VISUAL INFORMATION-GUIDED PARALLEL DECODING FOR DIFFUSION MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion multimodal large language models (dMLLMs) have recently emerged as a new decoding paradigm for multimodal generation. Starting from a fully masked sequence, dMLLMs progressively decode the sequence by unmasking a subset of the remaining masked positions at each step. Since the selected tokens serve as the prediction context for subsequent steps, deciding which tokens to decode is crucial to the quality of the final output. The most common strategy prioritizes tokens based on a certainty measure that tends to favor tokens frequently observed in the training data. Recent approaches instead order tokens according to their influence on subsequent predictions, but do not explicitly account for the input image. We propose the Visual Information-Guided Sampler (VIG-Sampler), which prioritizes tokens based on their attention to image tokens. We further impose a constraint that penalizes candidate tokens whose image-attention distributions are similar to those of previously selected tokens, thereby increasing the information gain of the decoded subset. Extensive experiments on 7 captioning and VQA benchmarks with 3 open-source dMLLMs demonstrate the effectiveness of VIG-Sampler, which outperforms the Info-Gain Sampler by an average of 19.3 CIDEr points across the captioning benchmarks and surpasses it on COCO Caption while using only half as many decoding steps.

1 INTRODUCTION

Recent advances in diffusion large language models (dLLMs) have led to the emergence of diffusion multimodal large language models (dMLLMs) (Li et al., 2026; You et al., 2026; Yang et al., 2026b). By extending the diffusion-based text generation paradigm to multimodal inputs, dMLLMs provide a promising alternative to the conventional *autoregressive* multimodal large language models (AR-MLLMs). Unlike AR-MLLMs generating a sequence by decoding tokens from left to right, dMLLMs generate a sequence through an iterative decoding process without following a fixed positional order (Kim et al., 2025). In this decoding process, the choice of which tokens to decode is of great importance, as the resulting tokens become part of the conditioning context for the subsequent prediction and ultimately affect the quality of the final output (Gwak et al., 2025; Jazbec et al., 2025).

The commonly used strategy for this token selection is to rank tokens according to certainty measures derived from the model's predictions and then process them in the ordered sequence (Wu et al., 2025; Nie et al., 2026; Ben-Hamu et al., 2026). One well-known drawback is that this certainty-based selection does not account for the usefulness of the decoded tokens in predicting the remaining masked tokens (Fu et al., 2025b). Since models tend to assign high confidence to tokens frequently observed in the training data, semantically uninformative tokens such as copulas, punctuation marks, and end-of-text tokens are often decoded first in the early steps (Martinez et al., 2024; Huang et al., 2026). In such cases, these tokens fix the syntactic structure or ending of the response before the model determines what content to generate. Thus, subsequent predictions are constrained to follow the prematurely determined structure, potentially generating a structurally consistent but incorrect response (Huang et al., 2025; Yang et al., 2026a).

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.

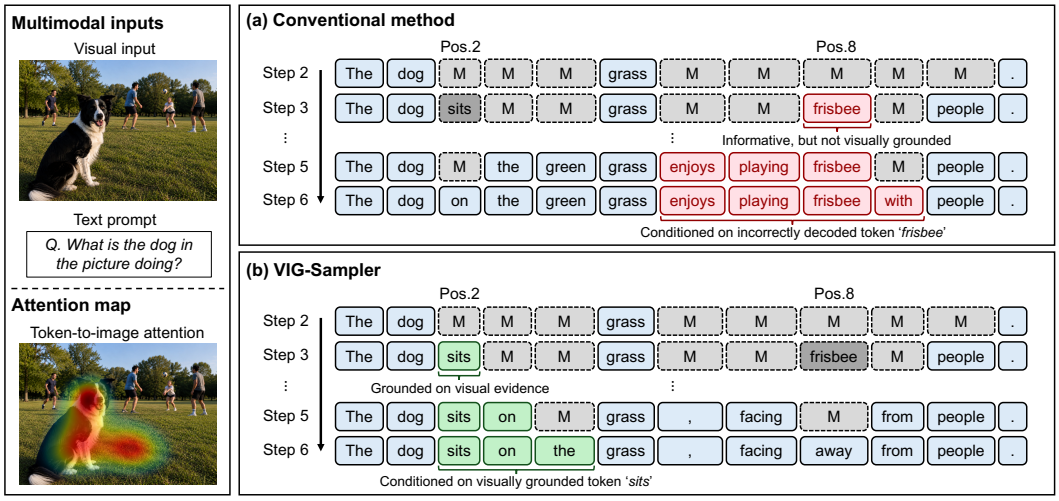


Figure 1: Comparison of conventional sampling and VIG-Sampler. The conventional method may prioritize the visually unsupported token *"frisbee"*, causing subsequent predictions to drift from the image. In contrast, VIG-Sampler uses image attention to prioritize the visually grounded token *"sits"*, leading to a response consistent with the visual input.

To overcome the shortcoming, several token selection approaches that account for the influence of decoded tokens on subsequent predictions have been proposed. One approach uses token frequencies in a fixed reference corpus to mitigate the excessive prevalence of less informative tokens in generated sequences (Huang et al., 2025). Another approach selects tokens based on each token's contribution to reducing sequence-level uncertainty (i.e., information gain). In this scheme, the contribution is quantified by the decrease in entropy over the remaining masked tokens after the token is decoded (Yang et al., 2026a). Although these approaches have improved the generation quality in language-only settings to some extent, they do not explicitly account for the input image and thus generate tokens without sufficient visual grounding.

To illustrate this scenario, consider a dMLLM is given an image of a sitting dog along with the question, *"What is the dog in the picture doing?"* (see Figure 1). At decoding step 3, the model may predict *"sits"* at position 2 and *"frisbee"* at position 8. Conventional token selection approaches may decode *"frisbee"* instead of *"sits"* if decoding the token at position 8 is estimated to yield greater information gain, even though the token is not grounded in the visual content. Once decoded, *"frisbee"* becomes a part of the context for subsequent predictions and may increase the likelihood of tokens such as *"playing"* and *"with"*. The model can thus generate a response inconsistent with the image, such as *"The dog on the green grass enjoys playing frisbee with people."* This example highlights a central principle of dMLLM decoding: visual evidence and language-based signals should be considered jointly.

An aim of this paper is to propose a novel image-grounded sampling strategy for parallel decoding in dMLLMs to select token sets that are both visually grounded and informative for subsequent predictions. To this end, we propose the Visual Information-Guided Sampler (VIG-Sampler), which leverages image attention readily available at each decoding step. Based on the observation that attention to image tokens serves as a reliable proxy for both visual grounding and informativeness, VIG-Sampler employs each token's image attention as an ordering score. This prioritizes visually informative tokens and thereby promotes an image-grounded decoding trajectory. In addition to the token-level ordering, VIG-Sampler increases the overall information gain from decoding multiple tokens in parallel. We empirically observe that set-level information gain can fall substantially below the sum of individual gains when selected tokens convey redundant information, and that this redundancy manifests as similarity between their token-to-image attention distributions. VIG-Sampler further translates this similarity into a penalty score, discouraging tokens with attention distributions similar to those of the tokens already selected.

Overall, VIG-Sampler jointly promotes visual grounding and complementary information gain during parallel decoding, without requiring additional training or model forward passes. Extensive ex-

periments on 7 captioning and visual question answering (VQA) benchmarks with 3 open-source dMLLMs show that VIG-Sampler consistently outperforms recent samplers that rely solely on the textual context, regardless of the number of tokens decoded per step. In particular, when decoding 8 tokens per step with LaViDa (Li et al., 2026), VIG-Sampler outperforms the Info-Gain Sampler by an average of 19.3 CIDEr points on captioning benchmarks and 7.3 accuracy points on VQA benchmarks. Notably, on COCO captioning (Lin et al., 2014), VIG-Sampler decoding 8 tokens per step still surpasses the Info-Gain Sampler decoding only 4 tokens per step by 5.3 CIDEr points, despite using half as many decoding steps. These results demonstrate the effectiveness of VIG-Sampler and show that grounding token selection in visual information is essential for dMLLMs.

