VIVA: VISUAL INFORMATION DISPERSION VIA ATTENTION GUIDANCE FOR PARALLEL DECODING IN DIFFUSION MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion multimodal large language models (dMLLMs) enable parallel decoding by predicting multiple masked tokens using bidirectional context, offering a promising alternative to autoregressive MLLMs. However, existing token selection strategies largely determine which tokens to decode based on prediction confidence at each step, without considering how visual conditioning influences the overall decoding trajectory. In this work, we analyze dMLLM decoding trajectories and identify two key issues: visually relevant tokens tend to be decoded in later stages, limiting their contribution as context for earlier predictions, and tokens describing visual content often influence each other's predictions, making their simultaneous decoding prone to prediction errors. Based on these observations, we propose **V**isual **I**nformation Dispersion **v**ia **A**ttention Guidance (VIVA), an inference-time decoding method that disperses visually informative tokens both across decoding steps and within each step. VIVA uses token-to-image attention to distribute visually relevant tokens across the decoding trajectory and limits their simultaneous selection within each step, while preserving the ability to decode tokens in parallel. Across six benchmarks and three open-source dMLLMs, VIVA demonstrates strong performance against existing decoding baselines, particularly when multiple tokens are decoded in parallel. Specifically, on LaViDa-Reason, VIVA improves CIDEr from 67.5 to 86.0 on COCO Captioning and MMBench accuracy from 43.0 to 51.4 over confidence-based decoding when decoding eight tokens per step. Further analyses and qualitative examples confirm that VIVA effectively mitigates both identified decoding issues.

1 INTRODUCTION

The success of diffusion large language models (dLLMs) (Nie et al., 2026; Ye et al., 2025) has driven their extension to multimodal inputs, leading to the development of diffusion multimodal large language models (dMLLMs) (Yang et al., 2026b; Li et al., 2026a; You et al., 2025). These models inherit the decoding process of dLLMs by generating responses through iterative denoising of masked tokens in the sequence. Distinct from standard autoregressive multimodal large language models (AR-MLLMs), which predict the next token from causal context and decode one token at a time (Bai et al., 2025; Wang et al., 2025), dMLLMs predict tokens at all masked positions using bidirectional context and then decode multiple tokens in parallel (Yang et al., 2026b; Li et al., 2026a; You et al., 2025). These differences offer the chance to capture rich multimodal context in the token prediction process while achieving speedups in inference. For successful decoding, it is of importance to identify the masked positions to decode, since the selected tokens serve as context for the remaining masked tokens and also affect subsequent predictions (Yang et al., 2026a).

Conventionally, masked positions are picked based on the model's certainty in its predictions, typically favoring those with high confidence (top-1 probability) or low predictive entropy (Ben-Hamu et al., 2026; Nie et al., 2026; Wu et al., 2025). A potential drawback of these criteria is that they make token selection decisions based only on the current step, without considering the overall decoding trajectory (Lee et al., 2025). In recent studies, approaches that analyze decoding trajectories and adjust token selection strategies based on the resulting observations have been proposed. For example, PC-Sampler identifies a task-agnostic bias in decoding order and adjusts token selection priorities

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.

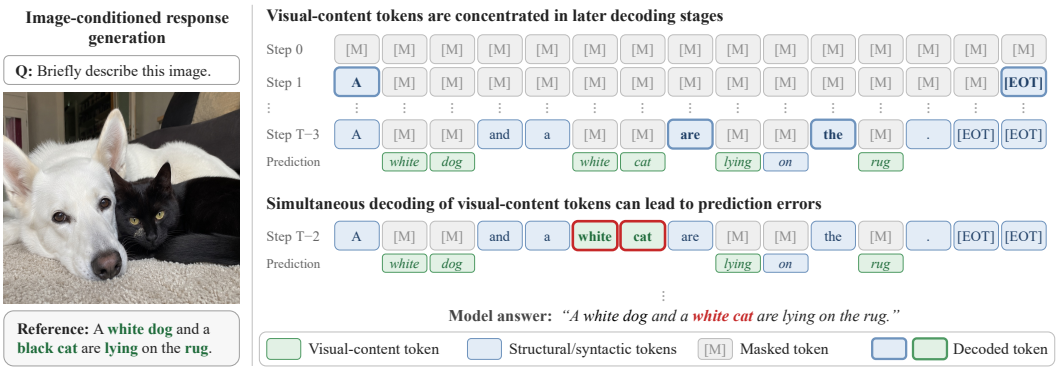

Figure 1: Illustration of two key issues in dMLLM decoding under visual conditioning. Structural and syntactic tokens are decoded early, leaving visual-content tokens concentrated in later stages. Simultaneously decoding these tokens can cause prediction errors by ignoring their mutual influence, which may further propagate to subsequent decoding steps.

to produce trajectories better suited to individual tasks (Huang et al., 2025). Another method, called LocalLeap, addresses unnecessary decoding delays that arise when tokens remain masked despite having consistent predictions across decoding steps (Kong et al., 2025). While these methods have improved decoding quality and efficiency, the effect of visual input on token selection and decoding trajectories remains largely unexplored.

In this work, we conduct an extensive analysis of decoding trajectories when images are provided as input and then identify two key issues degrading the decoding quality of dMLLMs. First, we observe that masked tokens that are highly correlated with the given image tend to be decoded later in the decoding trajectory. As illustrated in Figure 1, tokens that are easily predictable from linguistic context, such as structural or syntactic tokens, are decoded predominantly in the early stages, while tokens relying on visual information are decoded mainly in later stages. For this reason, tokens decoded in the early stages cannot benefit from the context provided by visually grounded tokens. Second, we observe that tokens associated with visual content often have a strong influence on each other's predictions. This phenomenon is further illustrated by the token predictions "white" and "cat", both describing the image content. When these positions are decoded in parallel, their predictions are made without reflecting their mutual influence, which can result in parallel prediction errors such as "white cat".

To address these challenges, we propose **V**isual **I**nformation Dispersion **v**ia **A**ttention Guidance (VIVA), a novel dMLLM decoding method that distributes visually relevant tokens across the decoding trajectory. By doing so, VIVA improves the utilization of visual information while reducing errors in parallel decoding. A salient feature of VIVA is that it uses masked-token-to-image attention as a visual relevance signal to adjust the decoding order, prioritizing visually relevant tokens for earlier decoding to prevent their concentration in later stages. Further, VIVA suppresses the simultaneous decoding of tokens with high visual relevance to reduce errors caused by their mutual influence, and also preserves parallelism by grouping tokens according to their image-attention patterns and applying the constraint only within each group. VIVA operates entirely at inference time, requiring no retraining or model modification while introducing only minimal computational overhead.

We evaluate VIVA on six benchmarks spanning image captioning, detailed image description, visual question answering, and multimodal reasoning, using three open-source dMLLMs. VIVA consistently improves generation quality, with particularly large gains when multiple tokens are decoded in parallel. For example, on LaViDa-Reason (Li et al., 2026a), compared with confidence-based decoding, VIVA improves CIDEr from 67.5 to 86.0 on COCO Captioning (Lin et al., 2014) and from 42.1 to 52.9 on Flickr30k (Young et al., 2014), while increasing MMBench (Liu et al., 2024) accuracy from 43.0 to 51.4 when decoding eight tokens per step. Additional analyses and qualitative examples further illustrate the effectiveness of VIVA in dispersing visually relevant tokens across decoding steps and reducing parallel prediction errors.

