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

1. 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 1 shows the complete think prompt templates used in all our experiments.

2. Additional Results

2.1. Additional Analysis

In this subsection, we conduct additional experiments on MMaDa. The results corresponding to Observation 1 and Observation 2 are presented in Figure 2 and Figure 3, respectively. As shown in Figure 2, 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 3 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.

2.2. Analysis on Varying Diffusion Steps

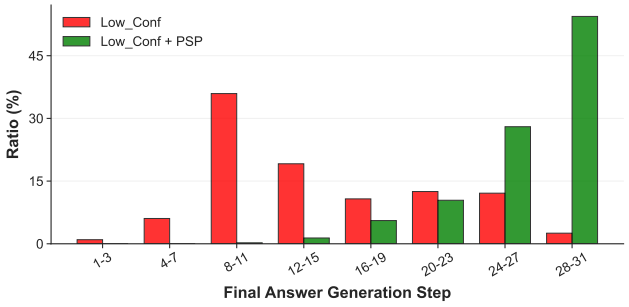
We further analyze LaViDa by fixing the generation length to $L = 64$ and varying the number of diffusion steps across three settings: $T = 8$, $T = 16$, and $T = 32$. The corresponding results for Early Answer Generation and visual prompt dependency are presented in Figure 4 and Figure 5, respectively.

As shown in Figure 4, 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.

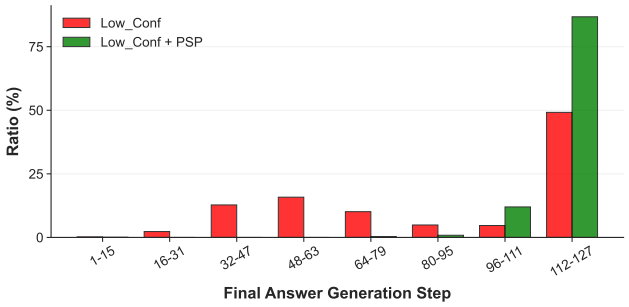
Similarly, Figure 5 shows the visual prompt dependency (PDM) under the same settings. When using the default re-

Table 1. 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**

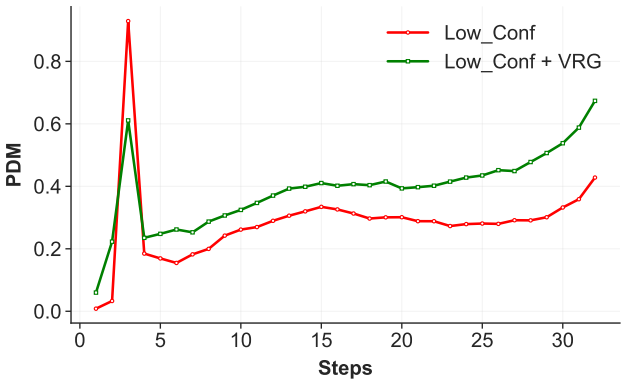


(a) Generation length L = 64 / diffusion step T =32

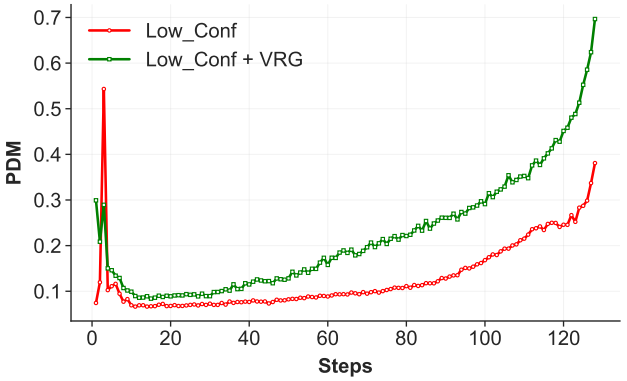


(b) Generation length L = 256 / diffusion step T = 128

Figure 2. 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

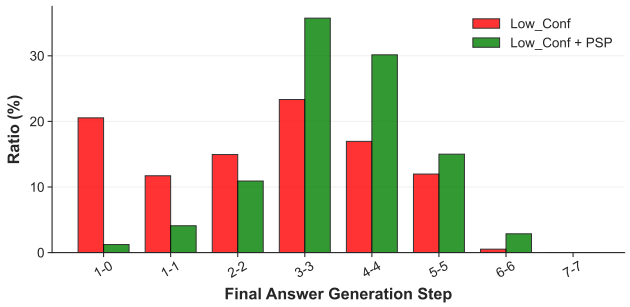
Figure 3. Comparison of PDM measurements on the M3CoT validation set between Low-conf and Low-conf + VRG using MMaDa.

masking 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.