In summary, our contributions are as follows:

- Building on our analytical findings, we develop VIG-Sampler, a decoding strategy that leverages readily available token-to-image attention to select visually grounded and informative tokens without incurring additional training or model forward passes.
- Through empirical analysis, we show that tokens decoded in parallel may provide redundant information, limiting the information gain of the selected set, and incorporate this observation into VIG-Sampler through attention-guided subset selection.
- Extensive experiments on 7 multimodal benchmarks with 3 open-source dMLLMs demonstrate that VIG-Sampler outperforms Info-Gain Sampler by 19.3 CIDEr points on image captioning and still achieves better performance when using only half the number of decoding steps.

2 Visual Information-Guided Sampler

2.1 Preliminaries

Decoding process of dMLLM. We consider a dMLLM p_θ that generates a textual response $\mathbf{y}$ conditioned on a visual input $\mathbf{I}$ and a text prompt $\mathbf{x}$. Starting from a length-N fully masked sequence, $\mathbf{y}_0 = [\text{MASK}]^N$, we perform T decoding steps to produce the final response $\mathbf{y}_T = \mathbf{y}$. At step t, the model predicts a distribution $p_\theta^i(\cdot \mid \mathbf{I}, \mathbf{x}, \mathbf{y}_t)$ for each masked position $i \in \mathcal{M}_t$, where $\mathcal{M}_t = \{i \in \{0, \dots, N-1\} \mid y_t^i = [\text{MASK}]\}$. For brevity, we write $p_\theta^i(\cdot \mid \mathbf{I}, \mathbf{x}, \mathbf{y}_t)$ as $p_\theta^i(\cdot \mid \mathbf{y}_t)$. From this distribution, we obtain $\hat{y}_t^i$ as the token with the highest probability and the confidence c_t^i as its corresponding probability:

$$\hat{y}_t^i = \arg\max_{v \in \mathcal{V}} p_\theta^i(v \mid \mathbf{y}_t), \qquad c_t^i = p_\theta^i(\hat{y}_t^i \mid \mathbf{y}_t), \tag{1}$$

where $\mathcal{V}$ denotes the vocabulary set. Based on these predictions, a subset $\mathcal{S} \subseteq \mathcal{M}_t$ of k positions is selected, and the selected positions are updated as $y_{t+1}^i = \hat{y}_t^i$ for $i \in \mathcal{S}$. A common strategy for selecting $\mathcal{S}$ is to choose the masked positions with the highest confidence scores (Nie et al., 2026).

Attention mechanism in dMLLM. At each decoding step t, the token embeddings of $\mathbf{I}$, $\mathbf{x}$, and $\mathbf{y}_t$ are concatenated into a single input sequence of length L. Transformer-based dMLLMs (Li et al., 2026; Yang et al., 2026b; You et al., 2026) process this sequence through multiple self-attention layers, enabling tokens to exchange information across modalities. For each layer and attention head, the self-attention matrix $A \in \mathbb{R}^{L \times L}$ is computed as

$$A = \text{softmax}\left(QK^\top\right), \tag{2}$$

where Q and K denote the query and key matrices, respectively. Each element $A_{i,j}$ represents the attention weight from query position i to key position j, where a larger value indicates that token i receives more information from token j (Vaswani et al., 2017; Zhang et al., 2025b). Throughout this work, we use A to denote the self-attention matrix from the last layer, averaged over all attention heads.

Information gain in token selection. In dLLM decoding, selecting informative tokens for subsequent predictions facilitates the decoding process (Fu et al., 2025a; Yang et al., 2026a; Huang et al., 2025). Such informativeness can be quantified using *information gain*, which measures the

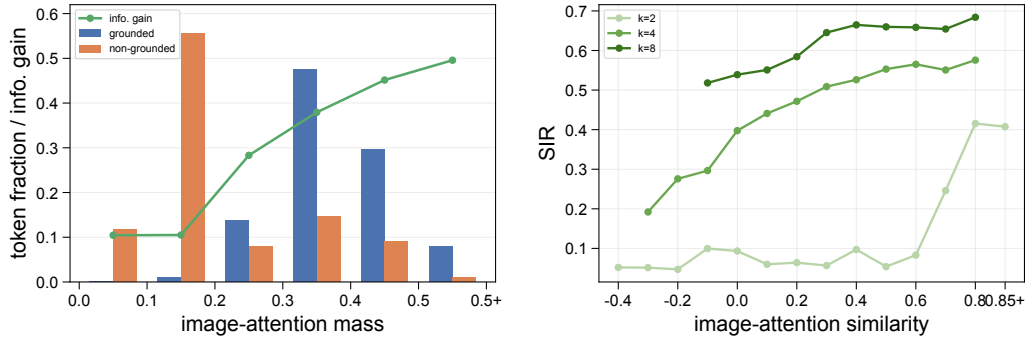


Figure 2: Motivating observations on how image attention relates to visual grounding and information gain during dMLLM decoding. Top: distributions of visually grounded and non-grounded tokens across image-attention-mass ranges, together with the mean information gain. Bottom: shared information ratio (SIR) versus mean pairwise image-attention similarity for selected token sets of size $k \in \{2, 4, 8\}$.

reduction in uncertainty of subsequent predictions (Yang et al., 2026a). To formalize this, for a set $\mathcal{S} \subseteq \mathcal{M}_t$, let $\mathcal{M}_t^{(\mathcal{S})} = \mathcal{M}_t \setminus \mathcal{S}$, and let $\mathbf{y}_t^{(\mathcal{S})}$ denote the response obtained by replacing y_t^i with $\hat{y}_t^i$ for $i \in \mathcal{S}$. For each $\mathbf{y} \in \{\mathbf{y}_t, \mathbf{y}_t^{(\mathcal{S})}\}$, we measure this uncertainty as the average entropy over $\mathcal{M}_t^{(\mathcal{S})}$:

$$U\big(\mathcal{M}_t^{(\mathcal{S})}; \mathbf{y}\big) = \frac{1}{|\mathcal{M}_t^{(\mathcal{S})}|} \sum_{i \in \mathcal{M}_t^{(\mathcal{S})}} H\big(p_\theta^i(\cdot \mid \mathbf{y})\big), \tag{3}$$

where $H(\cdot)$ denotes the entropy of a predicted distribution over tokens. The information gain $\Delta U_t(\mathcal{S})$ is defined as the reduction in this average entropy after unmasking $\mathcal{S}$:

$$\Delta U_t(\mathcal{S}) = U(\mathcal{M}_t^{(\mathcal{S})}; \mathbf{y}_t) - U\big(\mathcal{M}_t^{(\mathcal{S})}; \mathbf{y}_t^{(\mathcal{S})}\big). \tag{4}$$

2.2 Motivating Observations

When generating a response conditioned on multimodal inputs, existing approaches for selecting informative tokens face two practical limitations. First, evaluating information gain across many candidate token subsets requires an additional forward pass for each subset. This becomes computationally expensive with long multimodal inputs, as the cost of self-attention grows quadratically with sequence length. Second, information gain does not necessarily reflect how well a token is grounded in the visual input. A token can yield substantial information gain solely from linguistic priors or textual dependencies within the response. To address these limitations, we use the attention weights from masked tokens to image tokens (i.e., image attention) as a readily available signal for assessing token informativeness while accounting for the visual input (Kang et al., 2025; Nguyen et al., 2026; Jian et al., 2025). We conduct analyses to examine how image attention can guide token selection during dMLLM decoding, leading to the following two observations.

Observation 1: total image attention reflects both visual grounding and token informativeness. We examine how the amount of image attention relates to visual grounding and information gain. To quantify this amount, we define the image-attention mass as the sum of attention weights over all image tokens. Throughout the decoding process, we measure the image-attention mass and information gain $\Delta U_t(\{i\})$ for each selected position $i \in \mathcal{S}$. At the same time, we identify tokens that rely on visual information by examining whether $\hat{y}_t^i$ changes when the image is removed from the input. A change indicates dependence on visual evidence, in which case we classify the token as *grounded*; otherwise, we classify it as *non-grounded*. As shown in Figure 2 (left), higher image-attention mass corresponds to both higher information gain and a greater proportion of *grounded* tokens. This suggests that image-attention mass can serve as a signal for identifying tokens that are both visually grounded and informative.

