Thinking Diffusion: Penalize and Guide Visual-Grounded Reasoning in Diffusion Multimodal Language Models

Keuntae Kim[1*] Mingyu Kang[2*] Yong Suk Choi[1,2†]
[1]Department of Computer Science, Hanyang University
[2]Department of Artificial Intelligence, Hanyang University
ktkpv94@hanyang.ac.kr, alsrb15788@hanyang.ac.kr, cys@hanyang.ac.kr

Abstract

Diffusion large language models (dLLMs) are emerging as promising alternatives to autoregressive (AR) LLMs. Recently, this paradigm has been extended to multimodal tasks, leading to the development of diffusion multimodal large language models (dMLLMs). These models are expected to retain the reasoning capabilities of LLMs while enabling faster inference through parallel generation. However, when combined with Chain-of-Thought (CoT) reasoning, dMLLMs exhibit two critical issues. First, we observe that dMLLMs often generate the final answer token at a very early timestep. This trend indicates that the model determines the answer before sufficient reasoning, leading to degraded reasoning performance. Second, during the initial timesteps, dMLLMs show minimal dependency on visual prompts, exhibiting a fundamentally different pattern of visual information utilization compared to AR vision-language models. In summary, these findings indicate that dMLLMs tend to generate premature final answers without sufficiently grounding on visual inputs. To address these limitations, we propose Position & Step Penalty (PSP) and Visual Reasoning Guidance (VRG). PSP penalizes tokens in later positions during early timesteps, delaying premature answer generation and encouraging progressive reasoning across timesteps. VRG, inspired by the classifier-free guidance, amplifies visual grounding signals to enhance the model's alignment with visual evidence. Extensive experiments across various dMLLMs demonstrate that our method achieves up to 7.5% higher accuracy while delivering more than 3× speedup compared to reasoning with four times more diffusion steps.

1. Introduction

Vision-language models (VLMs) have demonstrated remarkable capabilities in understanding visual inputs and generating contextually relevant textual descriptions [1, 3, 14, 15, 34, 44]. These models have achieved strong performance across a wide range of vision-language tasks, including visual question answering and visual reasoning. Furthermore, with the incorporation of Chain-of-Thought (CoT) prompting, VLM can perform step-by-step reasoning before producing the final answer [24, 39, 48].

Recently, diffusion large language models (dLLMs) [25, 41] have emerged as a promising alternative to autoregressive (AR) LLMs. In contrast to AR models that generate a single token at a time in a left-to-right manner, dLLMs restore multiple tokens in parallel, providing substantially faster inference. Extending this advantage to the multimodal model, diffusion multimodal large language models (dMLLMs) [16, 40, 42, 43] have been introduced, presenting a new paradigm that enables joint reasoning over textual and visual information.

However, the reasoning process of dMLLMs remains insufficiently understood. Since dMLLMs rely on a diffusion-based generation mechanism that reconstructs tokens in parallel, reasoning enhancement methods designed for AR VLMs cannot be directly applied. Moreover, while prior research has provided quantitative analyses on how AR VLMs utilize visual evidence [6, 12, 13], such analyses are largely absent for dMLLMs, leaving it unclear whether they effectively leverage visual information. This gap highlights the need to closely investigate how dMLLMs perform reasoning and to design effective methods to strengthen their reasoning ability.

In this work, we present the first quantitative analysis of the reasoning process of dMLLMs with CoT prompting. Our analysis presents two critical issues. First, dMLLMs exhibit Early Answer Generation, where the model generates final answer tokens prematurely at very early timesteps, particularly when the number of diffusion steps is small.

*Equal contribution.
†Corresponding author.

Query: Are the two animals in the picture the same color?
(A) Same
(B) Not the same
(C) Can't judge

State at Answer Generation:

[Base]

To determine if the _____ let's analyze _____
_____:
__ ____ ____ to have _____.
_____ _____ that _____.
_, the _ answer is A.

[PSP & VRG]

To determine if the two animals are the same color, let's analyze their colors:
1. The left rhino appears darker.
2. The right rhino _____ warmer.
It is reasonable _____ they are not the same color.
____ __ correct answer is B.

Final Response:

[Base]

To determine if the two animals in the picture are the same color, let's analyze their fur colors:
1. Both animals appear to have a similar light or gray fur color.
Given this analysis, we can conclude that the two animals are indeed the same color.
Therefore, the correct answer is A.

[PSP & VRG]

To determine if the two animals are the same color, let's analyze their colors:
1. The left rhino appears darker.
2. The right rhino appears warmer.
It is reasonable to conclude that they are not the same color.
Therefore, the correct answer is B.

Figure 1. Example responses of LaViDa (generation length = 64, diffusion steps = 32). **State at answer generation** refers to the output state at the moment when the final answer (A, B, C) is generated, while **Final Response** denotes the model's final response generated by each method.

For example, as shown in Figure 1, the model appears to determine the final answer before completing sufficient reasoning steps and then generates intermediate reasoning to justify the final answer. Second, we observe low dependency on visual prompts during the initial timesteps. Unlike AR VLMs, dMLLMs exhibit a fundamentally different pattern, relying little on visual information early on and incorporating visual cues only in later timesteps. Overall, these findings suggest that dMLLMs tend to generate answers too early without effectively leveraging visual evidence.

To address these issues, we propose two novel training-free methods that can be applied directly during inference: Position & Step Penalty (PSP) and Visual Reasoning Guidance (VRG). First, PSP discourages the premature generation of answer-position tokens at early timesteps by applying penalties, thereby encouraging a more progressive reasoning process. Second, VRG, inspired by classifier-free guidance, amplifies the conditional logits associated with visual prompts to strengthen the model's use of visual evidence.

We validate the effectiveness of our method through extensive experiments. Experimental results across various dMLLMs demonstrate that our method achieves up to 7.5% higher accuracy and over 3× faster inference compared to reasoning using four times more diffusion steps. Our contributions can be summarized as follows:

- We conduct the first quantitative analysis of the reasoning process in diffusion-based multimodal language models with CoT prompting, identifying Early Answer Generation and insufficient early visual grounding.
- We propose Position & Step Penalty (PSP) to promote timestep-wise reasoning progression and Visual Reasoning Guidance (VRG) to enhance visual grounding during inference.
- We introduce a training-free method that generalizes across different dMLLMs and achieves strong performance on multimodal benchmarks, outperforming models that use four times more diffusion steps with up to 7.5% higher accuracy and more than 3× faster inference.