2 PRELIMINARIES

2.1 DIFFUSION MULTIMODAL LARGE LANGUAGE MODELS

We consider an absorbing diffusion model p_θ that estimates the conditional distribution of masked tokens given an image I and a text prompt P (Li et al., 2026a; You et al., 2025). At decoding step $t \in \{0, 1, \ldots, T-1\}$, the model takes a partially masked token sequence $\boldsymbol{x}_t = (x_t^1, x_t^2, \ldots, x_t^L)$, together with encoded visual tokens $\boldsymbol{v}$ and prompt tokens $\boldsymbol{p}$, concatenated as $[\boldsymbol{v}; \boldsymbol{p}; \boldsymbol{x}_t]$. For each masked position $i \in \mathcal{M}_t$, where $\mathcal{M}_t := \{i \mid x_t^i = \texttt{[MASK]}\}$, the model estimates a categorical distribution over vocabulary tokens $w \in \mathcal{V}$, given by

$$p_\theta(x_{t+1}^i = w \mid \boldsymbol{x}_t, \boldsymbol{p}, \boldsymbol{v}). \tag{1}$$

For simplicity, we henceforth write $p(x_{t+1}^i = w)$ whenever no confusion arises.

2.2 TOKEN SELECTION CRITERIA AND PARALLEL PREDICTION ERROR IN DMLLMS

In parallel decoding of dMLLMs, a subset $C_t \subseteq \mathcal{M}_t$ of size k is selected from the masked positions at step t and unmasked using predictions from a single forward pass. A common approach ranks positions using scores derived from their predictive distributions and greedily chooses the highest-ranked ones (Li et al., 2026a; Yang et al., 2026b; You et al., 2025). Given the predictive distribution $p(x_{t+1}^i)$ at position i, the confidence and entropy scores are defined as follows:

$$\begin{aligned} s_t^{\text{conf}}(i) &= p(x_{t+1}^i = w_1), \\ s_t^{\text{ent}}(i) &= \sum_{w \in \mathcal{V}} p(x_{t+1}^i = w) \log p(x_{t+1}^i = w), \end{aligned} \tag{2}$$

where w_1 denotes the token with the largest probability under $p(x_{t+1}^i)$. These criteria are natural choices, as they use the model's predictive certainty to prioritize tokens whose predictions are considered reliable (Wu et al., 2025; Ben-Hamu et al., 2026).

Nonetheless, selecting tokens based solely on these criteria can lead to errors when they are decoded in parallel (Ben-Hamu et al., 2026). This is because parallel decoding uses a product of marginal distributions to approximate the joint conditional distribution of the selected tokens:

$$q(\boldsymbol{x}_{t+1}^{C_t}) := \prod_{i \in C_t} p(x_{t+1}^i). \tag{3}$$

where q denotes the product-of-marginals approximation to the joint conditional distribution, and $\boldsymbol{x}_{t+1}^{C_t} = (x_{t+1}^{i_1}, x_{t+1}^{i_2}, \ldots, x_{t+1}^{i_k})$, with $(i_1, i_2, \ldots, i_k)$ specifying an arbitrary order of the token positions in C_t. The resulting error, known as dependency error, can be decomposed using the chain rule of KL divergence as follows:

$$\begin{aligned} \mathcal{E}(C_t) &= D_{\text{KL}}\left(p(\boldsymbol{x}_{t+1}^{C_t}) \| q(\boldsymbol{x}_{t+1}^{C_t})\right) \\ &= \sum_{j=1}^{k} \mathbb{E}_{\boldsymbol{x}_{t+1}^{i<j}} \left[D_{\text{KL}}\left(p(x_{t+1}^{i_j} \mid \boldsymbol{x}_{t+1}^{i<j}) \| p(x_{t+1}^{i_j})\right)\right], \end{aligned} \tag{4}$$

where $\boldsymbol{x}_{t+1}^{i<j} = (x_{t+1}^{i_1}, x_{t+1}^{i_2}, \ldots, x_{t+1}^{i_{j-1}})$.

3 VISUAL INFORMATION DISPERSION FOR PARALLEL DECODING

3.1 OBSERVATION

Image attention and parallel prediction error along the decoding trajectory. To understand the decoding dynamics of dMLLMs, we examine the decoding order of tokens with high attention to the image and the stepwise variations in parallel prediction error. Specifically, we decode one token at a time based on confidence and record their decoding order. At each decoding step t, we measure the image-attention mass of the selected token x_t^i:

$$A_t^i = \sum_{j \in \boldsymbol{v}} \boldsymbol{A}_{i,j}, \tag{5}$$

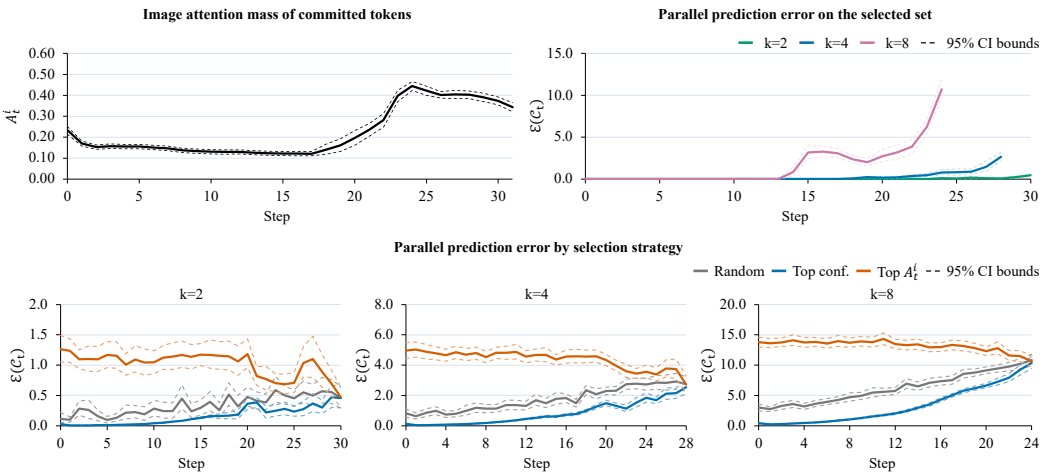

Figure 2: Image attention and parallel prediction error in LaViDa-Reason on the COCO captioning dataset. Top: Later decoding steps exhibit higher image attention and parallel prediction error. Bottom: Selecting tokens with higher image attention yields larger parallel prediction errors.

where $\boldsymbol{A}_{i,j}$ is the attention weight, averaged across heads in the last Transformer layer, from x_t^i to the j-th element of $\boldsymbol{v}$.

After generating the final sequence $\boldsymbol{x}_T$, we re-mask sets of $k \in \{2, 4, 8\}$ tokens that are consecutive in the recorded decoding order, keeping all other tokens fixed. For each set, we compare parallel predictions from a single forward pass with their sequential predictions under the same initial context to quantify parallel prediction error.

As shown in Figure 2 (top), tokens decoded later in the trajectory tend to exhibit higher A_t^i, while sets corresponding to later decoding steps exhibit higher parallel prediction error. If A_t^i serves as a proxy for the visual informativeness of x_t^i (further examined in Appendix B.2), these observations suggest that the decoding of visually informative tokens is concentrated in later steps, where parallel prediction error is also higher. Even so, these trends do not establish a direct link between visual informativeness and parallel prediction error, motivating the following question: *"Are visually informative tokens particularly prone to prediction errors when decoded in parallel?"*

Image-attention concentration and parallel prediction error. To address this question more directly, we examine the association between the sum of A_t^i over a token set and its parallel prediction error. At each step t, we select sets of $k \in \{2, 4, 8\}$ tokens from the remaining masked positions in $\boldsymbol{x}_t$ using random selection, top-k confidence, or top-k A_t^i and compare the resulting errors across these selection strategies.