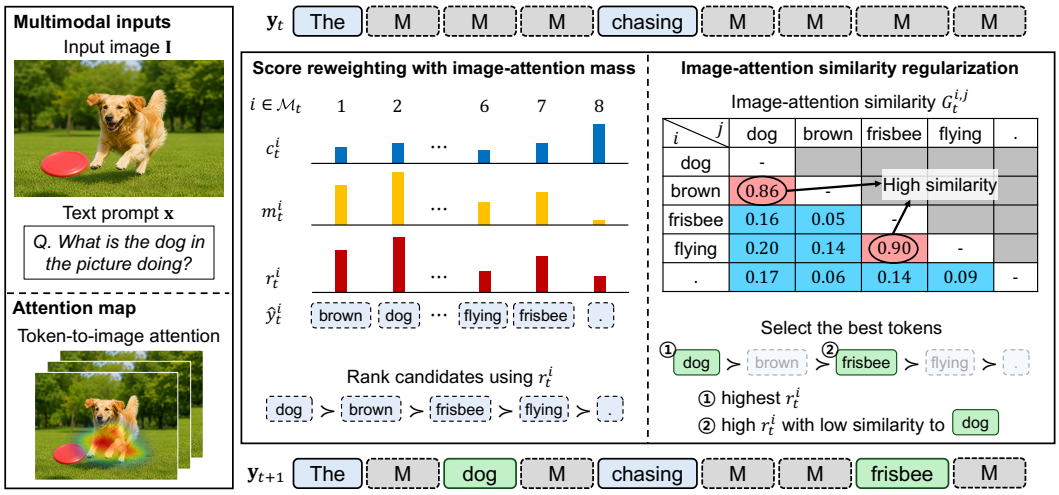


Figure 3: Overview of the proposed VIG-Sampler. At each decoding step, image-attention mass is used to adjust the confidence score of each masked position, prioritizing tokens that are both visually grounded and informative. The resulting scores and pairwise image-attention similarities are then used to select the token set expected to yield higher information gain.

Observation 2: similar image-attention patterns reflect shared information across tokens. In addition to the total amount of image attention, we consider the distribution of attention across image tokens. Intuitively, tokens with similar image-attention patterns may rely on shared visual evidence and thus provide overlapping information when decoded together. To quantify this overlap, we define the shared information ratio (SIR) as a measure that compares the information gain from decoding the tokens in $\mathcal{S}$ together with the sum of the information gains obtained by decoding each token individually:

$$\mathrm{SIR}_t(\mathcal{S}) = 1 - \frac{\Delta U_t(\mathcal{S})}{\sum_{i \in \mathcal{S}} \Delta U_t(\{i\})}. \tag{5}$$

A larger $\mathrm{SIR}_t(\mathcal{S})$ indicates a greater degree of information overlap among the tokens in $\mathcal{S}$. We then examine the relationship between image-attention similarity and SIR. For each subset size $k \in \{2, 4, 8\}$, we decode k tokens per step and compute the SIR and mean pairwise image-attention similarity of the selected subset $\mathcal{S}$. We then report the average SIR within each image-attention similarity range. As shown in Figure 2 (right), SIR increases with image-attention similarity, consistent with the intuition that similar attention patterns reflect shared information across tokens. This suggests that image-attention distributions can provide a cue for selecting token sets that yield higher information gain when decoded together. Additional experimental details for these observations are provided in the Appendix B.

2.3 VIG-Sampler

Motivated by these observations, we propose VIG-Sampler, a visually grounded decoding strategy for dMLLMs that leverages image attention as an additional cue. Specifically, VIG-Sampler guides token selection by prioritizing visually grounded and informative tokens based on image-attention mass, while penalizing high image-attention similarity within the selected set to reduce information overlap. An overview of VIG-Sampler is provided in Figure 3, with the overall decoding procedure detailed in the Appendix B.

Score reweighting with image-attention mass. Based on Observation 1, we use image-attention mass to reweight the confidence score c_t^i used for masked-token selection. For each masked position $i \in \mathcal{M}_t$, let $\mathbf{a}_t^i$ denote the attention weights from position i to the image-token positions $\mathcal{I}$, and let

m_t^i denote the image-attention mass:

$$\mathbf{a}_t^i = A_{i,\mathcal{I}}, \qquad m_t^i = \|\mathbf{a}_t^i\|_1 = \sum_{j\in\mathcal{I}} A_{i,j}. \tag{6}$$

To reflect the relative magnitude of m_t^i, we normalize it by m_t^{med}, the median image-attention mass over all masked positions. We then use the normalized image-attention mass to reweight the confidence score:

$$r_t^i = c_t^i \left(\frac{m_t^i}{m_t^{\text{med}}}\right)^{\gamma}, \tag{7}$$

where $\gamma \geq 0$ is a hyperparameter that controls the strength of image-attention guidance. This formulation gradually incorporates image-attention guidance into the confidence score as γ increases, starting from $r_t^i = c_t^i$ at $\gamma = 0$. For $\gamma > 0$, positions with higher image-attention mass receive larger weights, thereby prioritizing tokens that attend more strongly to the image. In this way, VIG-Sampler favors visually grounded and informative tokens while retaining the model's original preference (i.e., confidence-based preference). The resulting scores are then used in the subsequent set-selection objective.

Set selection via image-attention similarity regularization. Building on Observation 2, we use pairwise image-attention similarity as a regularization term when selecting the token set $\mathcal{S}$ at each decoding step. To better capture how each image-attention pattern differs from the overall pattern, we center them by subtracting the mean over $\mathcal{M}_t$:

$$\tilde{\mathbf{a}}_t^i = \mathbf{a}_t^i - \frac{1}{|\mathcal{M}_t|}\sum_{j\in\mathcal{M}_t} \mathbf{a}_t^j. \tag{8}$$

Using $\tilde{\mathbf{a}}_t^i$, we compute the pairwise image-attention similarity as

$$G_t^{i,j} = \max\left\{\left\langle \tilde{\mathbf{a}}_t^i, \tilde{\mathbf{a}}_t^j \right\rangle_{\cos}, 0\right\}, \tag{9}$$

where $\langle\cdot,\cdot\rangle_{\cos}$ denotes cosine similarity. We then formulate the set-selection objective based on the reweighted token scores while accounting for pairwise image-attention similarity:

$$\mathcal{S}_t^\star = \arg\max_{\substack{\mathcal{S}\subseteq\mathcal{M}_t \\ |\mathcal{S}|=k}} \left[\sum_{i\in\mathcal{S}} r_t^i - \frac{\lambda}{|\mathcal{S}|-1}\sum_{\substack{i,j\in\mathcal{S} \\ i<j}} G_t^{i,j}\right], \tag{10}$$

where $\lambda \geq 0$ is a hyperparameter controlling the strength of the similarity regularization. The second term is defined as zero for $|\mathcal{S}| = 1$. Since $G_t^{i,j}$ is non-negative, it acts solely as a penalty for similar image-attention patterns rather than rewarding dissimilar ones. VIG-Sampler thus discourages selecting tokens with similar image-attention patterns, reducing information overlap within the selected set.

Finding the optimal set $\mathcal{S}_t^\star$ in Equation 10 requires searching over a combinatorial number of candidate subsets, which is computationally impractical. We therefore use a greedy algorithm that selects the next position to maximize the marginal gain in the objective given the currently selected set (see the Appendix B for its derivation). After initializing $\mathcal{S}$ to $\emptyset$, we select the first position solely based on r_t^i. We then select the next position $i^\star$ as

$$i^\star = \arg\max_{i\in\mathcal{M}_t^{(\mathcal{S})}} \left[r_t^i - \frac{\lambda}{|\mathcal{S}|}\sum_{j\in\mathcal{S}} G_t^{i,j}\right]. \tag{11}$$

We iteratively add the selected position $i^\star$ to $\mathcal{S}$ until $|\mathcal{S}| = k$. The resulting set $\mathcal{S}$ determines the positions to be decoded with their predicted tokens $\hat{y}_t^i$ at the current step.