2. Background & Related Work

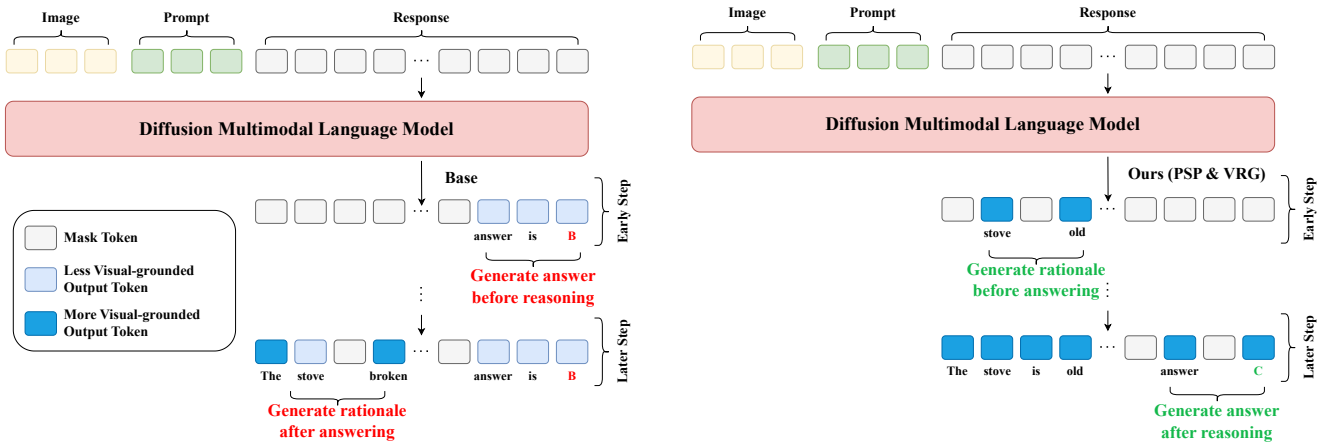
2.1. Diffusion Language Models

Diffusion models have demonstrated notable success in image generation [27, 28]. Motivated by their success, several studies have adapted diffusion methods to text generation by applying them to continuous text embeddings [5, 8, 17]. To handle the discrete nature of text, discrete diffusion models were introduced, operating directly on the discrete vocabulary space [20, 29]. Recently, dLLMs such as LLaDA [25] and Dream [41] achieved performance comparable to autoregressive LLMs at scale. These models allow flexible control over the speed–quality tradeoff by adjusting the number of diffusion steps.

Formally, given a discrete token sequence of length L, $X_0 = [X_0^1, X_0^2, \ldots, X_0^L]$, the forward process $q(X_t|X_s)$ progressively masks tokens over the time interval $[0, 1]$, where $1 \geq t \geq s \geq 0$. When $t = 1$, the sequence X_1 consists entirely of mask tokens $[M]$. The model p_θ parameterizes the reverse process $p(X_s|X_t)$, learning to reconstruct the original sequence from the fully masked state.

During inference, the model begins with a fully masked sequence $X_1 = [M, M, \ldots, M]$ and iteratively applies the learned reverse process $p_\theta(X_0|X_t)$ to gradually restore the original tokens. Sampling starts by setting the target generation length and initializing the response sequence X_1 entirely with $[M]$. The model then progressively transitions from state X_t to X_s (where $s < t$), decreasing the masking level and incrementally recovering the sequence.

Specifically, given the current state X_t, the model p_θ predicts all masked tokens $[M]$ based on the provided condi-



(a) Baseline dMLLM Inference.

(b) Our proposed dMLLM inference.

Figure 2. Overview of our method. (a) shows the base inference process, which exhibits early answer generation and less visual grounding. (b) illustrates our proposed inference process that introduces PSP and VRG to mitigate these two issues.

tions (e.g., multimodal input v, prompt c, and the partially masked sequence X_t). After prediction, a portion of tokens corresponding to the ratio s/t is remasked, while the remaining $(1 - s/t)$ tokens are retained as is.

2.2. Multimodal Chain-of-Thought

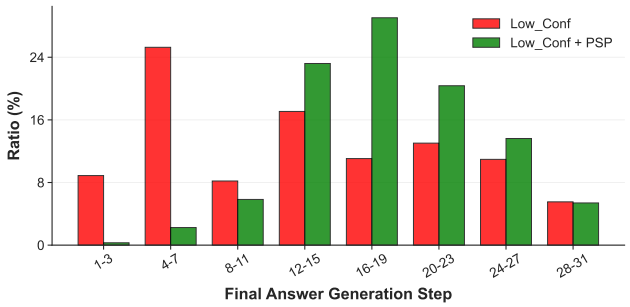
Chain-of-Thought (CoT) [35, 37, 46], which has greatly contributed to improving the reasoning ability of LLMs, has also been actively explored in MLLMs [21, 31, 36]. However, unlike CoT in LLMs, Multimodal CoT faces the challenge of handling inputs from different modalities, and various approaches have been proposed to address this issue [47]. The most common strategy is to leverage the strong capabilities of LLMs by generating image captions and feeding them, together with textual prompts, as inputs to the LLM. However, this approach heavily depends on the quality of the image captions and suffers from significant visual information loss during the image-to-text conversion process.

To overcome these limitations, there have been attempts to apply multimodal CoT directly to VLMs. Nevertheless, such methods still rely heavily on text-based rationales and struggle to capture fine-grained associations between visual and textual information [39, 45, 48]. Recent studies have thus focused on enriching visual descriptions or injecting structural information. For instance, CCoT [24] generates a scene graph to structurally encode object, relation, and spatial information into the prompt, thereby systematizing linguistic reasoning cues. DDCoT [49] decomposes complex inputs into reasoning and recognition and guides the model through a step-by-step reasoning process. ICoT [7] attempts to insert fine visual cues, which are difficult to express in text, directly into intermediate reasoning steps by interleaving textual rationales with image patches. Although these methods have proven effective in autoregressive (AR) VLMs, they fundamentally depend on stepwise reasoning through sequential token generation and therefore fail to demonstrate the same effectiveness in dMLLMs.

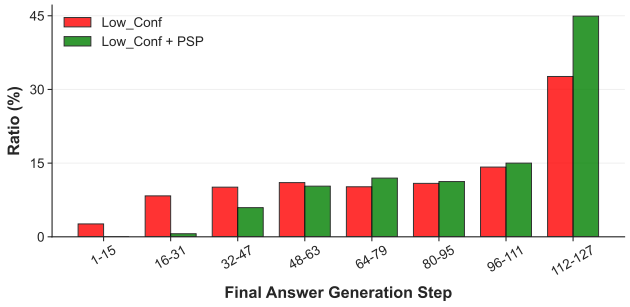
2.3. CoT in Diffusion Language Models

In AR based language models, CoT method has achieved remarkable success by generating intermediate reasoning steps [35, 37, 46]. AR CoT leverages the left-to-right token generation process, conditioning each new token on previously generated ones. This approach maximizes reasoning capability by allowing the model to generate and utilize its own rationales. However, it also has critical drawbacks. As the length of the intermediate reasoning increases, the inference time and computational cost of AR CoT grow linearly. Moreover, due to the irreversible generation structure, once a token is generated, it cannot be modified. Consequently, any error that occurs during generation continues to propagate through subsequent timesteps. To correct such errors, the framework must force the model to regenerate the sequence from the beginning, which requires additional inference calls and results in several times higher computational cost [11, 23, 26, 32].

In contrast, the reasoning process of dLLMs consists of iterative steps, where tokens are unmasked and the remaining tokens are remasked for progressive refinement at each timestep [30, 38]. Thus, the core of diffusion CoT lies in the remasking strategy [10, 33]. While intuitive strategies such as low-confidence, entropy, and margin-based methods have been explored, research on optimal remasking strategies for reasoning remains in its early stages. Although diffusion CoT has not yet reached the reasoning performance level of AR CoT, it offers two key advantages that AR CoT cannot provide: (1) faster reasoning through parallel gener-



(a) Generation length L = 64 / diffusion step T =32



(b) Generation length L = 256 / diffusion step T = 128

Figure 3. Results of the final answer generation step on the M3CoT validation set using LaViDa. The default remasking strategy is Low-confidence.

ation, (2) flexible control over the speed–quality tradeoff. In this work, we extend dMLLMs with CoT reasoning while effectively balancing the speed–quality tradeoff inherent in diffusion-based generation.

