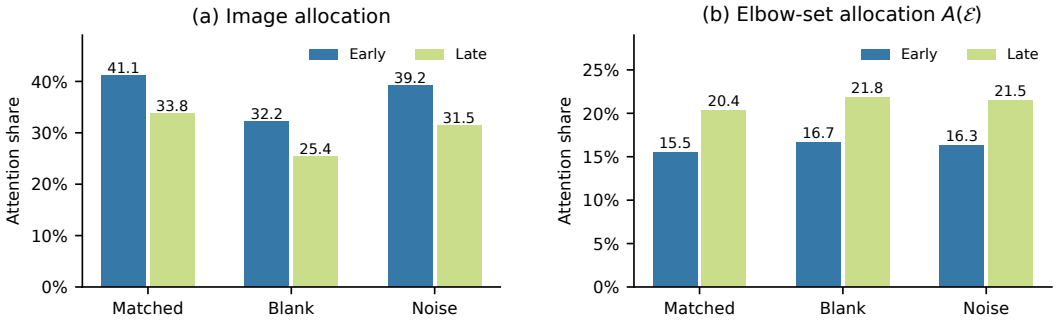



Figure 9: **Visual representations change allocation.** Early and late phase means over the same 200 images. Image share is the probability assigned to image positions, including blank and noise inputs. Concentration persists in all three conditions.

The cyclic permutation reuses the same images and generic prompt. It preserves the group-level image distribution, so it does not test semantic image–prompt mismatch.

## G COMPLETE ATTENTION-FLOW RESULTS

**Frozen-state redistribution.** We compare probabilities before and after EPS at the same state and with the same set $\mathcal{S}$. For each record, we compute removed mass $A_i(\mathcal{S}) - A'_i(\mathcal{S})$ and regional increments. When removal is positive, the destination fraction is $\rho_{\mathcal{R}} = \Delta A(\mathcal{R})/\text{removed mass}$. We then average within and across images.

EPS removes 0.102–0.115 attention mass from $\mathcal{S}$ across phases. Image keys receive 67.2%, 62.7%, and 59.7% of this mass. Other-output shares rise from 19.8% to 27.9%, while prompt shares remain near 13%. The image region is the main destination in every phase. Image, prompt, and other-output increments account for the released probability; Table 10 reports all regional changes.

**Full-rollout allocation.** Separate baseline and EPS runs start from the same input and seed. Figure 4 shows phase-averaged suppressed-set share $A(\mathcal{S})$ falling from 0.111/0.117/0.125 to 0.009/0.010/0.010. Image share rises from 0.411/0.359/0.338 to 0.480/0.426/0.407. Top-1 output fraction falls from 0.092/0.072/0.071 to 0.028/0.027/0.027, and Top-5 from 0.235/0.208/0.196 to 0.118/0.118/0.116 (Table 10). Thus the allocation change persists across all three phases of actual denoising.

Table 10: **Attention redistribution with EPS ($\alpha = 0.8$).** Sample-first means over 200 matched-image COCO samples, 128 invocations, and four layers. The elbow set $\mathcal{E}$ is detected first. Eligibility defines the actual suppressed set $\mathcal{S} = \mathcal{E} \cap \mathcal{L}_t$. All set shares and removed masses refer to $\mathcal{S}$; other output is $\mathcal{O}_{\text{gen}} \setminus \mathcal{S}$. Frozen-state effects compare attention before and after intervention at the same hidden state; full-rollout entries compare separate baseline and EPS trajectories. Each $\rho$ averages record-level ratios and need not equal the ratio of displayed means.