Table 1: Experimental results of LaViDa with parallel decoding budgets $k \in \{2, 4, 8\}$. The first and last Avg. columns report mean CIDEr and VQA scores, respectively. The best scores are shown in **bold**, and the second-best scores are underlined.

k	Sampler	COCO Cap.	Flickr30K	NoCaps	Avg.	DetailCaps	TextVQA	DocVQA	ChartQA	Avg.
		CIDEr	*CIDEr*	*CIDEr*	*CIDEr*	*CAPTURE*	*Acc.*	*ANLS*	*Relaxed Acc.*	*Macro Avg.*
2	Confidence	100.1	69.7	88.4	86.1	56.2	55.2	60.1	56.0	57.1
	Entropy	95.4	68.2	83.9	82.5	56.5	54.6	59.8	56.0	56.8
	Margin	105.8	71.7	88.7	88.7	56.9	55.0	61.6	56.6	57.7
	MPD-PAC	95.1	70.1	87.2	84.1	56.9	56.0	**62.8**	56.2	58.3
	PC-Sampler	101.7	69.8	88.9	86.8	57.1	54.3	58.9	55.0	56.1
	Info-Gain	106.0	71.2	89.6	88.9	55.6	57.4	61.4	**57.0**	58.6
	VIG-Sampler	**106.2**	**74.7**	**90.5**	**90.5**	**57.7**	**58.7**	62.6	55.8	**59.0**
4	Confidence	93.2	63.8	78.2	78.4	49.5	47.4	59.6	53.0	53.3
	Entropy	91.2	63.5	78.4	77.7	45.9	46.0	57.7	52.6	52.1
	Margin	93.7	65.5	78.0	79.1	51.5	47.4	59.7	53.2	53.4
	MPD-PAC	92.4	65.5	79.4	79.1	49.0	47.9	60.5	53.6	54.0
	PC-Sampler	95.8	66.6	82.8	81.7	49.6	52.2	57.1	54.0	54.4
	Info-Gain	94.8	65.9	81.9	80.9	49.2	50.2	60.2	54.0	54.8
	VIG-Sampler	**105.1**	**71.2**	**87.4**	**87.9**	**55.9**	**56.8**	**62.3**	**55.8**	**58.3**
8	Confidence	74.5	49.0	58.9	60.8	29.2	41.7	54.1	50.0	48.6
	Entropy	71.6	49.4	58.7	59.9	27.4	41.7	53.8	49.8	48.4
	Margin	73.2	46.9	59.8	60.0	34.0	41.7	54.4	50.0	48.7
	MPD-PAC	73.4	49.0	59.7	60.7	29.9	40.7	55.7	50.4	48.9
	PC-Sampler	83.9	56.2	67.9	69.3	30.2	50.8	56.8	54.0	53.9
	Info-Gain	76.7	50.6	60.9	62.7	29.5	42.7	54.4	49.8	49.0
	VIG-Sampler	**100.1**	**63.6**	**82.3**	**82.0**	**42.4**	**53.1**	**61.5**	**54.2**	**56.3**

3 Experiments

3.1 Models and Baselines

Our evaluation covers three representative dMLLMs: LaViDa-LLaDA-v1.0-Instruct (Li et al., 2026), MMaDA-8B-MixCoT (Yang et al., 2026b), and LLaDA-V (You et al., 2026). Across these models, we evaluate VIG-Sampler against six sampling baselines: Confidence (Chang et al., 2022), Entropy (Ben-Hamu et al., 2026), Margin (Kim et al., 2025), MPD-PAC (Hong et al., 2026), PC-Sampler (Huang et al., 2025), and Info-Gain (Yang et al., 2026a). The first three are widely used uncertainty-based heuristics, whereas MPD-PAC is designed to enhance visual grounding in dMLLMs. PC-Sampler incorporates global trajectory and content-aware guidance, whereas Info-Gain accounts for the future uncertainty reduction induced by each decoding decision. Note that we use fixed hyperparameters of $\gamma = 1$ and $\lambda = 3$ for all models. Further implementation and benchmark details are provided in the Appendix B.

3.2 Main Results

Tables 1 and 2 summarize the experimental results on LaViDa and MMaDA, respectively. VIG-Sampler achieves the highest mean CIDEr score across all model–budget settings and obtains the best DetailCaps and VQA averages in five of the six settings, with the second-best result in the remaining setting. Its consistent gains across captioning and VQA suggest that visual reward-guided token scoring effectively prioritizes visually grounded and informative tokens. The advantage of VIG-Sampler becomes more pronounced as the decoding budget increases. At $k = 8$, it outperforms the second-best baseline by 12.7 and 14.5 in mean CIDEr score on LaViDa and MMaDA, respectively, with corresponding CAPTURE gains of 8.4 and 4.7. Remarkably, VIG-Sampler at $k = 8$ matches or outperforms certainty-based sampling at $k = 2$ on COCO Caption with LaViDa and Flickr30K with MMaDA, despite a $4\times$ reduction in decoding steps. Together, these results show that joint-information-aware set selection enables VIG-Sampler to gain more information at each step, benefiting from larger commit budgets while preserving generation quality.

Table 2: Experimental results of MMaDA with parallel decoding budgets $k \in \{2, 4, 8\}$. The first and last Avg. columns report mean CIDEr and VQA scores, respectively. The best scores are shown in **bold**, and the second-best scores are underlined.

k	Sampler	COCO Cap. *CIDEr*	Flickr30K *CIDEr*	NoCaps *CIDEr*	Avg. *CIDEr*	DetailCaps *CAPTURE*	TextVQA *Acc.*	DocVQA *ANLS*	ChartQA *Relaxed Acc.*	Avg. *Macro Avg.*
2	Confidence	87.4	57.7	86.5	77.2	50.1	12.2	9.3	11.8	11.1
	Entropy	82.5	55.8	83.3	73.9	**50.5**	12.0	8.8	11.8	10.9
	Margin	88.0	58.7	88.0	78.2	**50.5**	12.2	9.5	11.6	11.1
	MPD-PAC	86.1	56.7	81.2	74.7	50.0	12.4	10.1	11.6	11.4
	PC-Sampler	87.1	57.4	80.3	74.9	49.3	12.4	8.6	11.8	10.9
	Info-Gain	**89.1**	59.8	86.9	78.6	48.8	**12.8**	**10.6**	11.8	**11.7**
	VIG-Sampler	87.9	**60.2**	**89.2**	**79.1**	50.4	12.4	10.5	**12.0**	11.6
4	Confidence	83.4	56.8	78.2	72.8	45.5	11.1	7.9	12.0	10.3
	Entropy	78.8	54.7	75.9	69.8	44.5	11.1	7.6	12.0	10.2
	Margin	81.1	57.2	80.7	73.0	46.6	11.1	8.0	11.6	10.2
	MPD-PAC	80.9	52.4	77.3	70.2	46.5	11.4	8.3	12.2	10.6
	PC-Sampler	81.5	51.4	73.5	68.8	45.8	12.4	8.6	12.0	11.0
	Info-Gain	83.8	58.0	81.4	74.4	45.8	11.4	8.9	12.0	10.8
	VIG-Sampler	**87.6**	**58.8**	**86.6**	**77.7**	**49.1**	**12.7**	**9.9**	**13.2**	**11.9**
8	Confidence	61.6	41.7	61.6	55.0	33.1	11.0	8.4	12.2	10.5
	Entropy	60.7	41.0	60.1	53.9	31.8	11.0	8.3	12.2	10.5
	Margin	66.4	44.8	64.0	58.4	35.7	11.0	8.5	12.2	10.6
	MPD-PAC	61.0	41.2	60.2	54.1	38.6	11.1	8.6	12.2	10.6
	PC-Sampler	62.9	43.1	62.8	56.3	33.5	**12.6**	8.4	12.0	11.0
	Info-Gain	65.4	43.8	65.2	58.1	33.1	11.0	8.4	11.8	10.4
	VIG-Sampler	**81.9**	**58.3**	**78.4**	**72.9**	**43.3**	12.3	**9.5**	**12.4**	**11.4**

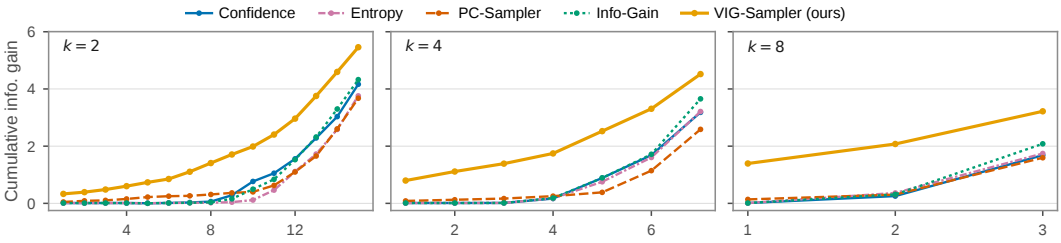


Figure 4: Cumulative information gain across decoding steps for $k \in \{2, 4, 8\}$

4 Analysis

4.1 Analysis of VIG-Sampler

Information gain induced by VIG-Sampler. Figure 4 compares the cumulative information gain of VIG-Sampler against competing sampling baselines across decoding steps on COCO Caption using LaViDa. To better isolate information gain related to visual content, we exclude punctuation, articles, and other function words, whose predictions are more strongly driven by linguistic context than by visual evidence. Although VIG-Sampler does not directly compute information gain, it consistently achieves the highest cumulative information gain across all commit budgets.