3. Analysis of Reasoning in dMLLMs

3.1. Experimental Setup for Case Study

To analyze the reasoning of dMLLMs with CoT prompting, we conduct a case study using M3CoT Validation Set [2]. M3CoT covers diverse domains such as science, mathematics, and commonsense, requiring multi-step reasoning based on multimodal inputs. We quantitatively analyze the inference process of dMLLMs under varying diffusion steps and response generation length. Our primary analysis focuses on LaViDa, while additional case studies on other dMLLMs are provided in the supplementary material.

3.2. Observation 1: Early Answer Generation

We quantitatively analyze when dMLLMs generate an answer under different diffusion steps T and generation length L. Specifically, for each sample, we record the timestep at which the final answer token first appears and visualize the overall distribution in Figure 3.

Figure 3a shows the distribution under L = 64 and T = 32. In this setting, the model tends to generate the final answer at very early steps. Specifically, it exhibits a strong Early Answer Generation tendency, producing the final answer before the 7th step in over 30% of cases, as shown in Figure 3a. This suggests that it often generates the answer before sufficient reasoning has been completed, and only afterward generates rationales based on that early answer. In contrast, as shown in the Figure 3b, with L = 256 and T = 128, most samples generate answers at later steps. This indicates that when both the diffusion steps and generation length are sufficiently large, the model tends to complete a more coherent reasoning process before generating the final answer.

Overall, these results suggest that while stable reasoning emerges with sufficiently large L and T, smaller configurations lead to a stronger Early Answer Generation, preventing the model from performing sufficient reasoning. Therefore, to ensure reliable reasoning performance even under limited L and T, a new remasking strategy is required to guide the model's reasoning progression more effectively.

3.3. Observation 2: Week Visual Grounding

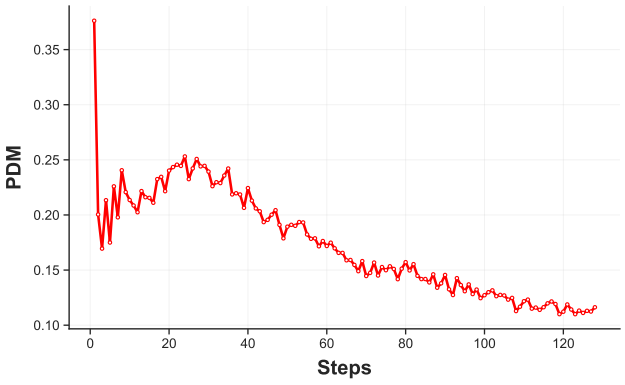
We further evaluate how visual information contributes to token generation. Following prior work [6], we quantify the dependency of each unmasked token on visual prompts using the visual prompt dependency measure (PDM):

$$\mathrm{PDM}(X_s) = \frac{1}{\sqrt{2}} \sqrt{\sum \left(\sqrt{p_\theta(X_s \mid X_t, c, v)} - \sqrt{p_\theta(X_s \mid X_t, c)} \right)^2} \quad (1)$$

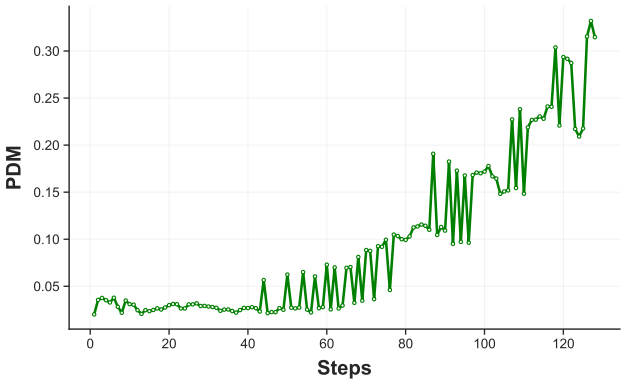
where v, c, and X_t denote the visual input, the prompt, and the partially masked sequence, respectively.

As shown in Figure 4, LLaVA-1.5 [18] (an AR-based VLM) presents high PDM at the early stages of generation, which gradually decline as reasoning progresses. This trend suggests that the model heavily relies on visual features at the beginning but reduces its dependence on them in later reasoning steps [6, 13]. In contrast, dMLLM shows a markedly different pattern. The overall PDM are lower, particularly at the early diffusion steps, where the model shows minimal sensitivity to visual inputs. As the diffusion process progresses, PDM gradually increases, indicating that the model begins to incorporate visual information only at later stages of generation.

This result suggests that, unlike AR VLMs that actively incorporate visual information from the earliest reasoning steps, dMLLMs generate their final answer already at very early timesteps, before visual prompt dependence has been sufficiently strengthened. Combining Observations 1 and 2, we find that dMLLMs often generate an answer without sufficiently referencing the visual input and then construct the following reasoning trajectory around that answer, which is generated before effective visual grounding.



(a) LLaVA-1.5



(b) LaVida with generation length $L = 256$ / step $T = 128$

Figure 4. Comparison of PDM measurements on the M3CoT validation set between the autoregressive-based model LLaVA-1.5 and the diffusion-based model LaViDa.

4. Method

4.1. Position & Step Penalty

In Section 3.2, we observe that dMLLMs tend to generate answers prematurely under short diffusion steps and limited response generation length. In other words, the model often generates the final answer tokens before completing sufficient reasoning steps, and subsequently generates rationales conditioned on this early answer.

To mitigate this issue, we introduce Position & Step Penalty (PSP), designed to delay answer generation to later timesteps and encourage gradual reasoning progression. As shown in Figure 2b, the model starts from an initial sequence X_1 consisting of L masked tokens and progressively unmasks them through K discrete timesteps $\{t_1, t_2, \ldots, t_K\}$, where $t_1 = 1$ and $t_K = 0$. At each step, the model samples:

$$X_0 \sim p_\theta(X_0 \mid X_{t_i}) \tag{2}$$

and then remasks the top $L \times t_{i+1}$ tokens to obtain $X_{t_{i+1}}$.

For each jth token at timestep t_i, with its confidence score C_j^i, we apply PSP as follows:

$$\text{rel}(j) \in [0, 1], \qquad \tau_i = \frac{i}{K} \in (0, 1] \tag{3}$$

$$\tilde{C}_j^i = C_j^i \cdot [1 - \gamma(1 - \tau_i)\,\text{rel}(j)] \tag{4}$$

where γ is a penalty strength coefficient, τ_i represents the relative progress of the diffusion process, and $\text{rel}(j)$ denotes the normalized positional index of token j within the response sequence (ranging from 0 to 1). Consequently, tokens positioned toward the end of the sequence receive a much stronger penalty in the early timesteps, preventing the model from prematurely generating an answer token. This penalty encourages the model to establish intermediate reasoning steps before generating the final answer.

As shown in Figure 3, applying PSP effectively shifts the answer generation step to later timesteps compared to the baseline. This indicates that the model performs more complete reasoning before generating the final answer.

4.2. Visual Reasoning Guidance

In Section 3.3, we observe that dMLLMs exhibit weak dependency on visual evidence during reasoning. To address this issue, we extend the principle of Classifier-Free Guidance (CFG) [9], widely used in diffusion models, and propose Visual Reasoning Guidance (VRG).