| Metric | Early | Middle | Late |
|---|---|---|---|
| *Frozen state: immediate redistribution on the EPS trajectory* | | | |
| Pre-intervention $\mathcal{S}$ share | 0.111 | 0.117 | 0.125 |
| Removed $\mathcal{S}$ mass (pre − post) | 0.102 | 0.107 | 0.115 |
| $\Delta$ image share | 0.069 | 0.067 | 0.069 |
| $\Delta$ prompt share | 0.013 | 0.013 | 0.014 |
| $\Delta$ other-output share | 0.020 | 0.026 | 0.032 |
| $\rho_{\text{image}}$ | 0.672 | 0.627 | 0.597 |
| $\rho_{\text{prompt}}$ | 0.130 | 0.126 | 0.124 |
| $\rho_{\text{other}}$ | 0.198 | 0.247 | 0.279 |
| *Full rollout: baseline → EPS* | | | |
| Suppressed-set share | $0.111 \to 0.009$ | $0.117 \to 0.010$ | $0.125 \to 0.010$ |
| Image share | $0.411 \to 0.480$ | $0.359 \to 0.426$ | $0.338 \to 0.407$ |
| Prompt share | $0.175 \to 0.188$ | $0.167 \to 0.180$ | $0.160 \to 0.174$ |
| Other-output share | $0.303 \to 0.323$ | $0.357 \to 0.384$ | $0.377 \to 0.409$ |
| Top-1 output fraction | $0.092 \to 0.028$ | $0.072 \to 0.027$ | $0.071 \to 0.027$ |
| Top-5 output fraction | $0.235 \to 0.118$ | $0.208 \to 0.118$ | $0.196 \to 0.116$ |
| Robust-event frequency | $0.961 \to 0.628$ | $0.945 \to 0.597$ | $0.948 \to 0.488$ |

Frozen-state rows hold the state and $\mathcal{S}$ fixed. Full-rollout rows compare independently evolving baseline and EPS states. On each trajectory, detection and eligibility define $\mathcal{S}$ from the current state; the baseline applies no shift. Post-EPS suppressed-set share uses the pre-update $\mathcal{S}$. Other output denotes $\mathcal{O}_{\text{gen}} \setminus \mathcal{S}$, including unsuppressed content positions. Top-1, Top-5, entropy, and concentration events are recomputed after the update. The summaries cover all 128 invocations with the 42/43/43 phase split.

Full-rollout image share increases by 6.9, 6.7, and 6.9 percentage points. Generation-quality effects are tested separately in Section 4.4.

# H Probability Guarantees and Numerical Implementation

## H.1 Proof of Proposition 3.1

*Proof.* For a visible anchor $a \notin \mathcal{S}$, the anchor logit is unchanged. Logits at positions in $\mathcal{S}$ receive the common shift $\log \gamma_i$, so

$$\frac{P'_{ij}}{P'_{ia}} = \gamma_i \frac{P_{ij}}{P_{ia}} \leq \frac{\alpha P_{ia}}{m_i + \epsilon} \frac{m_i}{P_{ia}} = \alpha \frac{m_i}{m_i + \epsilon} \leq \alpha. \tag{12}$$

This proves the anchor-relative bound. Averaging the bound over output-query rows also gives $v'_j \le \alpha v'_a$.

Let $D_i = \sum_\ell \exp Z_{i\ell}$ and $s_i = \sum_{j\in\mathcal{S}} P_{ij}$. The modified denominator is

$$D'_i = D_i - (1-\gamma_i)\sum_{j\in\mathcal{S}} \exp Z_{ij} = D_i[1-(1-\gamma_i)s_i]. \tag{13}$$

For $r \notin \mathcal{S}$, the numerator is unchanged, giving $P'_{ir} = P_{ir}/[1-(1-\gamma_i)s_i] \ge P_{ir}$. The denominator is positive and at most one after division by $D_i$, since $0 < \gamma_i \le 1$. Strictness follows when the original key probability and removed $\mathcal{S}$ mass are positive. Summation yields the same monotonicity for any non-suppressed region, including the image and prompt regions. □

### H.2 Numerical Floor

Algorithm 1 uses $\bar{\gamma}_i = \max(\delta, \gamma_i)$ with $\delta = 10^{-12}$ to avoid a non-finite logarithm. In exact arithmetic, the suppressed-position-to-anchor bound becomes

$$\frac{P'_{ij}}{P'_{ia}} \le \max\left\{\alpha\frac{m_i}{m_i+\epsilon}, \frac{\delta m_i}{P_{ia}}\right\}. \tag{14}$$

The ideal bound holds when the floor is inactive. Exceptionally small anchor probabilities can relax it. The non-suppressed formula remains valid because $0 < \bar{\gamma}_i \le 1$. Attention is analyzed in float32, and shifts are cast back to the logit dtype. Rounding introduces additional numerical error.