Qualitative case studies. Figure 5 shows that Confidence and Info-Gain commit tokens that convey overlapping information, such as *"penny"* and *"coin"*, without accounting for the joint information gain of the selected set. These early commitments bias subsequent predictions toward plausible but incorrect continuations, yielding phrases such as *"penny worth"* and semantically repetitive descriptions such as *"silver coin"* and *"silver penny"*. In contrast, VIG-Sampler selects complementary, visually grounded tokens and generates a caption consistent with the image.

Prompt: Provide a one-sentence caption for the provided image.



Confidence CIDEr 5.7	Info-Gain CIDEr 12.4	VIG-Sampler (ours) CIDEr 97.9
S11: A **penny** [M]…[M] **penny**.	**S12**: A [M][M] **is** [M][M] a [M] **penny**.	**S13**: A [M] of [M]…[M] **next** [M] a penny.
S12: A penny **worth** [M]…[M] a penny.	**S13**: A [M] **coin** is [M][M] **a** [M] penny.	**S14**: A **group** of **stones** [M][M] next to a penny.
Output: A penny worth of rocks that aren not worth a penny.	**Output**: A silver coin is next to a silver penny.	**Output**: A group of stones are sitting next to a penny.

Figure 5: Qualitative comparison on COCO Caption using LaViDa at $k = 2$. **S#** denotes the decoding step, and red and blue indicate incorrectly and correctly committed tokens, respectively.

4.2 Ablation Study

Hyperparameter sensitivity. Table 3 presents a sensitivity analysis of γ and λ on COCO Caption with LaViDa at $k = 2$. Performance generally improves as either γ or λ increases from zero, and their gains are largely additive when the two are combined. Performance remains robust across a broad range of values; although some configurations outperform the default, we fix $\gamma = 1$ and $\lambda = 3$ across all models and benchmarks for consistency.

Robustness to generation length. Table 4 compares VIG-Sampler and Info-Gain across different generation lengths at $k = 4$. VIG-Sampler consistently outperforms Info-Gain on both COCO Caption and TextVQA for all values of N, demonstrating robustness across generation lengths.

4.3 Analysis of Inference Cost

Table 5 summarizes the inference cost of each sampling method on COCO Caption using LaViDa, reporting peak GPU memory and average wall-clock time per sample. VIG-Sampler shows no measurable increase in peak memory over the baselines, while its wall-clock time remains close to confidence-based sampling across all values of k and substantially lower than Info-Gain. These results indicate that VIG-Sampler introduces only modest computational overhead despite incorporating visual guidance into token selection.

Table 3: Hyperparameter sensitivity on COCO Caption using LaViDa at $k = 2$. † denotes the default setting.

$\gamma\backslash\lambda$	0.0	1.0	2.0	3.0	4.0
0.0	100.2	100.9	102.1	102.9	101.7
0.5	101.9	105.2	106.9	106.5	105.2
1.0	101.5	103.8	105.4	106.2†	105.5
1.5	101.0	103.5	106.0	107.7	**107.8**
2.0	103.0	107.0	106.2	106.1	107.5

Table 4: Robustness to N using LaViDa at $k = 4$.

N	Sampler	COCO Cap.	TextVQA
16	Info-Gain	81.6	48.8
	VIG-Sampler	**88.4**	**55.0**
32	Info-Gain	94.8	50.2
	VIG-Sampler	**105.1**	**56.8**
48	Info-Gain	97.4	48.6
	VIG-Sampler	**102.5**	**57.4**
64	Info-Gain	96.3	49.5
	VIG-Sampler	**103.7**	**55.8**

Table 5: Inference cost of each sampler.

Sampler	Peak mem. (GB)	Wall-clock time (s)		
		k=2	k=4	k=8
Confidence	17.9	1.71	1.39	1.13
Info-Gain	17.9	3.25	2.21	1.52
VIG-Sampler	17.9	1.85	1.44	1.18

5 Conclusion

In this paper, we presented VIG-Sampler, a training-free parallel decoding strategy for dMLLMs that leverages token-to-image attention to guide token selection. By combining visual reward-guided token scoring with joint-information-aware set selection, VIG-Sampler prioritizes visually grounded and informative tokens during parallel decoding. Experiments across seven captioning and VQA benchmarks and three dMLLMs demonstrate consistent improvements over existing sampling baselines, with particularly strong gains under larger commit budgets. These results highlight the importance of incorporating visual information into the decoding order to preserve generation quality with fewer decoding steps.

AI USE STATEMENT

In this work, we used generative AI tools for translation and language editing. We have not used generative AI tools for experimental design or result generation. We have reviewed all AI-assisted work. We take responsibility for the final content of this work.

REFERENCES

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8948–8957, 2019.

Wenrui Bao, Zhiben Chen, Dan Xu, and Yuzhang Shang. Learning to parallel: Accelerating diffusion large language models via learnable parallel decoding. *arXiv preprint arXiv:2509.25188*, 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.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11315–11325, 2022.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

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*, 2025a.

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*, 2025b.

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.

Daehoon Gwak, Minseo Jung, Junwoo Park, Minho Park, ChaeHun Park, Junha Hyung, and Jaegul Choo. Reward-weighted sampling: Enhancing non-autoregressive characteristics in masked diffusion llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 34562–34582, 2025.

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 preprint arXiv:2508.13021*, 2025.

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

Metod Jazbec, Theo X Olausson, Louis Béthune, Pierre Ablin, Michael Kirchhof, João Monteiro, Victor Turrisi, Jason Ramapuram, and Marco Cuturi. Learning unmasking policies for diffusion language models. *arXiv preprint arXiv:2512.09106*, 2025.

Pu Jian, Junhong Wu, Wei Sun, Chen Wang, Shuo Ren, and Jiajun Zhang. Look again, think slowly: Enhancing visual reflection in vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9262–9281, 2025.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9339–9350, 2025.

Jaeyeon Kim, Kulin Shah, Vasilis Kontonis, Sham M Kakade, and Sitan Chen. Train for the worst, plan for the best: Understanding token ordering in masked diffusions. In *International Conference on Machine Learning*, pp. 30749–30768. PMLR, 2025.

Keuntae Kim, Mingyu Kang, and Yong Suk Choi. Thinking diffusion: Penalize and guide visual-grounded reasoning in diffusion multimodal language models. *arXiv preprint arXiv:2604.05497*, 2026a.

Seo Hyun Kim, Sunwoo Hong, Hojung Jung, Youngrok Park, and Se-Young Yun. Klass: Kl-guided fast inference in masked diffusion models. *Advances in Neural Information Processing Systems*, 38:92267–92301, 2026b.

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

Duo Li, Zuhao Yang, Xiaoqin Zhang, Ling Shao, and Shijian Lu. A comprehensive study on visual token redundancy for discrete diffusion-based multimodal large language models. *arXiv preprint arXiv:2511.15098*, 2025a.

Pengxiang Li, Yefan Zhou, Dilxat Muhtar, Lu Yin, Shilin Yan, Li Shen, Soroush Vosoughi, and Shiwei Liu. Diffusion language models know the answer before decoding. *arXiv preprint arXiv:2508.19982*, 2025b.