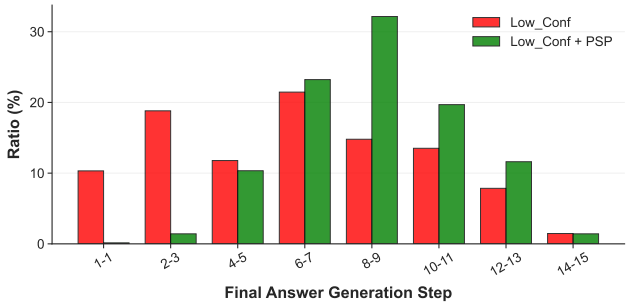
These findings indicate that the reasoning characteristics observed in the main paper, namely early answer generation and weak early visual grounding, persist across different diffusion step configurations. Furthermore, PSP and VRG reliably address these issues under all tested settings, confirming their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

2.3. Performance on Diffusion Steps

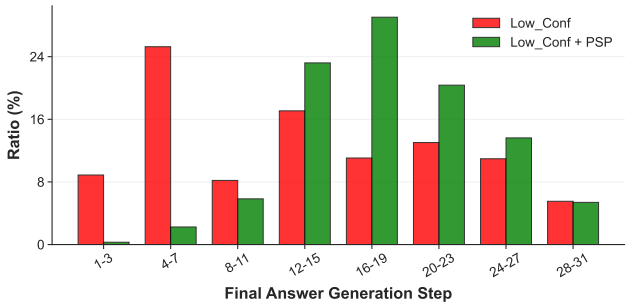
We further investigate the influence of the number of diffusion steps T on the reasoning performance. To isolate the effect of T, we fix the generation length to $L = 64$ for all experiments and vary only the number of diffusion



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



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



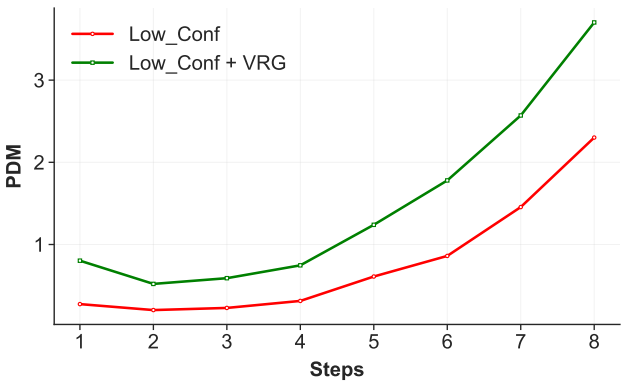
(c) Generation length $L = 64$ / diffusion step $T = 32$

Figure 4. Results of the final answer generation step on the M3CoT validation set using LaViDa, evaluated across different diffusion steps T.

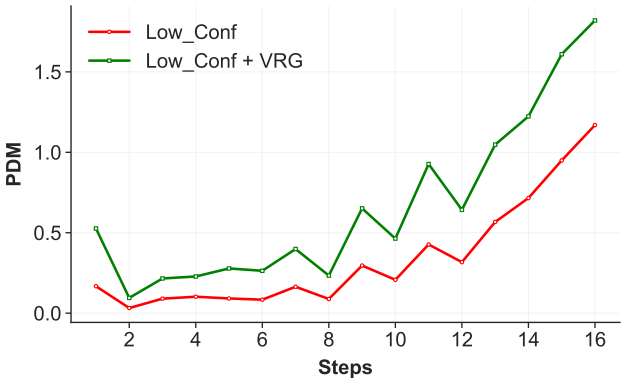
steps. Across all evaluated remasking strategies, we observe a trend that a reduction in the number of diffusion steps results in a consistent decrease in overall accuracy. Despite the reduction in performance at small T, our method remains highly effective. Across all tested values of T, our approach achieves state-of-the-art performance under very constrained diffusion schedules, as shown in Table 1. 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.

2.4. Experiments on Long-form Answers

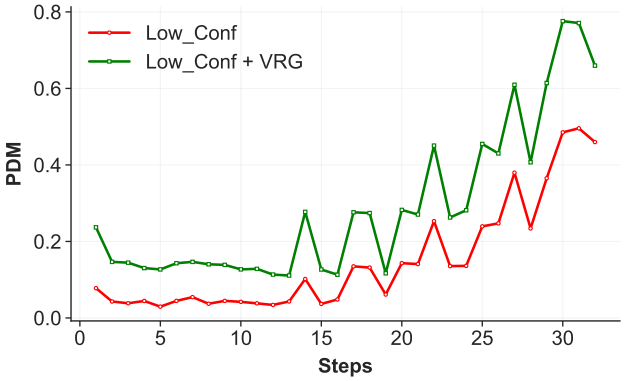
We further investigate whether the issue described in Observation 1 also arises in long-form answer settings. To



(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 5. 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. Results on MME and LLaVA-Bench.

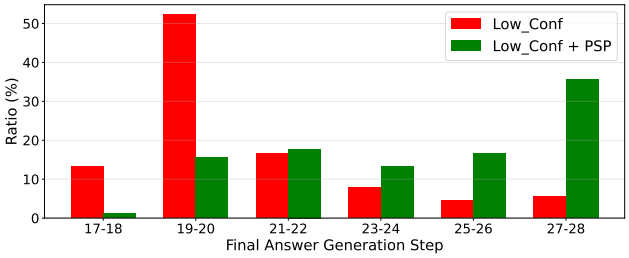


Figure 12. Observation 1 for LLaVA-Bench COCO.

2.5. 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.