### H.3 Measured Numerical Diagnostics

Table 11 reports float16 GPU diagnostics for the three suppression strategies. Unfiltered EPS uses $\mathcal{S} = \mathcal{E}$. Content-only EPS selects content positions within $\mathcal{E}$. Low-semantic EPS uses $\mathcal{S} = \mathcal{E} \cap \mathcal{L}_t$. Residuals are measured on each strategy's suppressed set. Low-semantic EPS has a mean floor rate of $1.21\times10^{-5}$ and mean bound residual of $2.48\times10^{-5}$. Its maximum summary residual is $1.36\times10^{-4}$. Its maximum summary row error is $2.14\times10^{-7}$. These maxima are over sample/phase/layer means. The small positive residuals are consistent with the numerical floor and rounding.

Table 11: **Numerical diagnostics from float16 interventions.** Means and maxima are across 2,400 sample/phase/layer summary rows per arm. Bound residual is the positive part of $\max_{j\in\mathcal{S}} P'_{ij} - \alpha P'_{ia}$, where $\mathcal{S}$ is the suppressed set for that strategy. Each summary row averages record-level diagnostics. The displayed maxima are maxima of these averages, not global per-update bounds. The shaded row is the main EPS configuration.

| Strategy | Mean floor rate | Mean residual | Max summary residual | Max summary row error |
|---|---|---|---|---|
| Unfiltered EPS | $1.60\times10^{-5}$ | $6.82\times10^{-5}$ | $2.21\times10^{-4}$ | $2.10\times10^{-7}$ |
| Content-only EPS | $1.66\times10^{-5}$ | $2.42\times10^{-5}$ | $2.75\times10^{-4}$ | $2.12\times10^{-7}$ |
| **Low-semantic EPS (Ours)** | $1.21\times10^{-5}$ | $2.48\times10^{-5}$ | $1.36\times10^{-4}$ | $2.14\times10^{-7}$ |

### H.4 Additional-Key Normalization

Adding keys alone cannot increase an existing key's attention if all original logits remain fixed. For an existing key $s$, old denominator $D$, and appended keys $\mathcal{V}$,

$$P_s^{\text{aug}} = \frac{e^{z_s}}{D + \sum_{r\in\mathcal{V}} e^{z_r}} \le \frac{e^{z_s}}{D}. \tag{15}$$

Visual conditioning also changes hidden states, queries, and keys. The fixed-logit identity isolates only the effect of adding keys to the denominator.

## I Implementation and Common-Budget Diagnostics

The mechanism runner applies EPS with other attention interventions disabled. It records probabilities and detects $\mathcal{E}$. It then filters to $\mathcal{S}$ and modifies the selected logits. Thirteen CPU unit tests cover

baseline identity, numerical EPS examples, and aggregation. They also check budget rules, CHAIR parsing, and paired metrics. Table 11 reports GPU diagnostics.