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*, 2025c.

Shufan Li, Jiuxiang Gu, Kangning Liu, Zhe Lin, Zijun Wei, Aditya Grover, and Jason Kuen. Sparse-lavida: Sparse multimodal discrete diffusion language models. *arXiv preprint arXiv:2512.14008*, 2025d.

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, 2026.

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.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024.

Richard Diehl Martinez, Zébulon Goriely, Andrew Caines, Paula Buttery, and Lisa Beinborn. Mitigating frequency bias and anisotropy in language model pre-training with syntactic smoothing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5999–6011, 2024.

Ahmed Masry, Jia Qing Tan, Shafiq Joty, Enamul Hoque, et al. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pp. 2263–2279, 2022.

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.

Tuan Dung Nguyen, Minh Khoi Ho, Qi Chen, Yutong Xie, Cam-Tu Nguyen, Minh Khoi Nguyen, Dang Huy Pham Nguyen, Anton van den Hengel, Johan Verjans, Phi Le Nguyen, et al. Beyond the global scores: Fine-grained token grounding as a robust detector of lvlm hallucinations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40235–40244, 2026.

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.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Qingyan Wei, Yaojie Zhang, Zhiyuan Liu, Puyu Zeng, Yuxuan Wang, Biqing Qi, Dongrui Liu, and Linfeng Zhang. Accelerating diffusion large language models with slowfast sampling: The three golden principles. *arXiv preprint arXiv:2506.10848*, 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. *arXiv preprint arXiv:2511.12428*, 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 ZHOU, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. Llada-v: Large language diffusion models with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10093–10105, 2026.

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.

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

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, 2025a.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19781–19791, 2025b.

A Related Work

A.1 Diffusion Multimodal Large Language Models

Multimodal large language models (MLLMs) project visual information into the representation space of a pretrained LLM, achieving strong performance on multimodal tasks such as captioning, visual question answering, and visual reasoning (Liu et al., 2023; 2024; Dai et al., 2023). With the emergence of dLLMs as an alternative to autoregressive LLMs (Nie et al., 2026; Gong et al., 2025; Ye et al., 2025), this paradigm has also been extended to multimodal models that denoise sequences conditioned on both visual and textual tokens (You et al., 2026; Li et al., 2026; Yu et al., 2025). More recent models jointly learn visual and textual tokens in a unified framework, rather than adapting visual features to language space, enabling balanced multimodal understanding and visual generation (Yang et al., 2026b; Li et al., 2025c; Shi et al., 2025).

Recent studies on dMLLM decoding have focused on improving inference efficiency by pruning redundant masked or visual tokens during generation (Li et al., 2025a;d; Xu et al., 2025). Another line of work improves the visual grounding of generated responses by adjusting token-level predictions through model-internal bias correction or classifier-free guidance (Hong et al., 2026; Kim et al., 2026a). Despite the growing body of research on dMLLMs, relatively little attention has been paid to which tokens should be decoded at each step and how these decoding decisions affect the subsequent generation trajectory.

A.2 Parallel Decoding Strategies in Diffusion Language Models

At each denoising step, diffusion language models predict all remaining masked positions and commit a subset of them for parallel decoding. The most common strategies select positions based on predictive certainty, measured by confidence (Nie et al., 2026; Li et al., 2026), entropy (Ye et al., 2025; Ben-Hamu et al., 2026), or the probability margin between the top two candidates (Kim et al., 2025; Li et al., 2025b). Based on these signals, several methods dynamically determine which and how many tokens to commit using threshold-based criteria (Wu et al., 2025; Yu et al., 2025; Ben-Hamu et al., 2026). Other training-free approaches leverage signals accumulated across denoising steps, such as changes in predictive distributions, convergence, and prediction stability (Kim et al., 2026b; Wei et al., 2025). Another class of approaches learns token selection policies from trace-level signals, such as future-stability labels derived from completed decoding traces or rewards based on final outputs (Bao et al., 2025; Jazbec et al., 2025). Recent studies have also explored alternative criteria, including information gain over the remaining masked positions, semantic priors, and search over candidate decoding trajectories (Yang et al., 2026a; Huang et al., 2025; Lee et al., 2025). Although these methods are generally effective in language-only settings, parallel decoding strategies for multimodal language models remain relatively underexplored, especially those that use visual grounding to guide token selection.

B Additional Details

B.1 Experimental Details for the Motivation Study

We conduct all experiments in the motivation study on a subset of 100 images from the COCO Captions validation set. For observation 1, we first generate captions using the confidence-based sampler with $k = 1$. For each generated caption, we mask one token at a time and run the model again with the image embeddings replaced by zero embeddings. If the prediction at the masked position changes, we classify the token as *visually grounded*; otherwise, we classify it as *non-grounded*. The image-attention mass and image-attention similarity used in the experiments are defined as m_t^i and $\langle \tilde{\mathbf{a}}_t^i, \tilde{\mathbf{a}}_t^j \rangle_{\cos}$, respectively.

B.2 Benchmarks

We evaluate VIG-Sampler on four image captioning benchmarks and three visual question answering benchmarks, covering a broad range of vision-language capabilities. For image captioning, we use COCO Caption (Lin et al., 2014), Flickr30K (Young et al., 2014), NoCaps (Agrawal et al., 2019),

and DetailCaps (Dong et al., 2024), which assess general image description, open-vocabulary generalization, and fine-grained, long-form visual description. For visual question answering, we use TextVQA (Singh et al., 2019), DocVQA (Mathew et al., 2020), and ChartQA (Masry et al., 2022), covering scene text understanding, document comprehension, and chart reasoning, respectively. We report CIDEr for COCO Caption, Flickr30K, and NoCaps, CAPTURE for DetailCaps, and accuracy, ANLS, and relaxed accuracy for TextVQA, DocVQA, and ChartQA, respectively. We additionally report the mean CIDEr score across the first three captioning benchmarks and the mean score across the three VQA benchmarks. For a consistent evaluation protocol, we use the `lmms-eval` package (Zhang et al., 2025a), evaluating DetailCaps on 500 samples and the remaining benchmarks on their lite subsets.

B.3 Implementation Details

Sampler configurations. We retain the default hyperparameter settings reported for prior methods whenever possible and make only minimal adjustments when differences in generation length or dataset characteristics lead to unstable decoding behavior. For MPD-PAC, we set the RoPE coefficient to $\beta = 0.01$, the threshold to $\tau_0 = 0.6$, $k = 3$, and the slope of the sigmoid schedule to 8. We use $\lambda_{\text{prior}} = 0.3$ for LaViDa, $\lambda_{\text{prior}} = 0.1$ for LLaDA-V, and $\lambda_{\text{prior}} = 0.03$ for MMaDA. For PC-Sampler, we set the positional decay coefficient to $\lambda = 0.025$ and set the clipping threshold to $\alpha = 10$. For Info-Gain, we use a position temperature of $\tau_{\text{pos}} = 0.1$, $N = 4$ candidates, and an acceleration threshold of 0.8. For VIG-Sampler, we set $\gamma = 1$ and $\lambda = 3$.

Generation settings. For LaViDa and MMaDA, we use a generation length of $N = 32$ on all benchmarks except DetailCaps. On DetailCaps, we employ block decoding for both models with a generation length of $N = 128$ and a block length of 16. For LLaDA-V, we use $N = 16$ on COCO Caption, Flickr30K, and NoCaps, $N = 128$ on DetailCaps, and $N = 32$ on the VQA benchmarks.

Evaluation setup. All experiments are conducted on NVIDIA RTX A5000 GPUs. We set the sampling temperature to zero and report single-run results.