CFG amplifies the difference between the conditioned distribution $\epsilon_\theta(x_t \mid c)$ and the unconditioned distribution $\epsilon_\theta(x_t)$, encouraging the model to generate samples more faithful to the given condition c (e.g., text prompt). Formally, it can be expressed as:

$$\epsilon_{\text{cfg}} = \epsilon_{\text{uncond}} + s\,(\epsilon_{\text{cond}} - \epsilon_{\text{uncond}}) \tag{5}$$

where s is the guidance scale controlling the influence strength of the condition.

Similarly, in the context of dMLLM reasoning, we compute the conditional logits logits_c (conditioned on the visual prompt v) and the unconditional logits logits_u in parallel, and apply the following VRG formulation:

$$\text{logits}_{\text{vrg}} = \text{logits}_u + (s_{\text{vrg}} + 1) \cdot (\text{logits}_c - \text{logits}_u) \tag{6}$$

where s_{vrg} denotes the visual guidance scale, which amplifies the influence of visual conditioning and strengthens the model's reliance on visual information during reasoning process.

The guided logits are then used to sample tokens at each timestep for prediction.

Finally, by combining VRG and PSP, the confidence score C_j^i used for the remasking strategy is defined as follows.

Table 1. Main result table. X/Y denotes the generation length L and step T, respectively. *Low-conf* refers to the Low-confidence remasking strategy, and *Ours* indicates the Low-conf strategy combined with PSP and VRG. **Bold** represents the best performance, and underline indicates the second-best performance.

Model	Method	M^{3}CoT			MMBench			SQA-IMG			V* Bench		
		64/32	128/64	256/128	64/32	128/64	256/128	64/32	128/64	256/128	64/32	128/64	256/128
LaViDa	DDCoT	45.7	46.7	46.7	72.7	72.8	73.7	71.1	71.1	71.7	41.3	42.4	43.9
	CCoT	45.2	46.5	47.7	72.8	73.9	73.4	71.2	71.1	72.3	42.9	43.9	42.4
	Entropy	46.4	46.9	46.8	72.6	72.6	73.2	70.9	71.4	72.4	42.4	41.8	42.9
	Margin	46.3	46.5	49.2	72.5	73.5	74.1	71.0	71.5	71.8	43.4	43.4	44.5
	Low-conf	45.8	46.2	49.0	72.8	73.2	74.3	71.0	71.1	72.2	42.9	43.4	44.5
	Ours (PSP)	47.6	47.3	**50.5**	74.3	74.6	75.0	72.0	72.5	72.7	44.5	45.5	**46.0**
	Ours (PSP & VRG)	**48.4**	**48.6**	50.3	**74.9**	**75.2**	**75.3**	**72.8**	**72.7**	**73.4**	**45.5**	**46.6**	**46.0**
MMaDa	DDCoT	34.1	34.0	34.1	55.7	55.8	55.5	56.2	56.4	55.8	35.0	34.5	36.1
	CCoT	34.7	31.8	33.3	54.7	55.0	55.0	54.6	56.9	56.9	36.1	36.1	36.6
	Entropy	34.1	33.6	34.3	56.2	55.7	55.5	56.0	56.7	57.3	36.1	35.6	35.0
	Margin	34.5	33.8	33.8	55.6	55.5	55.8	56.0	56.8	57.1	34.0	35.6	34.5
	Low-conf	33.7	33.8	34.6	56.1	56.0	55.7	56.4	56.7	57.3	35.6	35.0	34.5
	Ours (PSP)	35.6	35.2	35.8	57.4	57.5	56.9	**57.3**	57.5	57.6	37.7	38.2	37.1
	Ours (PSP & VRG)	**36.3**	**36.6**	**36.4**	**59.9**	**59.1**	**58.1**	56.9	**58.4**	**58.8**	**38.2**	**38.7**	**37.7**

$$C_j^i = \text{softmax}\big(\text{logits}_{\text{vrg}}\big)_j = \frac{\exp\big(\text{logits}_{\text{vrg},j}\big)}{\sum_{k\in\mathcal{C}} \exp\big(\text{logits}_{\text{vrg},k}\big)} \quad (7)$$

$$\tilde{C}_j^i = \frac{\exp\big(\text{logits}_{\text{vrg},j}\big)}{\sum_{k\in\mathcal{C}} \exp\big(\text{logits}_{\text{vrg},k}\big)} \left[1 - \gamma(1-\tau_i)\,\text{rel}(j)\right] \quad (8)$$

5. Experiments

5.1. Experimental Setup

Benchmarks. We evaluate our proposed method on four benchmarks: M3CoT [2], ScienceQA [22], MMBench [19], and V* Bench [4]. M3CoT is designed to assess multimodal CoT reasoning across various domains, including science, mathematics, and commonsense. ScienceQA evaluates the model's ability to integrate scientific knowledge with visual information for knowledge-based multimodal reasoning. MMBench includes diverse splits to evaluate a model's general visual perception and visual reasoning capabilities. V* Bench is designed to focus on evaluating the model's performance in high-resolution visual question answering tasks.

Models and Baselines. Our experiments are conducted based on two dMLLMs with strong reasoning capabilities: LaViDa-llada-reason [16] and MMaDA-8B-MixCoT [40]. Both models are additionally fine-tuned to generate rationales and are structured to generate progressive responses through an unmasking-based generation process. For a rigorous baseline setup, we include three remasking strategies: Low-confidence (Low-conf), Entropy, and Margin. We also incorporate two CoT-based methods, CCoT [24] and DDCoT [49], which have demonstrated strong performance in VLMs. Both CCoT and DDCoT employ the Low-conf strategy as the default remasking strategy of the dMLLMs.

Implementation Details. We evaluate our method by considering the speed-quality tradeoff, a key advantage of dMLLMs, across different generation lengths and diffusion steps (L / T). The penalty coefficient γ for PSP is fixed at 0.5, and the s_{vrg} for VRG is also set to 0.5. To ensure strict experimental reproducibility, we do not apply temperature scaling and report results obtained using greedy decoding. The performance for each remasking strategy is presented in Table 3, while Low-conf is used as the default remasking strategy in all other experiments. Details on prompt configurations and additional experimental setups are provided in the supplementary material.

5.2. Main Results

In Table 1, existing methods originally designed for VLMs, such as DDCoT and CCoT, fail to achieve strong performance despite generating additional rationales. Notably, both DDCoT, which emphasizes a stepwise divide and conquer approach, and CCoT, which focuses on compositional visual reasoning, show limited improvement. This suggests that dMLLMs require a fundamentally different approach from AR-based VLMs. Neither DDCoT nor CCoT outperforms our methods or the representative unmasking strategies (Entropy, Margin, and Low-conf) under any generation length or diffusion steps.

In contrast, our method consistently outperforms all baselines, including both the remasking strategies (Entropy, Margin, and Low-conf) and VLM CoT meth-

Table 2. LaViDa's ablation study results with generation length $L = 64$ / step $T = 32$. *Low-conf* denotes the Low-confidence remasking strategy, and PSP and VRG are combined with Low-conf. **Bold** represents the best performance, and underline indicates the second-best performance.

Method	M^{3}CoT	MMBench	SQA-IMG	V* Bench
Low-conf	45.8	72.8	71.0	42.9
Low-conf w/ PSP	47.6	74.3	72.0	44.5
Low-conf w/ VRG	47.8	**75.1**	72.1	45.0
Low-conf w/ PSP & VRG	**48.4**	74.9	**72.8**	**45.5**

Table 3. LaViDa's experimental results across different remasking strategies with generation length $L = 64$ / step $T = 32$. PSP and VRG are combined with each method.