A budget-matched control must be able to remove the requested mass from its selected columns. The count-matched random, mass-matched non-candidate, and content-prioritized variants often lack this capacity at the unconstrained EPS budget. The common-budget implementation therefore reduces the requested row budget to the minimum feasible amount across suppressed-set selections at the same state. Content selection uses non-candidate backfilling when too few content tokens are available; records with insufficient eligible columns are marked invalid. The adjustment retains about 14–15%

Table 12: **Common-budget implementation diagnostics.** Sample-first means over 200 images and four measured layers. A row-wise common feasible budget reduces the requested removal before applying each control; the limited-row rate is zero by construction after that reduction. MAE is mean absolute row-level matching error; invalid records are column-count mismatches. Budget ratio is the mean record-level common-budget/unconstrained-EPS ratio. These capacity-constrained variants characterize implementation behavior and do not establish candidate-specific harm.

| Control | Phase | Removed mass | MAE ($10^{-5}$) | Limited (%) | Invalid (%) | Budget ratio |
|---|---|---|---|---|---|---|
| Random | Early | 0.0173 | 0.77 | 0.0 | 3.24 | 0.142 |
| | Middle | 0.0220 | 1.22 | 0.0 | 1.92 | 0.137 |
| | Late | 0.0272 | 1.17 | 0.0 | 0.76 | 0.151 |
| Mass-matched | Early | 0.0173 | 2.05 | 0.0 | 3.23 | 0.143 |
| | Middle | 0.0219 | 3.33 | 0.0 | 1.89 | 0.137 |
| | Late | 0.0267 | 3.84 | 0.0 | 0.78 | 0.149 |
| Content | Early | 0.0173 | 1.37 | 0.0 | 3.23 | 0.143 |
| | Middle | 0.0220 | 1.91 | 0.0 | 1.87 | 0.137 |
| | Late | 0.0269 | 2.64 | 0.0 | 0.75 | 0.150 |

of the original EPS removal (Table 12). This is a probability-mass ratio. A low-capacity control reduces every selection's budget, including EPS-common. Separate rollouts still have different states and realized removal amounts. Quality comparisons appear in Section 4.4.

## J Visual Reallocation and Suppressed-Set Controls

Table 13 reports absolute quality scores for the visual-reallocation and suppressed-set controls discussed below.

**Visual reallocation.** The image-constrained control uses the EPS set $\mathcal{S}$ at $\alpha = 0.8$. It limits immediate visual reallocation in favor of non-image keys outside $\mathcal{S}$. The constraint acts within each update. Later hidden states and image attention remain free to change. Paired quality comparisons measure the benefit of EPS over this constrained variant.

**Suppressed-set selection.** Common-budget variants suppress the EPS set $\mathcal{S}$ or a mass-matched or random control set. Control columns come from outside $\mathcal{E}$. A shared rule sets the feasible removal budget at each state. Full-dose EPS uses the same low-semantic selection without the budget limit. Table 15 reports common-budget removal diagnostics. Those controls retain 14.0–14.4% of the full removal dose. Separate rollouts yield slightly different realized budgets. The full-sample quality contrasts follow below.

**Paired quality comparisons.** Table 14 and Figure 5 report quality contrasts and paired confidence intervals. EPS lowers $C_S$ by 3.6 points and raises Recall by 2.1 points over image-constrained EPS. The constrained variant also lowers $C_S$ from baseline. Its Recall interval includes zero. At a common reduced budget, targeting $\mathcal{S}$ outperforms random and mass-matched selections. Against mass matching, $C_S$ falls by 1.6 points (95% CI: $[-2.8, -0.4]$). Recall rises by 0.5 points ($[0.1, 0.9]$). Full-dose EPS further lowers $C_S$ by 3.5 points and raises Recall by 1.8 points over EPS-common. Recall intervals for full-dose EPS contrasts assume a fixed estimation-error distribution. The intervals for mass matching versus random selection include zero.