B.4 Derivation of the Greedy Selection Rule

We derive the greedy rule used to approximately optimize the set-level objective in Equation 10. For the current set $\mathcal{S}$ of n selected positions, i.e., $|\mathcal{S}| = n$, define

$$F_t(\mathcal{S}) = \sum_{i \in \mathcal{S}} r_t^i - \frac{\lambda}{|\mathcal{S}| - 1} \sum_{\substack{i,j \in \mathcal{S} \\ i<j}} G_t^{i,j}, \tag{12}$$

where the pairwise similarity term is defined as zero when $|\mathcal{S}| = 1$, and we set $F_t(\emptyset) = 0$. Consider adding a candidate $i \in \mathcal{M}_t^{(\mathcal{S})}$. Define the accumulated pairwise similarity within the current set as

$$P_t(\mathcal{S}) = \sum_{\substack{a,b \in \mathcal{S} \\ a<b}} G_t^{a,b}. \tag{13}$$

After adding candidate i, the pairwise similarity term additionally contains its similarities to all positions already in $\mathcal{S}$. Thus, for $n \geq 2$,

$$F_t(\mathcal{S} \cup \{i\}) = \sum_{j \in \mathcal{S}} r_t^j + r_t^i - \frac{\lambda}{n} \left[P_t(\mathcal{S}) + \sum_{j \in \mathcal{S}} G_t^{i,j} \right]. \tag{14}$$

Subtracting the current objective

$$F_t(\mathcal{S}) = \sum_{j \in \mathcal{S}} r_t^j - \frac{\lambda}{n-1} P_t(\mathcal{S}) \tag{15}$$

gives the marginal gain

$$F_t(\mathcal{S} \cup \{i\}) - F_t(\mathcal{S}) = r_t^i - \frac{\lambda}{n} \sum_{j \in \mathcal{S}} G_t^{i,j} + \frac{\lambda}{n(n-1)} P_t(\mathcal{S}). \tag{16}$$

Table 6: Experimental results of LLaDA-V with parallel decoding budgets $k \in \{2, 4, 8\}$. The first and last Avg. columns report mean CIDEr and VQA scores, respectively. The best scores are shown in **bold**, and the second-best scores are underlined.

k	Sampler	COCO Cap.	Flickr30K	NoCaps	Avg.	DetailCaps	TextVQA	DocVQA	ChartQA	Avg.
		CIDEr	*CIDEr*	*CIDEr*	*CIDEr*	*CAPTURE*	*Acc.*	*ANLS*	*Relaxed Acc.*	*Macro Avg.*
2	Confidence	92.8	83.9	91.6	89.4	59.9	67.3	81.1	76.2	74.9
	Entropy	91.0	77.8	89.4	86.1	59.0	63.9	82.0	76.6	74.2
	Margin	**94.5**	82.5	91.8	89.6	**61.1**	67.3	82.0	**77.6**	75.6
	MPD-PAC	92.4	83.6	91.2	89.1	60.5	65.0	81.4	75.8	74.1
	PC-Sampler	92.1	81.9	90.8	88.3	59.8	66.5	82.8	76.4	75.2
	Info-Gain	94.4	**86.1**	**93.8**	**91.4**	60.1	67.2	82.0	76.6	75.3
	VIG-Sampler	93.9	82.5	91.3	89.2	59.9	**67.8**	**83.4**	77.0	**76.1**
4	Confidence	89.4	77.9	84.9	84.1	58.4	63.2	80.2	75.6	73.0
	Entropy	85.7	73.3	83.8	80.9	56.3	62.2	78.6	73.4	71.4
	Margin	88.9	76.6	85.6	83.7	59.9	**63.8**	81.4	**76.4**	73.9
	MPD-PAC	89.0	76.0	84.2	83.1	58.8	62.5	79.5	74.8	72.3
	PC-Sampler	85.6	74.2	81.9	80.6	53.9	63.6	**82.6**	76.0	**74.1**
	Info-Gain	**90.8**	78.1	88.0	85.6	58.6	62.1	80.7	75.8	72.9
	VIG-Sampler	90.7	**78.5**	**90.4**	**86.5**	**60.3**	**63.8**	79.6	**76.4**	73.3
8	Confidence	72.3	61.4	66.8	66.8	49.0	51.2	73.3	72.0	65.5
	Entropy	66.6	59.5	65.0	63.7	36.8	43.7	69.0	69.0	60.6
	Margin	**74.3**	60.6	69.5	68.1	49.1	53.9	74.4	**73.0**	67.1
	MPD-PAC	72.3	61.1	66.2	66.5	47.3	48.1	71.0	71.8	63.6
	PC-Sampler	67.4	57.3	61.1	61.9	30.2	50.6	**75.7**	72.4	66.2
	Info-Gain	74.2	62.7	68.3	68.4	49.0	51.1	73.3	72.2	65.5
	VIG-Sampler	**74.3**	**64.0**	**71.8**	**70.0**	**59.5**	**56.1**	75.6	72.8	**68.2**

The last term depends only on the currently selected set $\mathcal{S}$ and is independent of candidate i. Therefore, it does not affect which candidate maximizes the marginal gain.

For $n = 1$, the current set has no pairwise similarity term, and adding i yields

$$F_t(\mathcal{S} \cup \{i\}) - F_t(\mathcal{S}) = r_t^i - \lambda \sum_{j \in \mathcal{S}} G_t^{i,j}, \tag{17}$$

which has the same candidate-dependent form as Equation 16 with $n = 1$.

Consequently, for any nonempty $\mathcal{S}$, the greedy selection rule is

$$i^\star = \arg \max_{i \in \mathcal{M}_t^{(\mathcal{S})}} \left[r_t^i - \frac{\lambda}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} G_t^{i,j} \right]. \tag{18}$$

When $\mathcal{S} = \emptyset$, the similarity term is defined as zero, so the first selected position is

$$i^\star = \arg \max_{i \in \mathcal{M}_t} r_t^i. \tag{19}$$

B.5 Overall Decoding Procedure of VIG-Sampler

For completeness, we summarize the overall decoding procedure of VIG-Sampler in Algorithm 1.

Algorithm 1 Visual Information-Guided Sampler

Require: Visual input $\mathbf{I}$, text prompt $\mathbf{x}$, response length N, commit budget k, and hyperparameters γ, λ
Ensure: Decoded response $\mathbf{y}_T$

$\mathbf{y}_0 \leftarrow [\mathrm{MASK}]^N$
$T \leftarrow N/k$
for $t = 0, \ldots, T-1$ **do**
  $\mathcal{M}_t \leftarrow \{i \mid y_t^i = [\mathrm{MASK}]\}$
  Run one forward pass conditioned on $(\mathbf{I}, \mathbf{x}, \mathbf{y}_t)$ to obtain $p_\theta^i(\cdot \mid \mathbf{y}_t)$ and A
  for all $i \in \mathcal{M}_t$ **do**
    $\hat{y}_t^i \leftarrow \arg\max_{v \in \mathcal{V}} p_\theta^i(v \mid \mathbf{y}_t)$
    $c_t^i \leftarrow p_\theta^i(\hat{y}_t^i \mid \mathbf{y}_t)$
    $\mathbf{a}_t^i \leftarrow A_{i,\mathcal{I}}$
    $m_t^i \leftarrow \sum_{j \in \mathcal{I}} A_{i,j}$
  end for
  $m_t^{\mathrm{med}} \leftarrow \mathrm{median}_{i \in \mathcal{M}_t} m_t^i$
  $\bar{\mathbf{a}}_t \leftarrow \dfrac{1}{|\mathcal{M}_t|} \sum_{j \in \mathcal{M}_t} \mathbf{a}_t^j$
  for all $i \in \mathcal{M}_t$ **do**
    $r_t^i \leftarrow c_t^i \left(\dfrac{m_t^i}{m_t^{\mathrm{med}}} \right)^\gamma$
    $\widetilde{\mathbf{a}}_t^i \leftarrow \mathbf{a}_t^i - \bar{\mathbf{a}}_t$
  end for
  for all $i, j \in \mathcal{M}_t,\ i \neq j$ **do**
    $G_t^{i,j} \leftarrow \max\left\{ \left\langle \tilde{\mathbf{a}}_t^i, \tilde{\mathbf{a}}_t^j \right\rangle_{\cos}, 0 \right\}$
  end for
  $\mathcal{S} \leftarrow \emptyset$
  while $|\mathcal{S}| < k$ **do**
    for all $i \in \mathcal{M}_t^{(\mathcal{S})}$ **do**
      if $\mathcal{S} = \emptyset$ **then**
        $q_t^i \leftarrow r_t^i$
      else
        $q_t^i \leftarrow r_t^i - \dfrac{\lambda}{|\mathcal{S}|} \sum_{j \in \mathcal{S}} G_t^{i,j}$
      end if
    end for
    $i^\star \leftarrow \arg\max_{i \in \mathcal{M}_t^{(\mathcal{S})}} q_t^i$
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{i^\star\}$
  end while
  $\mathbf{y}_{t+1} \leftarrow \mathbf{y}_t$
  for all $i \in \mathcal{S}$ **do**
    $y_{t+1}^i \leftarrow \hat{y}_t^i$
  end for
end for
return $\mathbf{y}_T$

C Additional Experiments and Analysis

C.1 Generalization to a Different Model Family

In Table 6, we present additional experiments on LLaDA-V (You et al., 2026) to evaluate the generalizability of VIG-Sampler to a different model family.