Method	M^{3}CoT	MMBench	SQA-IMG	V* Bench
Low-conf	45.8	72.8	71.0	42.9
Low-conf w/ PSP & VRG	48.4	74.9	72.8	45.5
Entropy	46.4	72.6	70.9	42.4
Entropy w/ PSP & VRG	48.0	74.9	72.6	46.0
Margin	46.3	72.5	71.0	43.4
Margin w/ PSP & VRG	48.1	74.8	72.9	45.0

ods (DDCoT / CCoT), across all experiments. First, our approach consistently delivers strong performance on both perception-focused benchmarks (MMBench and SQA-IMG) and reasoning-intensive benchmarks (M3CoT and V* Bench), demonstrating its robustness across different task types and difficulty levels. Second, despite the inherent difference in model reasoning capability between LaViDa and MMaDa, our method consistently improves performance across both models, showing that the effectiveness of our approach is not sensitive to model scale or base performance. Lastly, across all settings of generation length and diffusion steps, which are the key parameters of dMLLMs, our method achieves at least a 3% improvement in accuracy. This demonstrates its general applicability and effectiveness across diverse dMLLM configurations.

From the perspective of efficiency, our method achieves superior performance under the L/T = 64/32 compared to configurations using four times more diffusion steps. Given that dMLLMs inherently allow flexible control over the tradeoff between inference speed and response quality, our method effectively maximize the model's capability. For example, LaViDa achieves 74.3% accuracy using a generation length of 256 with the Low-conf strategy, whereas Low-conf + PSP & VRG achieves 74.9% accuracy with only a generation length of 64 in the MMBench benchmark. Similarly, on the same benchmark, MMaDa achieves 7.5% higher accuracy with a generation length of 64 using our method compared to using a generation length of 256 with the Low-conf strategy. These results demonstrate both superior efficiency and reasoning quality.

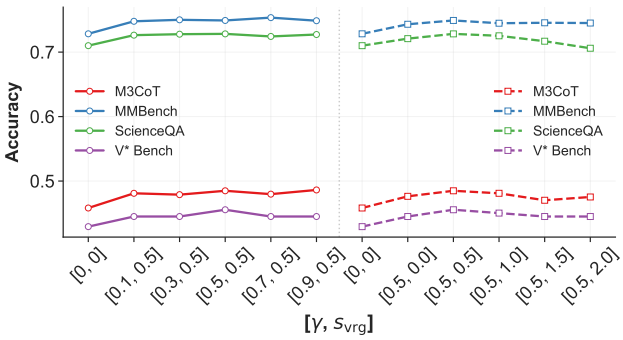


Figure 5. Performance comparison of LaViDa under different hyperparameter settings. On the horizontal axis, the notation $[x, y]$ represents the hyperparameter pair, where x denotes the penalty strength coefficient γ and y denotes the visual guidance scale s_{vrg}.

5.3. Ablation & Discussion

Ablation Study. In Table 2, we present the ablation study of PSP and VRG using LaViDa. Both PSP and VRG individually improve performance over the Low-conf (baseline), demonstrating that each method is effective on its own. Furthermore, applying them together yields the best performance across four benchmarks (M3CoT, MMBench, SQA-IMG, and V* Bench).

Generalization. Table 3 presents the results of applying PSP and VRG across three representative remasking strategies in dMLLMs. The remasking strategy is applied by replacing the confidence score in Equation 4 with entropy or margin. Experimental results show that our method consistently achieves significant performance improvements regardless of the remasking strategy. This suggests that PSP and VRG can operate in a plug-and-play manner with any remasking strategy, even as new strategies are introduced. In addition, Figure 5 shows the experimental results under different hyperparameter settings. Both PSP and VRG exhibit low sensitivity to hyperparameter variations while consistently outperforming the Low-conf (baseline) across all datasets. However, if s_{vrg} is set excessively large, the model may ignore textual information and rely solely on visual information, so it is necessary to choose an appropriate value.

Efficiency. Table 4 presents the results of the time cost analysis. For a fair comparison, we use Low-conf as the default remasking strategy. The results show that DDCoT, which separates model inputs into reasoning and recognition for inference, incurs the highest time cost across all generation length and diffusion steps settings. In contrast, CCoT and PSP exhibit similar time costs to the Low-conf (baseline). Notably, PSP achieves better or comparable efficiency while consistently outperforming the Low-conf, DD-

Table 4. Average time consumption of LaViDa on the M3CoT validation set. The default remasking strategy used was Low-confidence, and DDCoT, CCoT, PSP, VRG, and PSP & VRG were applied. The unit s denotes seconds.

Method	**64/32**	**128/64**	**256/128**
Low-conf	4.01s	6.07s	14.48s
Low-conf w/ DDCoT	6.03s	9.99s	25.95s
Low-conf w/ CCoT	4.05s	6.09s	14.59s
Low-conf w/ PSP	4.03s	6.06s	14.51s
Low-conf w/ VRG	4.73s	8.46s	22.05s
Low-conf w/ PSP & VRG	4.65s	8.43s	22.29s

Table 5. Comparison experimental results between PSP and L2R in LaViDa with generation length $L = 64$ / step $T = 32$. **Bold** indicates the better result between L2R & VRG and PSP & VRG, while underline denotes the better result between L2R and PSP.

Method	**M^{3}CoT**	**MMBench**	**SQA-IMG**	**V* Bench**
Low-conf	45.8	72.8	71.0	42.9
Low-conf w/ L2R	47.2	74.0	71.1	43.4
Low-conf w/ PSP	47.6	74.3	72.0	44.5
Low-conf w/ L2R & VRG	47.6	74.6	71.3	44.5
Low-conf w/ PSP & VRG	**48.4**	**74.9**	**72.8**	**45.5**

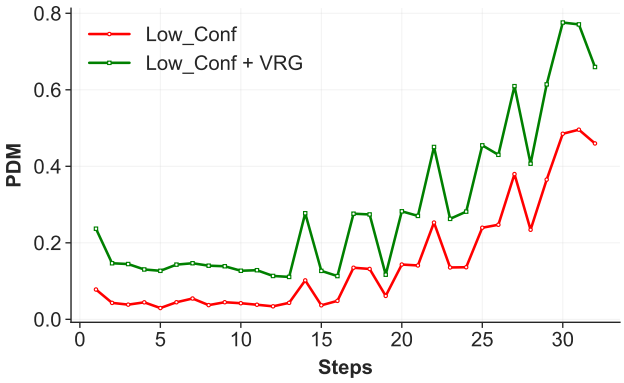
CoT, and CCoT in all settings. VRG introduces a slightly higher time cost due to the computation of conditional and unconditional logits, but it remains highly competitive. It also shows no significant increase in time under the L/T = 64/32 setting while still outperforming the L/T = 256/128 configuration.

Discussion. To compare PSP with AR generation, we conduct an experiment in which dLLMs generate tokens in a Left-to-Right (L2R) manner by unmasking the leftmost position at each timesteps. For example, when $L = 64$ and step $T = 32$, two leftmost remaining masked tokens are unmasked at each timesteps. As shown in Table 5, PSP consistently outperforms the L2R method across all datasets. Moreover, when combined with VRG, PSP & VRG achieve higher performance than L2R & VRG in every dataset.