Interestingly, for each $k \in \{2, 4, 8\}$, top-k selection by A_t^i yields the largest parallel prediction error among the three strategies across all evaluated decoding steps (see Figure 2 (bottom)). Taken together, these observations suggest that the elevated parallel prediction error for tokens decoded in later steps is associated with jointly decoding tokens with high A_t^i. This motivates a decoding scheme that distributes their decoding more evenly across the trajectory while limiting their concentration within each step.

3.2 Visual Information Dispersion via Attention Guidance

Motivated by these observations, we introduce VIVA (**V**isual **I**nformation dispersion **v**ia **A**ttention guidance), a decoding scheme that balances the emergence of visually informative tokens across the decoding trajectory. VIVA first combines prediction confidence $s_t^{\text{conf}}(i)$ with an attention-derived score to reorder masked positions, reducing the tendency for tokens with high image attention A_t^i to be decoded late. Following this order, VIVA regulates joint decoding of visual information within each step by constraining the sum of A_t^i over selected tokens. For fine-grained regulation, we group adjacent tokens into segments based on their image-attention distributions and impose joint con-

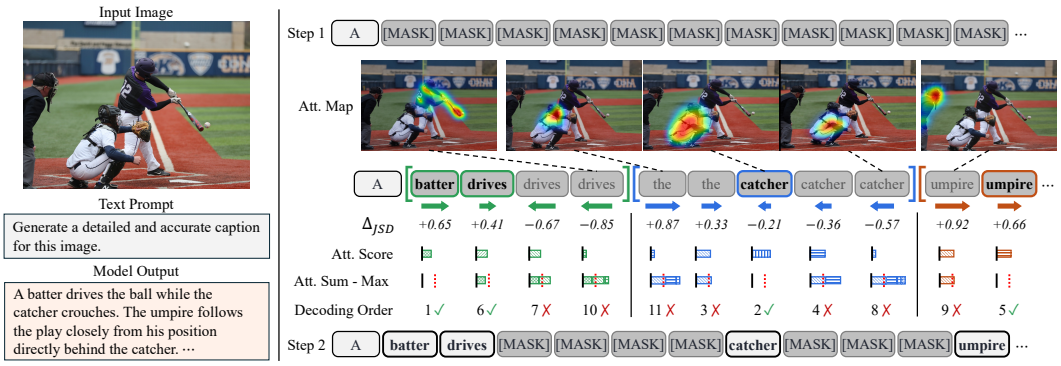

Figure 3: Overview of VIVA. VIVA uses image attention to guide token selection, groups adjacent tokens by their image-attention distributions, and applies segment- and sequence-level constraints to disperse visual information across decoding steps while limiting its concentration within each step.

straints at the segment and sequence levels. This hierarchical design limits both local concentration within each step and temporal concentration across the decoding trajectory.

Promoting tokens with high image-attention mass. We modify the decoding order so that tokens with high A_t^i are not concentrated in later steps. For each masked position $i \in \mathcal{M}_t$, we normalize A_t^i over the remaining masked positions to obtain a relative image-attention score:

$$s_t^{\text{attn}}(i) = \frac{A_t^i}{\sum_{j \in \mathcal{M}_t} A_t^j}. \tag{6}$$

We then incorporate $s_t^{\text{attn}}(i)$ into prediction confidence using an exponential weighting factor:

$$s_t^{\text{vg}}(i) = \exp\big(\lambda\, s_t^{\text{attn}}(i)\big)\, p(x_{t+1}^i = w_1), \tag{7}$$

where λ controls the contribution of the image-attention score. This weighting follows the reward-guided sampling formulation, treating normalized prediction confidence as the base distribution and $s_t^{\text{attn}}(i)$ as the reward (see Appendix C.1 for derivation).

Grouping adjacent tokens by changes in image-attention distributions. In diffusion-based parallel decoding, it has been observed that jointly decoded tokens closer in the sequence exhibit larger prediction errors, suggesting stronger local dependencies among their predictions (Hersche et al., 2026; Kim et al., 2026). Intuitively, in image-conditioned text generation, nearby tokens may become more interdependent when they attend to similar visual regions. We therefore partition adjacent tokens into segments based on changes in their image-attention distributions to capture the concentration of A_t^i at the segment level.

Let $\boldsymbol{a}_t^i = \boldsymbol{A}_{i,\boldsymbol{v}}/A_t^i$ denote the normalized image-attention distribution of token i over $\boldsymbol{v}$. We quantify the relative change in image-attention distributions around each position i as

$$\Delta(i) = \text{JSD}\big(\boldsymbol{a}_t^i, \boldsymbol{a}_t^{i-1}\big) - \text{JSD}\big(\boldsymbol{a}_t^i, \boldsymbol{a}_t^{i+1}\big), \tag{8}$$

where $\text{JSD}(\cdot,\cdot)$ denotes the Jensen–Shannon divergence. A larger $\Delta(i)$ indicates that $\boldsymbol{a}_t^i$ is more similar to the distribution on its right than to that on its left. We therefore define a segment boundary between positions i and $i{+}1$ when $\Delta(i) < 0$ and $\Delta(i{+}1) > 0$, indicating a sharp transition between their image-attention distributions.

Adaptive token selection with hierarchical constraints. The preceding components determine the priority and local structure of masked tokens; we now use them to adaptively determine which tokens are decoded together at each step. For a subset $\mathcal{S} \subseteq \mathcal{M}_t$, we measure its cumulative image-attention mass excluding the largest contribution:

$$M(\mathcal{S}) = \sum_{i \in \mathcal{S}} A_t^i - \max_{i \in \mathcal{S}} A_t^i, \tag{9}$$

with $M(\varnothing) = 0$. By constraining $M(\mathcal{S})$ below a threshold τ_t, we regulate the concentration of A_t^i within the selected tokens. The threshold is adaptively determined from the remaining masked tokens and the number of remaining decoding steps. Let $\boldsymbol{x}_t = [\boldsymbol{g}_1; \boldsymbol{g}_2; \ldots; \boldsymbol{g}_N]$ denote the segmentation of $\boldsymbol{x}_t$ into N contiguous segments. We define the segment- and sequence-level thresholds as

$$\tau_{t,n}^{\text{seg}} = \frac{M(\mathcal{M}_t \cap \mathcal{G}_n)}{T-t}, \qquad \tau_t^{\text{seq}} = \frac{M(\mathcal{M}_t)}{T-t}, \tag{10}$$

where $\mathcal{G}_n$ denotes the set of positions belonging to segment $\boldsymbol{g}_n$. These constraints regulate local concentration within the sequence and temporal concentration across decoding steps, respectively.

Exceeding the segment-level threshold stops further selection from that segment, while exceeding the sequence-level threshold ends selection for the current decoding step, after which the selected tokens are decoded in parallel. The overall selection procedure is summarized in Algorithm 1.

4 Interpreting Attention-Guided Token Selection

Image attention as a signal for inter-token dependency. The dependency error defined in Equation 4 occurs when tokens decoded together are not independent. In other words, if the prediction of one token changes significantly when conditioned on another token, parallel decoding can introduce larger errors by predicting them independently. This suggests that analyzing parallel prediction error requires considering not only prediction confidence but also each token's influence on the predictions of the remaining masked tokens. Let X^i denote the random variable at position i, and $\boldsymbol{X}^{-i}$ the remaining masked-token random vector. The dependency between X^i and $\boldsymbol{X}^{-i}$ can be expressed by the mutual information:

$$\begin{aligned} I(X^i; \boldsymbol{X}^{-i}) &= \mathbb{E}_{\boldsymbol{X}^{-i}} \left[D_{\text{KL}} \left(p(X^i \mid \boldsymbol{X}^{-i}) \| p(X^i) \right) \right] \\ &= H(\boldsymbol{X}^{-i}) - H(\boldsymbol{X}^{-i} \mid X^i). \end{aligned} \tag{11}$$

We hypothesize that revealing visually informative tokens substantially constrains the predictions of the remaining masked tokens and reduces their uncertainty. Although A_t^i is not a direct estimate of mutual information, this relationship supports its use as a lightweight proxy for token dependency. Thus, we upper-bound the parallel prediction error using $I(X^i; \boldsymbol{X}^{-i})$ and approximate this bound with the image-attention mass A_t^i, leading to $M(\mathcal{S})$.