Table 7: Robustness to N using LaViDa at $k = 2$.

N	Sampler	COCO Cap.	TextVQA
16	Info-Gain	87.2	**58.0**
	VIG-Sampler	**92.4**	57.5
32	Info-Gain	106.0	57.4
	VIG-Sampler	**106.2**	**58.7**
48	Info-Gain	**106.8**	56.5
	VIG-Sampler	104.2	**59.4**
64	Info-Gain	104.6	57.3
	VIG-Sampler	**105.9**	**59.0**

Table 8: Hyperparameter sensitivity on COCO Caption using LaViDa at $k = 4$ (left) and $k = 8$ (right). † denotes the default setting.

$\gamma\backslash\lambda$	0.0	1.0	2.0	3.0	4.0
0.0	93.2	98.0	101.1	**105.2**	103.6
0.5	96.4	100.6	101.1	101.8	104.5
1.0	96.3	99.5	102.7	105.1†	101.2
1.5	95.6	101.1	101.3	102.7	101.2
2.0	95.0	102.0	104.7	**105.2**	104.3

$\gamma\backslash\lambda$	0.0	1.0	2.0	3.0	4.0
0.0	74.8	85.4	92.7	96.2	99.9
0.5	76.7	86.3	92.8	95.9	98.5
1.0	80.2	87.4	94.7	**100.1**†	97.9
1.5	82.3	87.8	95.1	99.0	98.8
2.0	83.2	88.5	98.1	98.2	99.2

C.2 Additional Ablation Study on Generation Length

As shown in Table 7, we present additional comparisons between VIG-Sampler and Info-Gain across different generation lengths with $k = 2$. Consistent with the results at $k = 4$, VIG-Sampler maintains strong overall performance across different generation lengths, demonstrating its robustness to the choice of N.

C.3 Additional Hyperparameter Sensitivity Analysis

Additional hyperparameter sensitivity results across different values of k and models are reported in Tables 8 and 9. Performance is robust across a wide range of hyperparameter values, and we use $\gamma = 1$ and $\lambda = 3$ for all models and benchmarks for consistency despite some gains from other configurations.

C.4 Additional Qualitative Case Studies and Analysis of VIG-Sampler

In Figure 6, we provide additional qualitative comparisons that highlight the effectiveness of VIG-Sampler.

We further report the image-attention patterns of the selected tokens in Figure 7. In this example, the confidence-based method selects uninformative tokens with diffuse attention. Unlike the confidence-based method, VIG-Sampler selects tokens that are critical to generating the final answer and grounds them in distinct image regions. Additionally, we investigate the effect of a penalty based on image-attention similarity on token selection during decoding. Under the $\lambda = 0$ ablation setting (i.e., with the penalty term disabled), the selected tokens attend to nearly identical image regions and provide little additional information beyond one another. By contrast, in our default setting with the penalty term, the selected tokens attend to distinct image regions and faithfully capture different visual content.

Table 9: Hyperparameter sensitivity on COCO Caption using MMaDA at $k = 4$ (left) and $k = 8$ (right). † denotes the default setting.

$\gamma\backslash\lambda$	0.0	1.0	2.0	3.0	4.0
0.0	83.4	84.8	84.4	86.0	86.1
0.5	79.9	83.4	85.7	86.5	87.4
1.0	80.5	82.7	86.5	87.6†	**88.7**
1.5	79.9	83.5	87.3	88.3	87.0
2.0	80.9	86.0	88.6	87.1	88.2

$\gamma\backslash\lambda$	0.0	1.0	2.0	3.0	4.0
0.0	61.7	69.6	73.8	82.1	79.9
0.5	61.9	71.7	79.3	80.8	81.6
1.0	64.3	75.7	80.3	81.9†	82.0
1.5	64.9	75.0	80.6	82.9	80.4
2.0	66.1	74.3	81.8	**85.8**	82.2

Prompt: Provide a one-sentence caption for the provided image.



Confidence CIDEr 44	Info-Gain CIDEr 67	VIG-Sampler (ours) CIDEr 75
S13: A woman riding a skateboard **with a** [M]…[M].	**S14**: A woman riding a skateboard **down** a [M][M] **the** street.	**S14**: A woman riding a skateboard **on** a [M][M].
S14: A woman riding a skateboard with a [M][M] **on it**.	**S15**: A woman riding a skateboard down a **sidewalk on** the street.	**S15**: A woman riding a skateboard on a **cement street**.
Output: A woman riding a skateboard with a **blue design on it**.	**Output**: A woman riding a skateboard down a **sidewalk** on the street.	**Output**: A woman riding a skateboard on a **cement street**.

Prompt: Provide a one-sentence caption for the provided image.



Confidence CIDEr 148	Info-Gain CIDEr 91	VIG-Sampler (ours) CIDEr 256
S12: Two white trucks [M]…[M].	**S13**: Two white trucks **are** [M][M] **side** [M][M].	**S11**: Two white trucks are **parked** [M]…[M]
S13: Two white trucks **are** [M] **side** [M]…[M].	**S14**: Two white trucks are **side** [M] **side** [M] **side**.	**S12**: Two white trucks are parked **next to** [M]…[M]
Output: Two white trucks are parked **side** by **side** out**side**.	**Output**: Two white trucks are **side** by **side** by **side**.	**Output**: Two white trucks are parked **next to each other.**

Prompt: Provide a one-sentence caption for the provided image.



Confidence CIDEr 247	Info-Gain CIDEr 191	VIG-Sampler (ours) CIDEr 178
S07: A [M]…[M] **[EOT]**	**S07**: A [M]…[M] **[EOT]**	**S07**: A young man [M] **a** [M]…[M]
S08: A [M]…[M] [EOT]**[EOT]**	**S08**: A [M]…[M] [EOT]**[EOT]**	**S08**: A young man [M] a **red** [M]…[M]
Output: A young man in the snow standing on a snowboard.	**Output**: A young man in a hoodie standing on a snowboard.	**Output:** A young man in a **red** hoodie standing on a snow board.

Prompt: Provide a one-sentence caption for the provided image.



Confidence CIDEr 391	Info-Gain CIDEr 391	VIG-Sampler (ours) CIDEr 230
S07: A [M]…[M] **[EOT]**	**S07**: A [M]…[M] **[EOT]**	**S07**: A [M] eg [M]…[M] **[EOT]**
S08: A [M]…[M] [EOT]**[EOT]**	**S08**: A [M]…[M] [EOT]**[EOT]**	**S08**: A [M] **egret** [M]…[M]
Output: A white bird is standing on the roof of a car.	**Output**: A white bird is standing on the roof of a car.	**Output:** A white **egret** standing on the roof of a **parked** car.

Figure 6: Qualitative examples on COCO Caption using LaViDa at $k = 2$. The top two rows illustrate decoding failures by conventional methods without explicit visual guidance, in which incorrect tokens are produced during generation. The bottom two rows present detailed captions that are visually well aligned with the images from a human perspective, despite receiving relatively low CIDEr scores. **S#** denotes the decoding step, while red and blue indicate incorrectly and correctly committed tokens, respectively.



Figure 7: Qualitative analysis of token selection and image-attention patterns. Confidence-based selection tends to choose uninformative tokens with diffuse attention, whereas VIG-Sampler selects answer-relevant tokens grounded in distinct image regions. Without the penalty term, VIG-Sampler selects tokens with similar image-attention patterns, providing little additional information beyond one another.