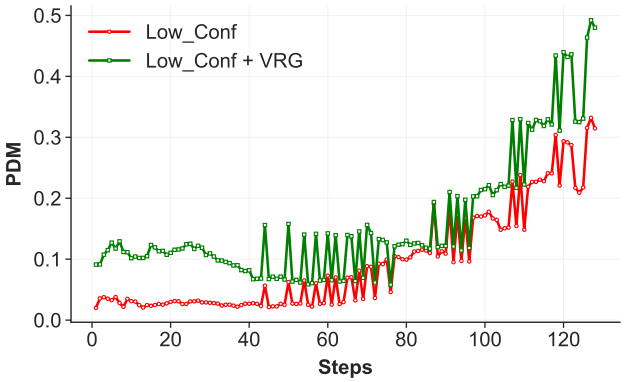
Figure 6 shows how the average PDM vary over timesteps on the M3CoT validation set when using LaViDa with Low-conf (base) and VRG. In the Low-conf, PDM remains relatively low during the early timesteps, indicating that the model initially relies on limited visual information. When VRG is applied, PDM is consistently higher across timesteps, demonstrating that visual grounding is effectively strengthened throughout the generation process.

6. Conclusion

In this paper, we analyze the reasoning process of dMLLMs and propose a novel method to address their limitations. Our analysis shows that dMLLMs exhibit two major issues: (1) Early Answer Generation, where the model tends



(a) LaVida with generation length L = 64 / step T = 32



(b) LaVida with generation length L = 256 / step T = 128

Figure 6. Comparison of PDM measurements on the M3CoT validation set between Low-conf and Low-conf + VRG using LaViDa. the visual guidance scale s_{vrg} = 0.5.

to generate the final answer prematurely before undergoing sufficient reasoning steps, and (2) low dependency on visual prompts during the early timesteps. These findings suggest that, due to the parallel token restoration mechanism inherent to dMLLMs, it is challenging to directly apply reasoning enhancement methods originally designed for AR models. To mitigate these issues, we propose Position & Step Penalty (PSP), which encourages dMLLMs to perform progressive reasoning throughout the diffusion process rather than generating final answers too early. In addition, we introduce Visual Reasoning Guidance (VRG) to strengthen the model's reliance on visual evidence. As a result, we achieve improvements in both efficiency and reasoning quality without additional training. Our study provides a promising direction for enabling dMLLMs to approach the reasoning capabilities of AR VLMs, and we expect it will further promote the development and utilization of dMLLMs in future research.

Acknowledgements

This work was supported by the Institute of Information and communications Technology Planning and evaluation (IITP) grant (No.RS-2025-25422680, No. RS-2020-II201373), and the National Research Foundation of Korea (NRF) grant (No. RS-2025-00520618) funded by the Korean Government (MSIT).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1

[2] Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M3 cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *arXiv preprint arXiv:2405.16473*, 2024. 4, 6

[3] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024. 1

[4] Zixu Cheng, Jian Hu, Ziquan Liu, Chenyang Si, Wei Li, and Shaogang Gong. V-star: Benchmarking video-llms on video spatio-temporal reasoning. *arXiv preprint arXiv:2503.11495*, 2025. 6

[5] Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, et al. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022. 2

[6] Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312, 2024. 1, 4

[7] Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19520–19529, 2025. 3

[8] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022. 2

[9] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5

[10] Zemin Huang, Yuhang Wang, Zhiyang Chen, and Guo-Jun Qi. Don't settle too early: Self-reflective remasking for diffusion language models. *arXiv preprint arXiv:2509.23653*, 2025. 3

[11] Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, 2023. 3

[12] Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25004–25014, 2025. 1

[13] Mingi Jung, Saehyung Lee, Eunji Kim, and Sungroh Yoon. Visual attention never fades: Selective progressive attention recalibration for detailed image captioning in multimodal large language models. *arXiv preprint arXiv:2502.01419*, 2025. 1, 4

[14] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023. 1

[15] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1

[16] 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. *arXiv preprint arXiv:2505.16839*, 2025. 1, 6

[17] Xiang Li, John Thickstun, Ishaan Gulrajani, Percy S Liang, and Tatsunori B Hashimoto. Diffusion-lm improves controllable text generation. *Advances in neural information processing systems*, 35:4328–4343, 2022. 2

[18] 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*, pages 26296–26306, 2024. 4

[19] 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*, pages 216–233. Springer, 2024. 6

[20] Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution, 2024. *URL https://arxiv. org/abs/2310.16834*. 2

[21] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 3

[22] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 6

[23] Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017, 2023. 3

[24] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431, 2024. 1, 3, 6

[25] Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025. 1, 2

[26] Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies. *arXiv preprint arXiv:2308.03188*, 2023. 3

[27] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[29] Subham Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37:130136–130184, 2024. 2

[30] Subham Sekhar Sahoo, Justin Deschenaux, Aaron Gokaslan, Guanghan Wang, Justin Chiu, and Volodymyr Kuleshov. The diffusion duality. *arXiv preprint arXiv:2506.10892*, 2025. 3

[31] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024. 3

[32] Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. Llm-check: Investigating detection of hallucinations in large language models. *Advances in Neural Information Processing Systems*, 37:34188–34216, 2024. 3

[33] Guanghan Wang, Yair Schiff, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Remasking discrete diffusion models with inference-time scaling. *arXiv preprint arXiv:2503.00307*, 2025. 3

[34] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 1

[35] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. 3

[36] Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025. 3

[37] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 3

[38] 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. 3

[39] Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2087–2098, 2025. 1, 3

[40] Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. Mmada: Multimodal large diffusion language models. *arXiv preprint arXiv:2505.15809*, 2025. 1, 6

[41] 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. 1, 2

[42] 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. 1

[43] Runpeng Yu, Xinyin Ma, and Xinchao Wang. Dimple: Discrete diffusion multimodal large language model with parallel decoding. *arXiv preprint arXiv:2505.16990*, 2025. 1

[44] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024. 1

[45] Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1662, 2025. 3

[46] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022. 3

[47] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023. 3

[48] Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, et al. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings*

of the Computer Vision and Pattern Recognition Conference, pages 1702–1713, 2025. 1, 3

[49] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. *Advances in Neural Information Processing Systems*, 36:5168–5191, 2023. 3, 6

Thinking Diffusion: Penalize and Guide Visual-Grounded Reasoning in Diffusion Multimodal Language Models

Supplementary Material

Input:

What is the most likely purpose of the tall red chair with a horse on it?
A. A playground for children
B. A sculpture or art installation
C. A seat for a giant
D. A prop for a movie set

Please reason step by step, and answer the question with option letter from given choices in the format of Answer: <option letter>.

(a) LaViDa

Input:

You should first think about the reasoning process in the mind and then provide the user with the answer. The reasoning process is enclosed within <think> </think> tags, i.e. <think> reasoning process here </think> answer here

What can you infer about the person from the image?
A. The person likes to eat out often
B. The person lives alone
C. The person eats a lot of frozen meals
D. The person likes to keep their surroundings clean

(b) MMaDa

Figure 7. Example of a reasoning prompt based on each model's reference implementation.

7. Prompting Details

This section provides additional details about our prompting used for diffusion-based multimodal reasoning. Following LaViDa and MMaDa, we adopt the think prompt to encourage structured, step-by-step reasoning during generation. The think prompt guides the model to first produce intermediate reasoning before generating the final answer, thereby improving interpretability and mitigating early answer generation. Figure 7 shows the complete think prompt templates used in all our experiments.