Attention-based approximation of parallel prediction error. For a set of tokens $\mathcal{S} = \{i_1, \ldots, i_k\}$ decoded in parallel, the dependency error is represented by $I(X^{i_j}; \boldsymbol{X}^{i<j})$. Since $\boldsymbol{X}^{i<j}$ includes a subset of the variables in $\boldsymbol{X}^{-i_j}$, each term is upper-bounded by the dependency between X^{i_j} and $\boldsymbol{X}^{-i_j}$:

$$\begin{aligned} \mathcal{E}(\mathcal{S}) = \sum_{j=2}^{k} I\big(X^{i_j}; \boldsymbol{X}^{i<j}\big) &\le \sum_{j=2}^{k} I\big(X^{i_j}; \boldsymbol{X}^{-i_j}\big) \\ &= \sum_{i \in \mathcal{S}} I(X^i; \boldsymbol{X}^{-i}) - \max_{i \in \mathcal{S}} I(X^i; \boldsymbol{X}^{-i}), \end{aligned} \tag{12}$$

where the last equality follows by choosing $i_1 = \arg\max_{i \in \mathcal{S}} I(X^i; \boldsymbol{X}^{-i})$.

This bound provides a dependency-based interpretation of $M(\mathcal{S})$ as an attention-based counterpart to the dependency quantity on the right-hand side of Equation 12.

Hierarchical Constraints on Token Selection. Constraining $M(\mathcal{S})$ over the entire sequence does not account for how the selected high-attention tokens are distributed across sequence positions. Even with the same $M(\mathcal{S})$, selecting adjacent tokens can induce stronger local dependency than selecting tokens that are farther apart.

To account for this effect, VIVA forms local segments and separately constrains the cumulative attention mass within each segment using $M(\mathcal{S} \cap \mathcal{G}_n)$. When dependencies are stronger within local groups, applying group-specific bounds can also yield a tighter dependency-based bound than using a single sequence-level bound. We provide the corresponding analysis in Appendix C.2.

Table 1: Experimental results on VQAv2 and MMBench (Accuracy) and DocVQA (ANLS) at decoding-step ratios $r \in \{0.125, 0.25, 0.5\}$. Best and second-best results within each model and setting are shown in **bold** and underline, respectively.

Model	Method	VQAv2			DocVQA			MMBench		
		0.125	0.25	0.5	0.125	0.25	0.5	0.125	0.25	0.5
LaViDa	Confidence	59.60	59.00	61.04	46.65	52.40	55.77	43.00	47.80	56.40
	Entropy	59.46	59.08	60.68	46.66	52.25	54.77	34.40	40.00	48.00
	VRCD	59.92	61.02	62.14	50.71	53.44	55.75	42.80	52.20	51.60
	VIVA (Ours)	**62.82**	**63.98**	**64.20**	**55.07**	**54.53**	**56.39**	**51.40**	**59.80**	**64.20**
MMaDA	Confidence	52.58	52.42	53.88	6.08	6.59	7.48	57.40	56.40	57.00
	Entropy	52.58	52.42	53.70	5.88	6.24	7.66	**57.60**	56.40	57.00
	VRCD	52.60	52.36	53.92	5.90	6.77	**7.89**	56.40	56.60	56.60
	VIVA (Ours)	**53.40**	**53.72**	**56.26**	**6.97**	**6.96**	7.40	56.80	**57.40**	**57.40**
LLaDA-V	Confidence	67.50	69.32	70.72	79.25	**82.34**	81.89	58.80	64.40	65.80
	Entropy	65.74	70.36	70.40	78.35	82.15	82.00	52.40	66.60	66.00
	VRCD	67.26	**71.58**	71.54	80.26	82.07	**84.18**	**61.60**	**67.20**	66.60
	VIVA (Ours)	**69.48**	70.28	**71.84**	**80.49**	80.82	83.33	61.20	65.20	**66.80**

While the sequence-level constraint limits the amount of visual information decoded within each step and thereby disperses it across the decoding trajectory, the segment-level constraint additionally accounts for its positional concentration within the sequence. Together, the two constraints distribute visual information both across decoding steps and across sequence positions.

5 Experiments

5.1 Experiment Setup

Models. We evaluate VIVA on three representative open-source dMLLMs: LaViDa-Reason (Li et al., 2026a), MMaDA-8B-MixCoT (Yang et al., 2026b), and LLaDA-V (You et al., 2025).

Benchmarks. We evaluate VIVA on six vision-language benchmarks: VQAv2 (Goyal et al., 2017), DocVQA (Mathew et al., 2020), and MMBench (Liu et al., 2024) for multimodal understanding and reasoning, and DetailCaps (Dong et al., 2024), COCO Caption (Lin et al., 2014), and Flickr30K (Young et al., 2014) for image description and captioning. We report accuracy for VQAv2 and MMBench, ANLS for DocVQA, CAPTURE for DetailCaps, and CIDEr for COCO Caption and Flickr30K. All evaluations are conducted using the lmms-eval framework (Zhang et al., 2025) on the lite subsets, with DetailCaps evaluated on 500 samples.

Baselines. We compare VIVA with two certainty-based token selection strategies, Confidence and Entropy, which prioritize masked positions based on top-1 prediction probability and predictive entropy, respectively. We additionally compare against VRCD (Yuan et al., 2026), a training-free dMLLM decoding method that controls visual redundancy among tokens selected for parallel decoding.

Implementation Details. For COCO Caption and Flickr30K, we set the generation length L to 32 for LaViDa-Reason and MMaDA and 16 for LLaDA-V, while using 128 for all models on the remaining benchmarks. Across all models and benchmarks, we evaluate decoding-step ratios $r = T/L \in \{0.125, 0.25, 0.5\}$, where T denotes the total number of decoding steps. For VIVA, we set $\lambda = 20$ on COCO Caption and Flickr30K and $\lambda = 100$ on the remaining benchmarks, while evaluating VRCD with its default configuration. All experiments use greedy decoding with zero temperature and are conducted on NVIDIA RTX A5000 GPUs.

Table 2: Experimental results on DetailCaps (CAPTURE) and COCO Caption and Flickr30K (CIDEr) at decoding-step ratios $r \in \{0.125, 0.25, 0.5\}$. Best and second-best results within each model and setting are shown in **bold** and underline, respectively.