Table 13: **Quality effects of visual reallocation and suppressed-set selection.** Both panels use the full CHAIR sample and $\alpha = 0.8$. EPS suppresses $\mathcal{S} = \mathcal{E} \cap \mathcal{L}_t$. EPS-common uses the same selection under a common reduced budget. The second panel compares it with two control sets. Scores are percentages. Bold marks the best score within each panel.

| Strategy | $C_S \downarrow$ | $C_I \downarrow$ | Recall↑ |
|---|---|---|---|
| *Visual reallocation* | | | |
| Baseline | 30.0 | 7.6 | 66.5 |
| Image-constrained EPS | 26.6 | 7.3 | 67.1 |
| EPS, full dose | **23.0** | **6.0** | **69.2** |
| *Suppressed-set selection and intervention strength* | | | |
| EPS-common | 26.5 | 7.4 | 67.4 |
| Mass-matched non-candidates | 28.1 | 7.6 | 66.9 |
| Random non-candidates | 29.6 | 7.8 | 66.4 |
| EPS, full dose | **23.0** | **6.0** | **69.2** |

Table 14: **Quality differences on the full CHAIR evaluation sample at $\alpha = 0.8$.** Differences and 95% confidence intervals are in percentage points. EPS denotes the low-semantic method at full dose. EPS-common uses the same selection under a common reduced budget.

| Comparison | $\Delta C_S$ | 95% CI | ΔRecall | 95% CI |
|---|---|---|---|---|
| *Visual reallocation* | | | | |
| EPS − Baseline | −7.0 | [−10.1, −3.9] | +2.7 | [1.4, 4.0] |
| Image-constrained − Baseline | −3.4 | [−6.2, −0.6] | +0.6 | [−0.5, 1.7] |
| EPS − Image-constrained | −3.6 | [−5.9, −1.2] | +2.1 | [1.2, 3.0] |
| *Suppressed-set selection and intervention strength* | | | | |
| EPS − EPS-common | −3.5 | [−5.8, −1.2] | +1.8 | [1.1, 2.6] |
| EPS-common − Random | −3.1 | [−5.4, −0.8] | +1.0 | [0.2, 1.8] |
| EPS-common − Mass-matched | −1.6 | [−2.8, −0.4] | +0.5 | [0.1, 0.9] |
| Mass-matched − Random | −1.5 | [−3.2, 0.3] | +0.5 | [−0.2, 1.1] |

Table 15: **Common-budget suppression diagnostics at $\alpha = 0.8$.** Removed mass is measured on each strategy's suppressed set. Image-share changes are probability fractions. Relative dose uses the full-dose EPS removal of 0.1690 as its 100% reference. Quality comparisons appear in Table 14.

| Strategy | Removed mass | Relative dose | Δ image share |
|---|---|---|---|
| EPS-common | 0.0242 | 14.4 | +0.0108 |
| Mass-matched non-candidates | 0.0239 | 14.2 | +0.0062 |
| Random non-candidates | 0.0236 | 14.0 | +0.0041 |

## K Additional Evaluation Details

Table 2 combines the seven benchmark families for LLaDA-V and LaViDA. LLaVA-Bench reports four separate dimensions: Conv, Details, Complex, and All. All is scored independently of the other three dimensions. Table 3 uses the same benchmark suite for LaViDA-Dream. Score highlighting denotes numerical rank, not statistical significance.

**Suppression strength.** The strength ablation in Table 16 shows a trade-off. Reducing $\alpha$ from 0.8 to 0.5 lowers $C_S$ from 23.0 to 22.0. Recall falls from 69.2 to 68.8, and all four LLaVA-Bench scores decrease. At $\alpha = 0.1$, Complex falls to 63.6 and All to 66.0. The default $\alpha = 0.8$ leads on six of seven metrics, including $C_I = 6.0$ and Recall$= 69.2$.

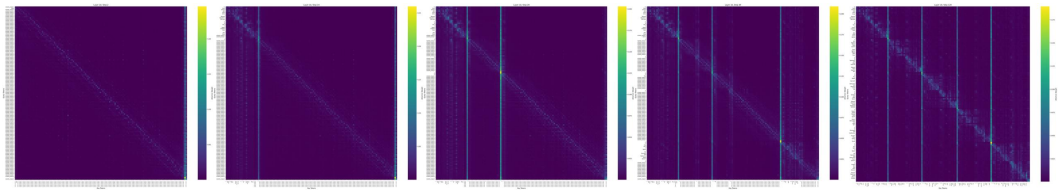
Table 16: **Suppression-strength ablation on LLaDA-V.** Smaller $\alpha$ gives stronger suppression. $\alpha = 0.5$ gives the lowest $C_S$. The default $\alpha = 0.8$ gives the best score on all six remaining metrics and matches Table 2. Bold denotes the best score per metric.