8. Additional Results

8.1. Additional Analysis

In this subsection, we conduct additional experiments on MMaDa. The results corresponding to Observation 1 and Observation 2 are presented in Figure 8 and Figure 9, respectively. As shown in Figure 8, MMaDa exhibits a clear Early Answer Generation, similar to what we observe in LaViDa. The model frequently generates the final answer at very early timesteps, indicating premature answer determination before sufficient reasoning. However, when PSP is applied, the distribution shifts toward later timesteps, encouraging more gradual reasoning and effectively mitigating early answer generation.

Likewise, Figure 9 shows that MMaDa demonstrates low visual dependence during early timesteps, again consistent with the property seen in LaViDa. The model begins to meaningfully incorporate visual information only in later steps. Applying VRG significantly increases visual dependency across the diffusion process, reinforcing visual grounding and enabling earlier and more consistent use of visual evidence. These results indicate that PSP and VRG improve reasoning progression and visual grounding not only in LaViDa but also in MMaDa, demonstrating their general applicability across different dMLLMs.

8.2. Analysis on Varying Diffusion Steps

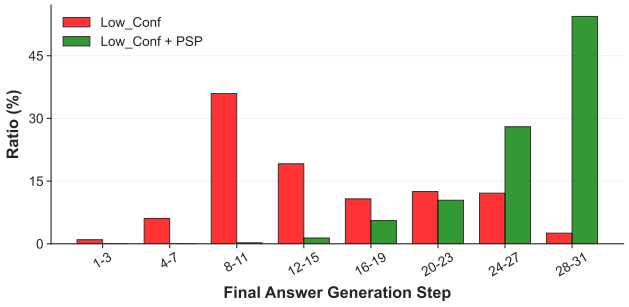
We further analyze LaViDa by fixing the generation length to $L = 64$ and varying the number of diffusion steps across three settings: $T = 8$, $T = 16$, and $T = 32$. The corresponding results for Early Answer Generation and visual prompt dependency are presented in Figure 10 and Figure 11, respectively.

As shown in Figure 10, LaViDa consistently exhibits strong Early Answer Generation when operating under low diffusion steps. With $T = 8$, the model frequently generates the final answer within the first few steps, indicating premature answer formation. Increasing the number of diffusion steps to $T = 16$ and $T = 32$ shifts the answer-generation distribution toward later timesteps, but the early-answer tendency remains visible. Applying PSP effectively suppresses this behavior across all three settings, pushing the answer generation toward later stages and encouraging more gradual reasoning even when the diffusion schedule is highly constrained.

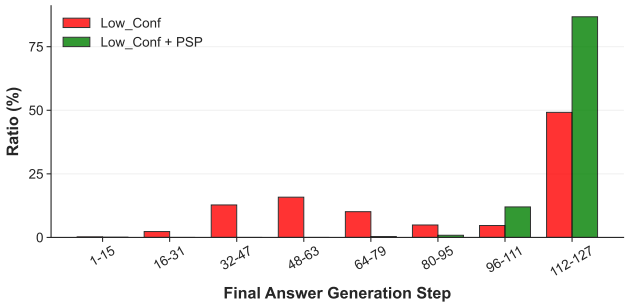
Similarly, Figure 11 shows the visual prompt dependency (PDM) under the same settings. When using the de-

Table 6. Comparison of dMLLM under varying numbers of diffusion steps T with a fixed generation length $L = 64$. X/Y denotes the generation length L and step size T, respectively.

Model	Method	M^{3}CoT				MMBench				SQA-IMG			
		64/8	64/16	64/32	64/64	64/8	64/16	64/32	64/64	64/8	64/16	64/32	64/64
LaViDa	Entropy	46.2	46.3	46.4	47.0	72.7	72.8	72.6	72.7	70.8	70.9	70.9	71.2
	Margin	46.4	46.5	46.3	46.9	72.3	72.9	72.5	73.0	71.3	71.0	71.0	71.5
	Low-conf	45.3	45.7	45.8	46.4	72.6	72.9	72.8	72.8	71.1	70.6	71.0	71.3
	Ours	**47.9**	**47.7**	**48.4**	**48.5**	**74.4**	**74.9**	**74.9**	**74.8**	**72.1**	**72.3**	**72.8**	**72.7**

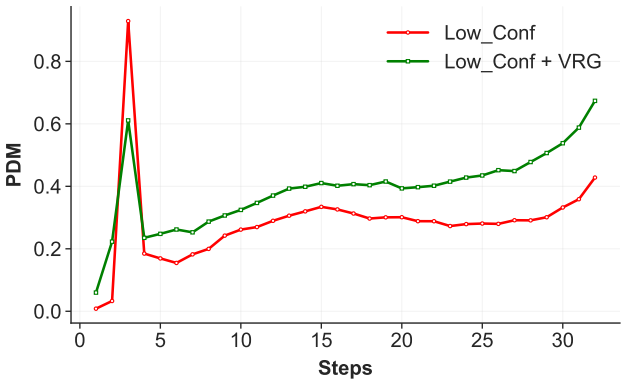


(a) Generation length $L = 64$ / diffusion step $T = 32$

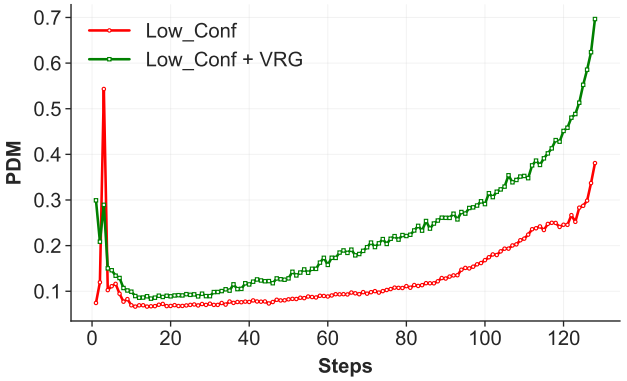


(b) Generation length $L = 256$ / diffusion step $T = 128$

Figure 8. Results of the final answer generation step on the M3CoT validation set using MMaDa. The default remasking strategy is Low-confidence.



(a) MMaDa with generation length $L = 64$ / step $T = 32$



(b) MMaDa with generation length $L = 256$ / step $T = 128$

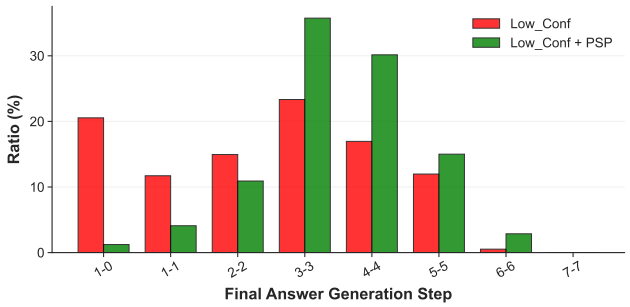
Figure 9. Comparison of PDM measurements on the M3CoT validation set between Low-conf and Low-conf + VRG using MMaDa.

fault remasking strategy, LaViDa shows weak visual dependence at early steps across all values of T. The PDM gradually increases as the diffusion progresses, but the early-stage visual grounding remains minimal. With the application of VRG, the PDM curves consistently rise across all diffusion steps, demonstrating substantially stronger and earlier integration of visual information. Notably, the improvement becomes more pronounced as T increases, showing that VRG enhances visual grounding regardless of the diffusion schedule.