Model	Method	DetailCaps			COCO Caption			Flickr30K		
		0.125	0.25	0.5	0.125	0.25	0.5	0.125	0.25	0.5
LaViDa	Confidence	38.67	48.73	**52.62**	67.50	83.63	92.55	42.10	54.99	59.26
	Entropy	39.08	43.84	42.94	63.77	80.82	88.11	43.08	52.01	57.07
	VRCD	41.11	**50.69**	50.30	66.22	87.88	93.27	42.30	55.14	**61.02**
	VIVA (Ours)	**41.99**	43.73	49.75	**85.97**	**89.03**	**95.37**	**52.90**	**56.55**	60.82
MMaDA	Confidence	35.60	27.80	17.27	62.81	**83.25**	87.92	41.37	56.79	58.08
	Entropy	40.39	37.92	23.18	60.56	79.26	83.44	41.34	54.29	56.19
	VRCD	40.31	37.79	28.64	63.52	82.84	**88.68**	45.14	**59.49**	60.93
	VIVA (Ours)	**43.89**	**44.18**	**43.86**	**77.05**	82.71	86.87	**51.81**	57.86	**62.40**
LLaDA-V	Confidence	48.13	58.03	58.77	71.85	**89.13**	92.55	61.54	76.28	**84.10**
	Entropy	42.58	55.85	58.39	66.54	85.58	91.78	59.78	71.44	77.37
	VRCD	50.53	58.45	**59.52**	70.93	87.66	**94.00**	**61.56**	76.97	83.55
	VIVA (Ours)	**55.48**	**58.98**	59.39	**72.09**	87.89	92.43	57.92	**77.55**	80.05

5.2 Main Results

Multimodal Understanding and Reasoning. Table 1 shows the performance of VIVA on multimodal understanding and reasoning tasks. VIVA improves over confidence-based decoding in 24 of the 27 evaluated settings across the three model backbones, demonstrating consistent gains across different tasks and decoding configurations. The improvements are particularly clear on LaViDa, where VIVA achieves the best performance across all decoding-step ratios and benchmarks. Notably, at $r = 0.125$, VIVA outperforms the strong VRCD baseline by 4.36 points on DocVQA and 8.60 points on MMBench.

Image Description and Captioning. Table 2 presents VIVA's performance on image description and captioning tasks. VIVA exhibits a particularly clear advantage under highly parallel decoding, achieving the best performance in eight of the nine settings at $r = 0.125$. For example, on LaViDa, VIVA outperforms VRCD by 19.75 CIDEr points on COCO Caption and 10.60 points on Flickr30K, while corresponding gains of 13.53 and 6.67 points are observed on MMaDA. These results support the effectiveness of visual-information dispersion when multiple tokens are decoded in parallel.

6 Analysis

6.1 Effect of hierarchical constraints.

To isolate the effect of hierarchical constraints from attention-guided token reordering, we evaluate three variants with different levels of constraint on token selection after reordering: unconstrained selection, sequence-level constrained selection, and hierarchically constrained selection. As shown in Table 3, the effect of hierarchical constraints becomes more pronounced as decoding becomes more parallel. At $r = 0.5$, attention-guided reordering alone remains competitive, whereas at $r = 0.125$, the full hierarchical constraint improves COCO CIDEr from 68.32 to 85.97 and MMBench accuracy from 49.40 to 51.40. The sequence-level constraint alone does not yield comparable gains, suggesting that the segment-level constraint plays an important role under highly parallel decoding.

6.2 Hyperparameter sensitivity.

We analyze the sensitivity of VIVA to the guidance strength λ, which controls the contribution of A_t^i to the token selection score, using 500 samples from COCO Caption and MMBench on LaViDa-Reason. As shown in Table 4, VIVA generally outperforms confidence-based decoding and VRCD

Table 3: Ablation study of the hierarchical constraints in VIVA on LaViDa-Reason.

Bound	COCO Caption			MMBench		
	0.125	0.25	0.5	0.125	0.25	0.5
w/o Constraint	68.32	84.79	91.31	49.40	**60.00**	**65.40**
Sequence-level Constraint	70.73	87.96	93.53	44.40	58.40	64.40
Hierarchical Constraint	**85.97**	**89.03**	**95.37**	**51.40**	59.80	64.20

Table 4: Sensitivity analysis of VIVA with respect to λ on LaViDa-Reason.

Method	λ	COCO Caption			MMBench		
		0.125	0.25	0.5	0.125	0.25	0.5
Confidence	–	67.50	83.63	92.55	43.00	47.80	56.40
VRCD	–	66.22	87.88	93.27	42.80	52.20	51.60
VIVA	100	83.83	86.69	91.73	**51.40**	**59.80**	**64.20**
	50	85.20	89.38	93.36	48.80	57.00	60.80
	20	**85.97**	89.03	95.37	46.20	52.20	55.20
	10	85.89	**89.89**	**98.42**	44.20	49.00	52.00

Table 5: End-to-end inference cost on COCO Caption with LaViDa, averaged over 100 samples.

Method	Peak mem. (GB)	Wall-clock time (s)		
		k=4	k=8	k=16
Confidence	16.59	0.928	1.132	1.468
VRCD	16.59	0.946	1.125	1.493
VIVA	16.59	0.962	1.145	1.539

across a broad range of λ values on both datasets. These results indicate that the performance gains of VIVA are not tied to a particular choice of guidance strength, demonstrating robustness to λ.

6.3 Inference Cost Analysis

We evaluate the inference time and GPU memory usage of VIVA. All measurements are performed on COCO Caption with LaViDa-Reason over 100 samples. For each sample, we measure the full inference pipeline from image preprocessing through generation and text decoding, with CUDA synchronization before and after timing. As shown in Table 5, VIVA incurs only a small increase in wall-clock time while requiring essentially the same peak GPU memory as confidence-based decoding. These results show that VIVA introduces little additional cost to the overall inference pipeline.

7 Conclusion

In this work, we identify two key issues in dMLLM decoding under visual conditioning: the late decoding of visually relevant tokens and prediction errors caused by their simultaneous decoding. We propose VIVA, an inference-time decoding method that uses token-to-image attention to disperse visual information across decoding steps while preserving parallel decoding. Experiments across six benchmarks and three open-source dMLLMs demonstrate consistent improvements in generation quality, supported by additional analyses and qualitative results. Further analyses highlight the contribution of hierarchical constraints and show that VIVA introduces only a small increase in inference time with essentially unchanged peak GPU memory usage.

AI USE STATEMENT

In this work, we used generative AI tools to assist with translation, manuscript revision, readability improvement, and identification of relevant literature. We did not use generative AI to generate experimental data or results. All AI-assisted content was reviewed and verified by the authors. We take responsibility for the final content of this work.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Heli Ben-Hamu, Itai Gat, Daniel Severo, Niklas S Nolte, and Brian Karrer. Accelerated sampling from masked diffusion models via entropy bounded unmasking. *Advances in Neural Information Processing Systems*, 38:55981–56007, 2026.

Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. Benchmarking and improving detail image caption. *arXiv preprint arXiv:2405.19092*, 2024.

Hengyu Fu, Baihe Huang, Virginia Adams, Charles Wang, Venkat Srinivasan, and Jiantao Jiao. From bits to rounds: Parallel decoding with exploration for diffusion language models. *arXiv preprint arXiv:2511.21103*, 2025.

Shansan Gong, Shivam Agarwal, Yizhe Zhang, Jiacheng Ye, Lin Zheng, Mukai Li, Chenxin An, Peilin Zhao, Wei Bi, Jiawei Han, et al. Scaling diffusion language models via adaptation from autoregressive models. In *International Conference on Learning Representations*, volume 2025, pp. 5046–5073, 2025.

Google DeepMind. Gemini diffusion. `https://deepmind.google/models/gemini-diffusion/`, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Michael Hersche, Nicolas Menet, Ronan Tanios, and Abbas Rahimi. Locally coherent parallel decoding in diffusion language models. *arXiv preprint arXiv:2603.20216*, 2026.

Sujung Hong, Chanyong Yoon, and Seong Jae Hwang. Mitigating mask prior drift and positional attention collapse in large diffusion vision-language models. *arXiv preprint arXiv:2605.14530*, 2026.

Pengcheng Huang, Shuhao Liu, Zhenghao Liu, Yukun Yan, Shuo Wang, Zulong Chen, and Tong Xiao. Pc-sampler: Position-aware calibration of decoding bias in masked diffusion models. *arXiv e-prints*, pp. arXiv–2508, 2025.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. *arXiv preprint arXiv:2503.03321*, 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.

Jiyeon Kim, Sungik Choi, Yongrae Jo, Moontae Lee, and Minjoon Seo. Early decisions matter: Proximity bias and initial trajectory shaping in non-autoregressive diffusion language models. *arXiv preprint arXiv:2604.10567*, 2026.