| Method | CHAIR | | | LLaVA-Bench | | | |
|---|---|---|---|---|---|---|---|
| | $C_S \downarrow$ | $C_I \downarrow$ | Recall↑ | Conv↑ | Details↑ | Complex↑ | All↑ |
| LLaDA-V | 30.0 | 7.6 | 66.5 | 72.2 | 62.4 | 71.8 | 70.8 |
| + EPS ($\alpha = 0.8$) | 23.0 | **6.0** | **69.2** | **75.2** | **67.2** | **74.7** | **72.1** |
| + EPS ($\alpha = 0.5$) | **22.0** | 7.2 | 68.8 | 74.3 | 63.6 | 70.6 | 70.3 |
| + EPS ($\alpha = 0.1$) | 24.0 | 6.2 | 67.3 | 73.0 | 62.6 | 63.6 | 66.0 |

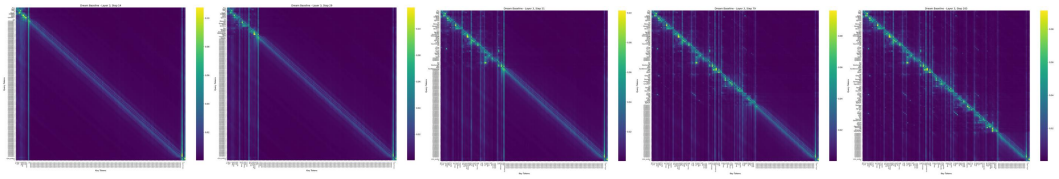
## L Attention across LaViDA Backbones

LaViDA results appear in Table 2; LaViDA-Dream generalization results appear in Table 3. Both use the same seven benchmark families.

Figure 10 shows output-column concentration during selected denoising trajectories from both backbones.



(a) LaViDA

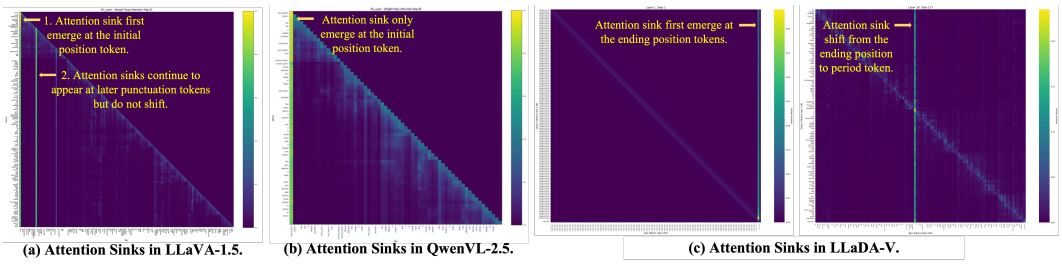


(b) LaViDA-Dream

Figure 10: **Attention concentration in dMLLMs.** Selected denoising trajectories show vertical output-column concentration in LaViDA and LaViDA-Dream. Each panel shows its denoising step and color scale.

## M ATTENTION IN AUTOREGRESSIVE MLLMS AND DMLLMS

Figure 11 compares attention maps from autoregressive MLLMs and dMLLMs. LLaVA-1.5 and Qwen2.5-VL show concentration at the initial token. LLaVA-1.5 also shows later punctuation-associated columns. LLaDA-V shows a boundary column at layer 1, step 2. At layer 10, step 127, it shows an interior full-stop column. Section 2.4 quantifies concentration and movement across images.