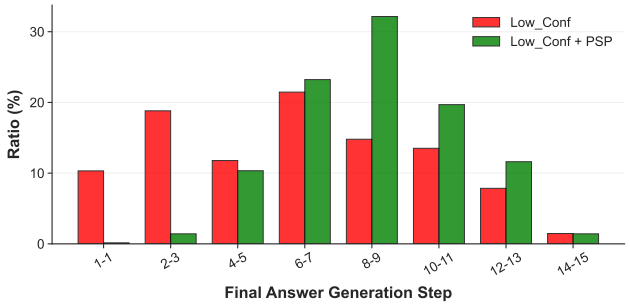
These findings indicate that the reasoning characteristics observed in the main paper, namely early answer generation and weak early visual grounding, persist across different diffusion step configurations. Furthermore, PSP and VRG reliably address these issues under all tested settings, confirming their robustness even when the diffusion budget is significantly limited.

8.3. Performance on Diffusion Steps

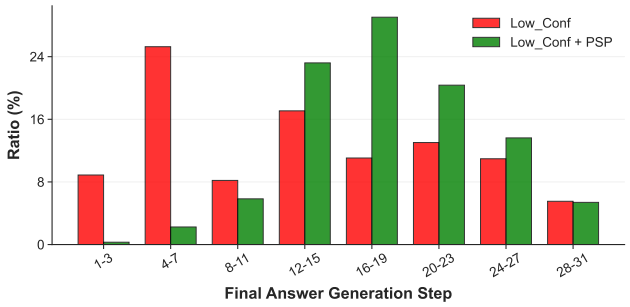
We further investigate the influence of the number of diffusion steps T on the reasoning performance. To isolate the effect of T, we fix the generation length to $L = 64$ for all experiments and vary only the number of diffusion



(a) Generation length $L = 64$ / diffusion step $T = 8$



(b) Generation length $L = 64$ / diffusion step $T = 16$



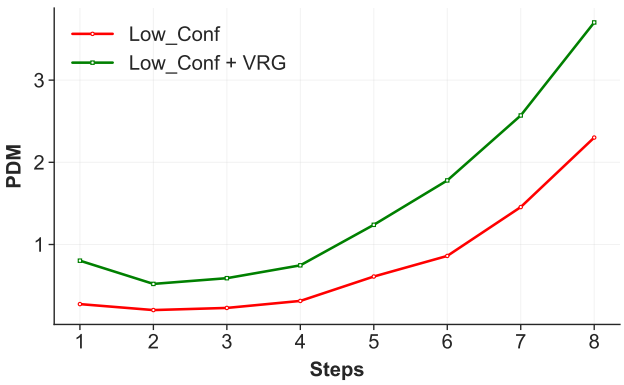
(c) Generation length $L = 64$ / diffusion step $T = 32$

Figure 10. Results of the final answer generation step on the M3CoT validation set using LaViDa, evaluated across different diffusion steps T.

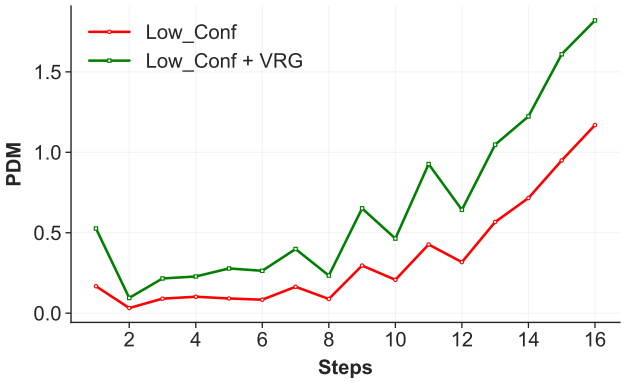
steps. Across all evaluated remasking strategies, we observe a trend that a reduction in the number of diffusion steps results in a consistent decrease in overall accuracy. Despite the reduction in performance at small T, our method remains highly effective. Across all tested values of T, our approach achieves state-of-the-art performance under very constrained diffusion schedules, as shown in Table 6. These results indicate that our method not only improves reasoning quality but also provides robustness to the choice of diffusion steps, enabling stable performance across a wide range of inference-time configurations.

8.4. Experiments on Long-form Answers

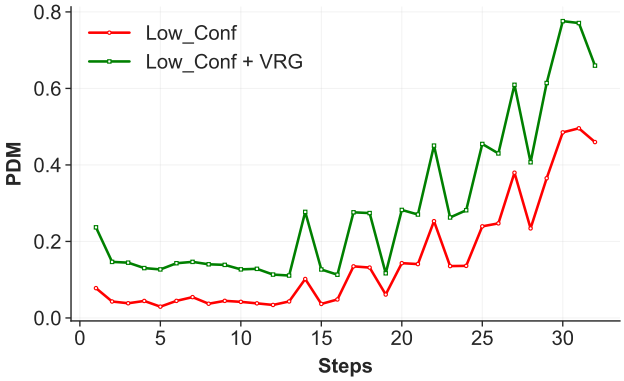
We further investigate whether the issue described in Observation 1 also arises in long-form answer settings. To



(a) LaVida with generation length $L = 64$ / step $T = 8$



(b) LaVida with generation length $L = 64$ / step $T = 16$



(c) LaVida with generation length $L = 64$ / step $T = 32$

Figure 11. Comparison of PDM measurements on the M3CoT validation set between Low-conf and Low-conf + VRG using LaViDa, evaluated across different diffusion steps T.

this end, we conduct the same analysis on LLaVA-Bench COCO with long-form answers. By enforcing the output format `<Mask><Mask> Answer: <Mask><Mask>`, we define the first timestep at which 75% of the tokens following "Answer:" are filled as the answer generation point. As shown in Figure 12, Early Answer Generation is consistently observed.

	MME Exist. ↑	MME Count ↑	MME Pos. ↑	MME Color ↑	MME Total ↑	LLaVA-Bench ↑
Low Conf.	183.33	133.33	86.67	141.67	545.00	16.5
CCoT	185.00	126.67	90.55	148.33	550.55	17.2
DDCoT	176.67	137.22	81.67	149.17	544.73	15.4
PSP & VRG	**187.00**	**143.33**	**91.67**	**150.00**	**570.00**	**20.0**

Table 7. Experimental results on the MME benchmark for LaViDa with the 64/128 setting.

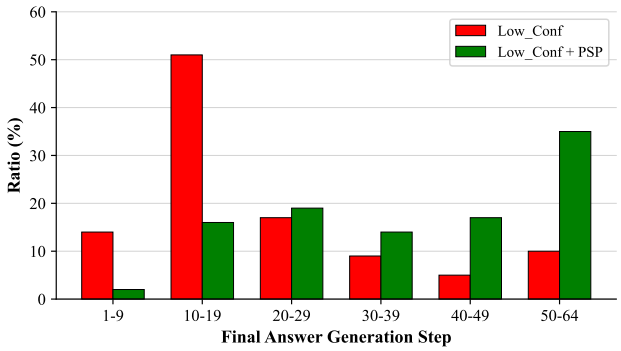


Figure 12. Observation 1 for LLaVA-Bench COCO.

8.5. Results on Complex Metrics / Datasets

To evaluate a broader range of perception and cognition abilities beyond simple accuracy-based measurements, we conduct experiments on the MME benchmark and LLaVA-Bench. We additionally evaluate image perception and cognition using MME and conduct experiments on LLaVA-Bench with descriptive responses. As shown in Table 7, PSP & VRG consistently achieves the best performance across all metrics.