Fanheng Kong, Jingyuan Zhang, Yahui Liu, Zirui Wu, Yu Tian, Guorui Zhou, et al. Accelerating diffusion llm inference via local determinism propagation. *arXiv preprint arXiv:2510.07081*, 2025.

Sanghyun Lee, Seungryong Kim, Jongho Park, and Dongmin Park. Lookahead unmasking elicits accurate decoding in diffusion language models. *arXiv preprint arXiv:2511.05563*, 2025.

Shufan Li, Jiuxiang Gu, Kangning Liu, Zhe Lin, Zijun Wei, Aditya Grover, and Jason Kuen. Lavida-o: Elastic large masked diffusion models for unified multimodal understanding and generation. *arXiv preprint arXiv:2509.19244*, 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 language model for multimodal understanding. *Advances in Neural Information Processing Systems*, 38: 105101–105134, 2026a.

Shufan Li, Yuchen Zhu, Jiuxiang Gu, Kangning Liu, Zhe Lin, Yongxin Chen, Molei Tao, Aditya Grover, and Jason Kuen. Lavida-r1: Advancing reasoning for unified multimodal diffusion language models. *arXiv preprint arXiv:2602.14147*, 2026b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Minesh Mathew, Dimosthenis Karatzas, R Manmatha, and CV Jawahar. Docvqa: A dataset for vqa on document images. corr abs/2007.00398 (2020). *arXiv preprint arXiv:2007.00398*, 2020.

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. *Advances in Neural Information Processing Systems*, 38:50608–50646, 2026.

Qingyu Shi, Jinbin Bai, Zhuoran Zhao, Wenhao Chai, Kaidong Yu, Jianzong Wu, Yunhai Tong, Xiangtai Li, Xuelong Li, and Shuicheng Yan. Muddit: Liberating generation beyond text-to-image with a unified discrete diffusion model. *arXiv preprint arXiv:2505.23606*, 2025.

Yuxuan Song, Zheng Zhang, Cheng Luo, Pengyang Gao, Fan Xia, Hao Luo, Zheng Li, Yuehang Yang, Hongli Yu, Xingwei Qu, et al. Seed diffusion: A large-scale diffusion language model with high-speed inference. *arXiv preprint arXiv:2508.02193*, 2025.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Chengyue Wu, Hao Zhang, Shuchen Xue, Zhijian Liu, Shizhe Diao, Ligeng Zhu, Ping Luo, Song Han, and Enze Xie. Fast-dllm: Training-free acceleration of diffusion llm by enabling kv cache and parallel decoding. *arXiv preprint arXiv:2505.22618*, 2025.

Jingqi Xu, Jingxi Lu, Chenghao Li, Sreetama Sarkar, Souvik Kundu, and Peter A Beerel. Red-vtp: Training-free acceleration of diffusion vision-language models inference via masked token-guided visual token pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2783–2792, 2026.

Xinhao Xu, Hui Chen, Mengyao Lyu, Sicheng Zhao, Yizhe Xiong, Zijia Lin, Jungong Han, and Guiguang Ding. Mitigating hallucinations in multi-modal large language models via image token attention-guided decoding. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1571–1590, 2025.

Kaisen Yang, Jayden Teoh, Kaicheng Yang, Yitong Zhang, and Alex Lamb. Improving sampling for masked diffusion models via information gain. *arXiv preprint arXiv:2602.18176*, 2026a.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *Advances in Neural Information Processing Systems*, 38:138867–138907, 2026b.

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.

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.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Yulin Yuan, Hongshuo Zhao, and Xiangming Meng. Visual-redundancy-controlled parallel decoding for diffusion-based multimodal large language models. *arXiv preprint arXiv:2605.25820*, 2026.

Lunbin Zeng, Jingfeng Yao, Bencheng Liao, Hongyuan Tao, Wenyu Liu, and Xinggang Wang. Diffusionvl: Translating any autoregressive models into diffusion vision language models. In *European Conference on Computer Vision*, pp. 331–348. Springer, 2026.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 881–916, 2025.

Qicheng Zhao, Qi Sun, and Zheyu Yan. Seeing the end at step zero: Accelerating diffusion mllms via mlp sparsity-aware truncation. *arXiv preprint arXiv:2607.14557*, 2026.

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

Haomin Zuo, Yidi Li, Luoxiao Yang, and Xiaofeng Zhang. Diffusion-cam: Faithful visual explanations for dmllms. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12085–12101, 2026.

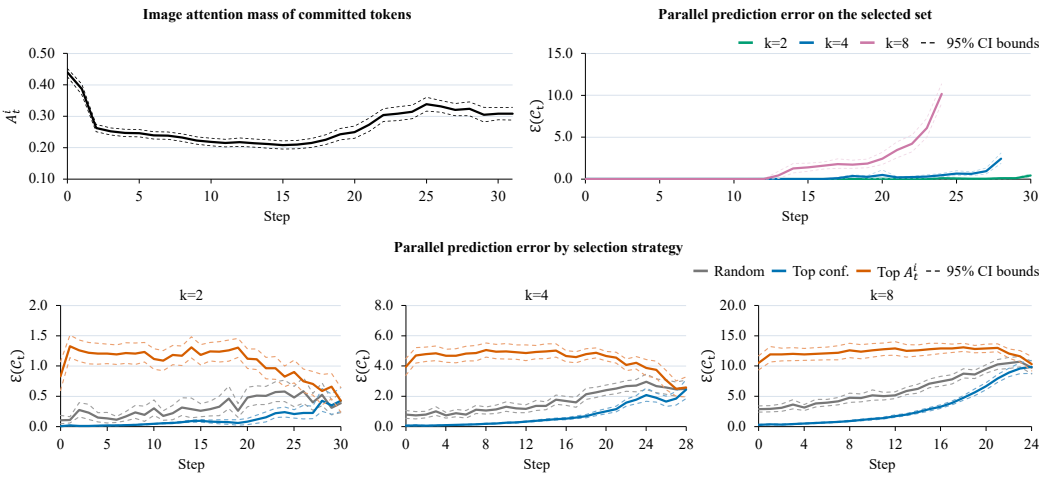

Figure 4: Image attention and parallel prediction error in MMaDA on the COCO captioning dataset. Top: Later decoding steps exhibit higher image attention and parallel prediction error. Bottom: Selecting tokens with higher image attention yields larger parallel prediction errors.

A Related Work

A.1 Diffusion Multimodal Large Language Models

Diffusion large language models (dLLMs) have emerged as an alternative to autoregressive models, leveraging iterative masked-token refinement to support flexible generation orders and parallel decoding. These properties have driven rapid progress in both closed-source (Khanna et al., 2025; Google DeepMind, 2025; Song et al., 2025) and open-source dLLMs (Nie et al., 2026; Zhu et al., 2025; Gong et al., 2025; Ye et al., 2025). This paradigm has further been extended to multimodal understanding, as exemplified by LaViDa and LLaDA-V (Li et al., 2026a; You et al., 2025). DiffusionVL further demonstrates that existing autoregressive vision-language models can be effectively adapted to diffusion-based generation (Zeng et al., 2026). Beyond multimodal understanding, recent models jointly model visual and textual tokens within unified diffusion frameworks, supporting both understanding and generation (Yang et al., 2026b; Shi et al., 2025; Li et al., 2025). Recent work on dMLLMs has explored several complementary directions. Some studies focus on multimodal reasoning (Li et al., 2026b), while others improve inference efficiency through visual-token pruning and sequence truncation (Xu et al., 2026; Zhao et al., 2026). In parallel, other work focuses on improving the interpretability of dMLLMs (Zuo et al., 2026).