(a) Attention Sinks in LLaVA-1.5. (b) Attention Sinks in QwenVL-2.5. (c) Attention Sinks in LLaDA-V.

Figure 11: **Attention concentration in autoregressive MLLMs and dMLLMs.** From left to right: LLaVA-1.5, Qwen2.5-VL, and two denoising stages of LLaDA-V. Annotations identify concentrated columns.

## N QUALITATIVE EXAMPLES AND COMPUTATIONAL COMPLEXITY

**Complexity.** For each head and step, probability computation and aggregation cost $O(ML)$. Sorting costs $O(N \log N)$. Elbow detection costs $O(N)$, and logit updates cost $O(M|\mathcal{S}|)$. EPS adds no trainable parameters. Materializing attention and disabling fused attention kernels add memory and runtime costs.

**Punctuation substitution.** After full-stop tokens are suppressed, concentration appears at "of" and "to" (Figure 12). These examples show concentration shifting to other low-semantic tokens as the generated context changes.

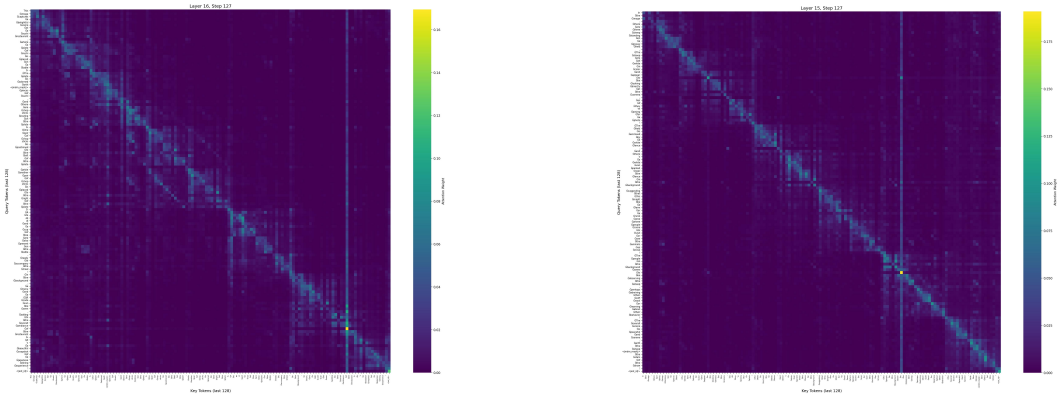


Figure 12: **Punctuation-suppression examples.** Concentrated columns appear at "of" (left) and "to" (right) after full-stop tokens are suppressed.

**Generation length.** Figure 13 shows output concentration at 64- and 256-token budgets. The examples differ in available positions and generated context. EPS adapts the size and membership of $\mathcal{S}$ at each step.

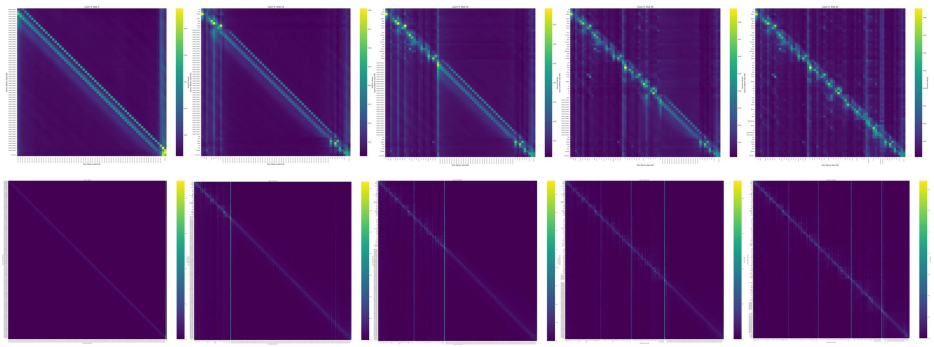


Figure 13: **Output concentration at different generation lengths.** Top: an example with 64 output slots. Bottom: an example with 256 output slots.