A.2 Decoding Strategies in Diffusion Large Language Models

dLLMs generate text through iterative refinement, repeatedly predicting tokens for masked positions and decoding a subset of them at each denoising step. Conventional decoding strategies typically prioritize positions based on prediction certainty, measured by confidence or entropy (Ben-Hamu et al., 2026; Nie et al., 2026; Wu et al., 2025). Beyond these step-wise criteria, recent studies have explored token selection from a broader trajectory-level perspective. Some methods leverage decoding-order patterns or prediction consistency across denoising steps to adjust token selection (Huang et al., 2025; Kong et al., 2025), while others explicitly consider future unmasking paths (Lee et al., 2025). From an information-theoretic perspective, subsequent work selects tokens according to their contribution to reducing uncertainty in later predictions or promotes exploratory decoding to increase the information acquired across decoding rounds (Yang et al., 2026a; Fu et al., 2025). Recent dMLLM studies have begun to address modality-specific decoding issues, including visual redundancy in parallel decoding (Yuan et al., 2026) and mask prior drift and positional attention collapse (Hong et al., 2026). Despite these recent advances, the influence of visual input on token ordering and parallel decoding dynamics remains relatively underexplored.

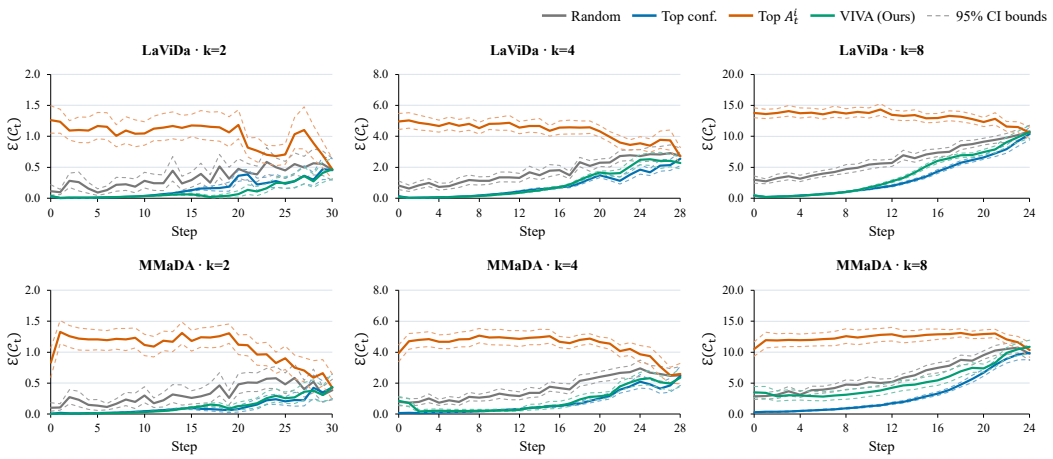

Figure 5: Parallel prediction error under different token selection strategies on LaViDa and MMaDA. VIVA maintains low error despite prioritizing high-image-attention tokens.

Table 6: Relationship between image attention and KL divergence on DetailCaps.

LaViDa		MMaDA	
A_t^i **range**	**Mean KL**	A_t^i **range**	**Mean KL**
0.02–0.16	0.19	0.01–0.13	0.05
0.16–0.21	0.27	0.13–0.19	0.14
0.21–0.26	0.40	0.19–0.25	0.26
0.26–0.32	0.63	0.25–0.31	0.41
0.32–0.52	1.06	0.31–0.79	0.48

B Additional Experiments and Analysis

B.1 Generalization of Observations Across Models

To examine whether our observations in Section 3.1 generalize beyond LaViDa-Reason, we repeat the same analysis using MMaDA on the COCO captioning dataset. We observe consistent trends: later decoding steps exhibit higher image-attention mass and parallel prediction error, and selecting tokens with the highest image-attention mass yields larger errors than random or confidence-based selection (see Figure 4).

B.2 Image attention and visual informativeness

Image attention has been widely used in MLLMs as a signal for identifying or guiding visually grounded generation (Kang et al., 2025; Xu et al., 2025). To further examine whether a high A_t^i reflects the contribution of image context to token prediction in dMLLMs, we analyze the change in predictive distributions induced by visual input.

We generated captions with LaViDa and MMaDA on 100 DetailCaps images and compared token predictive distributions under the original and white images while keeping the textual context fixed. We quantify their difference using KL divergence. As shown in Table 6, tokens with higher A_t^i tend to show larger KL divergence, suggesting that A_t^i captures the degree to which token prediction relies on visual context and can serve as a proxy for visual informativeness.

B.3 Effect of Hierarchical Constraints on Parallel Prediction Error

In Section 3.1, we observe that directly selecting tokens with high image-attention mass leads to substantially larger parallel prediction error than confidence-based or random selection. Here, we

Table 7: Sensitivity to generation length on LaViDa-Reason. Higher is better for all metrics.

Length	COCO Caption			MMBench		
	0.125	0.25	0.5	0.125	0.25	0.5
32	85.97	89.03	95.37	52.60	60.80	66.00
64	83.78	90.06	94.93	59.20	61.80	66.80
128	73.32	89.00	96.68	51.40	59.80	64.20

examine whether the hierarchical constraints in VIVA mitigate this error while still prioritizing visually informative tokens.

Figure 5 compares the parallel prediction error along the decoding trajectory for random selection, confidence-based selection, direct top-A_t^i selection, and VIVA. Although VIVA prioritizes tokens with high image-attention mass through attention-guided reordering, its parallel prediction error remains close to that of confidence-based and random selection across different values of k and on both LaViDa and MMaDA. In contrast, directly selecting the tokens with the largest A_t^i consistently results in substantially higher error.

These results show that attention-guided reordering alone can expose the model to strong inter-token dependencies when multiple visually informative tokens are decoded simultaneously, whereas the hierarchical constraints in VIVA prevent such tokens from being overly concentrated within the same decoding step. Thus, VIVA can prioritize visually informative tokens without incurring the large parallel prediction error associated with unconstrained top-A_t^i selection.

B.4 Generation length sensitivity.

We evaluate VIVA with different generation lengths on LaViDa-Reason. Table 7 reports the results on COCO Caption and MMBench for generation lengths of 32, 64, and 128.

B.5 Qualitative Examples

We provide qualitative examples comparing the decoding trajectories of VIVA with Confidence and VRCD in Figure 6. The baselines tend to postpone visually relevant content until later decoding steps, which can lead to repetition, incomplete descriptions, or incorrect relations. In contrast, VIVA introduces visually relevant tokens earlier and distributes them across the decoding trajectory, allowing decoded visual content to provide context for subsequent predictions. These examples qualitatively illustrate how VIVA promotes more coherent generation by dispersing visual information across decoding steps.

C Supporting Evidence for the Design Choices in VIVA

C.1 Image-attention-guided token sampling

Following the KL-regularized formulation of Lee et al. (2025), we formulate image-attention-guided token selection as follows. Given a base distribution $p_0(z)$ and a guidance score $r(z)$, we consider the KL-regularized objective

$$p^* = \arg\max_p \left[\mathbb{E}_{z\sim p}[r(z)] - \frac{1}{\lambda} D_{\mathrm{KL}}(p\|p_0) \right], \tag{13}$$

where λ controls the strength of the guidance relative to the KL regularization. The resulting distribution is

$$p^*(z) \propto p_0(z) \exp\left(\lambda\, r(z)\right). \tag{14}$$

For token selection at decoding step t, we use normalized prediction confidence as the base distribution and the normalized image-attention score $s_t^{\mathrm{attn}}(i)$ as guidance, yielding

$$p^*(i) \propto \frac{p(x_{t+1}^i = w_1)}{\sum_{j\in\mathcal{M}_t} p(x_{t+1}^j = w_1)} \exp\left(\lambda\, s_t^{\mathrm{attn}}(i)\right). \tag{15}$$

Since token selection depends only on the relative scores among masked positions, the normalization term in $p_0(i)$ can be omitted.

$$s_t^{\mathrm{vg}}(i) = \exp\left(\lambda s_t^{\mathrm{attn}}(i)\right) p(x_{t+1}^i = w_1), \tag{16}$$

which is the visual-guidance score used in Equation 7. A larger λ places greater emphasis on image attention when determining the decoding order.

C.2 Dependency-based analysis of hierarchical constraints

We provide a dependency-based interpretation of the hierarchical constraints used in VIVA. As in Section 4, we omit conditioning on the current decoding state for brevity.

Consider a set of masked positions $\mathcal{S}$ selected for parallel decoding. Recall from Section 4 that the dependency error satisfies

$$\mathcal{E}(\mathcal{S}) \le B_{\mathrm{seq}}(\mathcal{S}) := \sum_{i\in\mathcal{S}} I(X^i; \boldsymbol{X}^{-i}) - \max_{i\in\mathcal{S}} I(X^i; \boldsymbol{X}^{-i}). \tag{17}$$

This sequence-level bound depends only on the selected set $\mathcal{S}$ and does not account for how the selected positions are distributed across local segments.

Let $\{\mathcal{G}_n\}_{n=1}^N$ denote the segments defined in Section 3.2, and define

$$\mathcal{S}_n := \mathcal{S} \cap \mathcal{G}_n, \qquad \mathcal{A}(\mathcal{S}) := \{n : \mathcal{S}_n \neq \emptyset\}. \tag{18}$$

The dependency error can be decomposed exactly into within-segment and cross-segment dependencies:

$$\mathcal{E}(\mathcal{S}) = \sum_{n\in\mathcal{A}(\mathcal{S})} \mathcal{E}(\mathcal{S}_n) + R(\mathcal{S}), \tag{19}$$

where

$$R(\mathcal{S}) := D_{\mathrm{KL}}\left(p(\boldsymbol{X}^{\mathcal{S}}) \,\middle\|\, \prod_{n\in\mathcal{A}(\mathcal{S})} p(\boldsymbol{X}^{\mathcal{S}_n})\right) \tag{20}$$

measures the residual dependency among the selected token groups from different segments.

Applying Equation 17 separately to each $\mathcal{S}_n$ gives

$$\mathcal{E}(\mathcal{S}) \le B_{\mathrm{hier}}(\mathcal{S}) := \sum_{n\in\mathcal{A}(\mathcal{S})} \left[\sum_{i\in\mathcal{S}_n} I(X^i; \boldsymbol{X}^{-i}) - \max_{i\in\mathcal{S}_n} I(X^i; \boldsymbol{X}^{-i})\right] + R(\mathcal{S}). \tag{21}$$

To compare the two bounds, let

$$m_n := \max_{i\in\mathcal{S}_n} I(X^i; \boldsymbol{X}^{-i}), \qquad m := \max_{i\in\mathcal{S}} I(X^i; \boldsymbol{X}^{-i}). \tag{22}$$

Their difference is

$$B_{\mathrm{seq}}(\mathcal{S}) - B_{\mathrm{hier}}(\mathcal{S}) = \sum_{n\in\mathcal{A}(\mathcal{S})} m_n - m - R(\mathcal{S}), \tag{23}$$

$$B_{\mathrm{hier}}(\mathcal{S}) < B_{\mathrm{seq}}(\mathcal{S}) \iff R(\mathcal{S}) < \sum_{n\in\mathcal{A}(\mathcal{S})} m_n - m. \tag{24}$$

Thus, the hierarchical bound can be tighter when cross-segment dependency is sufficiently small. This provides a dependency-based justification for applying the constraint separately across local segments.



Reference captions

"A box full of matching, ridged donuts with glaze."
"A BOX OF DOUGHNUTS JUST SHORT OF A DOZEN"

Confidence

Step 0	▪ ▪
Step 1	▪ ▪
Step 2	▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪
Step 3	**A tray of** ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ **.**
Step 4	A tray of **cinnamon on on on on on a table**.

VRCD

Step 0	▪ ▪
Step 1	▪ ▪
Step 2	▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪
Step 3	**A tray of** ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ **.**
Step 4	A tray of **cinnamon on on on on on a table**.

VIVA (Ours)

Step 0	▪ ▪
Step 1	**A ▪ of** ▪
Step 2	A **tray** of **cinnamon** ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪
Step 3	A tray of cinnamon **rolls ▪ a on** ▪ ▪ **.** ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪
Step 4	A tray of cinnamon rolls **on** a on **a table**.



Reference captions

"An old man making duck lips while holding an orange."
"A man puckering his lips holding an orange"

Confidence

Step 0	▪ ▪
Step 1	▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪
Step 2	▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪
Step 3	**A old** ▪ ▪ ▪ ▪ ▪ ▪ **orange orange** ▪
Step 4	A old **man with a face a an** orange orange**.**

VRCD

Step 0	▪ ▪
Step 1	▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪
Step 2	▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪ ▪
Step 3	**A old** ▪ ▪ ▪ **orange** ▪ ▪ ▪ **orange** ▪
Step 4	A old **man eating an** orange **and to an** orange**.**

VIVA (Ours)

Step 0	▪ ▪
Step 1	▪ **old with** ▪ ▪ ▪ ▪ **orange** ▪ ▪ ▪ ▪
Step 2	▪ old with ▪ **hair** ▪ **an** orange ▪ ▪ ▪ ▪
Step 3	▪ old with **white** hair **holding** an orange ▪ **his** ▪**.**
Step 4	**Man** old with white hair holding an orange **to** his **mouth**.

Figure 6: Qualitative comparison of VIVA with conventional parallel decoding methods.

Algorithm 1 Adaptive Token Selection in VIVA

Require: Masked positions $\mathcal{M}_t$, segment position sets $\{\mathcal{G}_n\}_{n=1}^{N}$, image-attention masses $\{A_t^i\}_{i\in\mathcal{M}_t}$, ordering scores $\{s_t^{\mathrm{vg}}(i)\}_{i\in\mathcal{M}_t}$, total decoding steps T
Ensure: Selected positions $\mathcal{C}_t$

$\mathcal{C}_t \leftarrow \varnothing$
$\mathcal{H}_t \leftarrow \varnothing$
for $n = 1, \ldots, N$ **do**
    $\tau_{t,n}^{\mathrm{seg}} \leftarrow \dfrac{M(\mathcal{M}_t \cap \mathcal{G}_n)}{T - t}$
end for
$\tau_t^{\mathrm{seq}} \leftarrow \dfrac{M(\mathcal{M}_t)}{T - t}$
Sort $\mathcal{M}_t$ in descending order of $s_t^{\mathrm{vg}}(i)$
for each i in the sorted order **do**
    Let $n(i)$ denote the segment index such that $i \in \mathcal{G}_{n(i)}$
    if $n(i) \in \mathcal{H}_t$ **then**
        continue
    end if
    if $M(\mathcal{C}_t \cup \{i\}) > \tau_t^{\mathrm{seq}}$ **then**
        break
    end if
    if $M((\mathcal{C}_t \cap \mathcal{G}_{n(i)}) \cup \{i\}) > \tau_{t,n(i)}^{\mathrm{seg}}$ **then**
        $\mathcal{H}_t \leftarrow \mathcal{H}_t \cup \{n(i)\}$
        continue
    end if
    $\mathcal{C}_t \leftarrow \mathcal{C}_t \cup \{i\}$
end for
return $\mathcal{C}_t$