HANYANG UNIVERSITY

CVPR JUNE 3-7, 2026 DENVER COLORADO

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

**Keuntae Kim*, Mingyu Kang*, Yong Suk Choi†**

Hanyang University, Seoul, Korea

(*: Equal contribution, †: Corresponding author)

HYU Artificial Intelligence Laboratory

Contact us: ktkpv94@hanyang.ac.kr

# Introduction

## Background

- **Diffusion (M)LLMs** are emerging as **promising alternatives to autoregressive LLMs**
- By **generating tokens in parallel with a pre-defined output length**, diffusion-based models enable more efficient generation than sequential autoregressive decoding

## Motivation

- Despite their efficiency, **dMLLMs struggle with visual-grounded Chain-of-Thought reasoning**
- They often **determine the final answer before sufficient reasoning** and **rely weakly on visual prompts in early diffusion steps**
- This motivates methods that delay premature answer generation and strengthen visual grounding during reasoning

# Introduction

## Overview

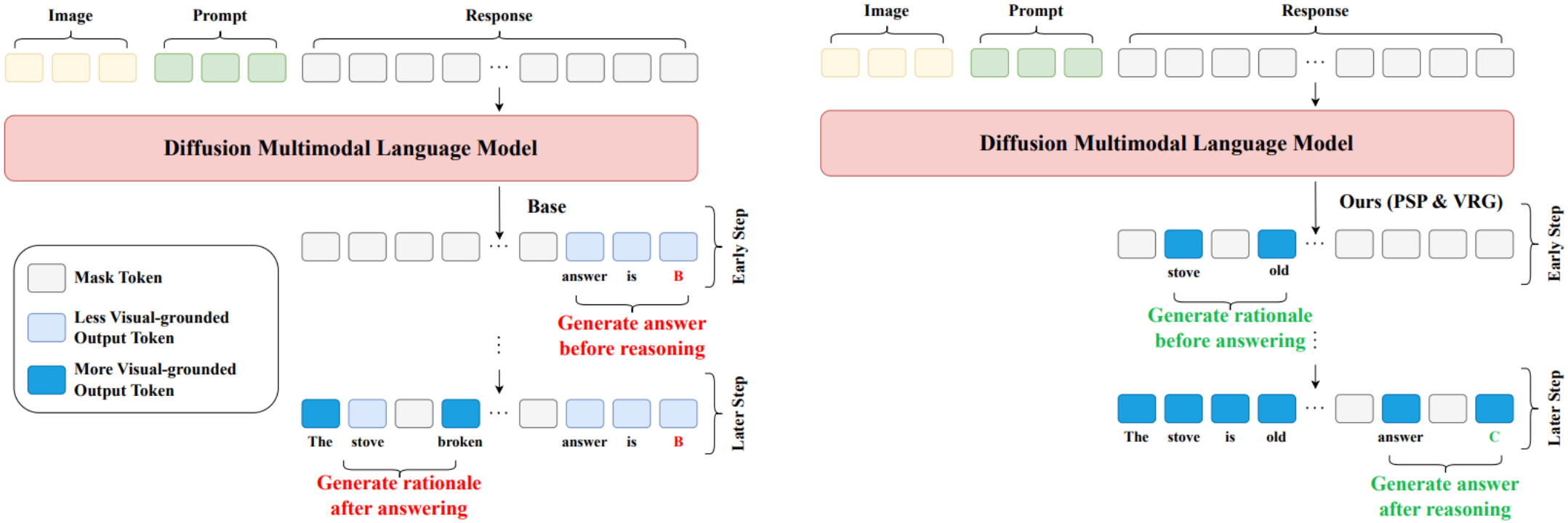
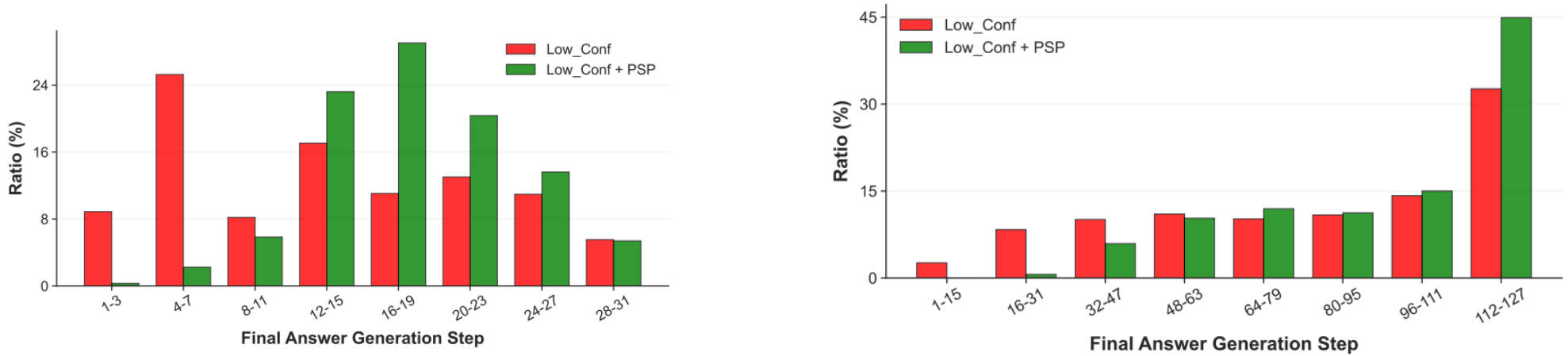



**Figure 1.** Overview of our method and comparison with the baseline

# Analysis of Reasoning in dMLLMs

CVPR JUNE 3-7, 2026 DENVER COLORADO

## Observation 1: Early Answer Generation




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

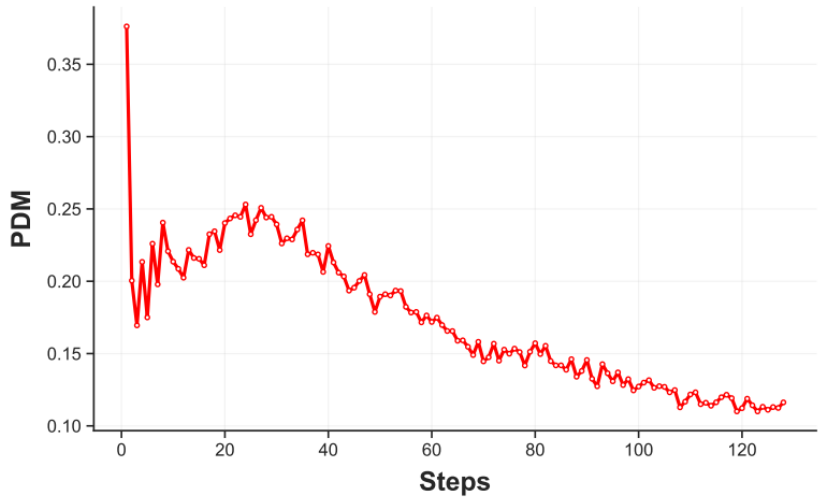
(b) Generation length $L = 256$ / diffusion step $T = 128$

**Figure 2.** Result of the final answer generation step in the M3CoT validation set using LaViDa

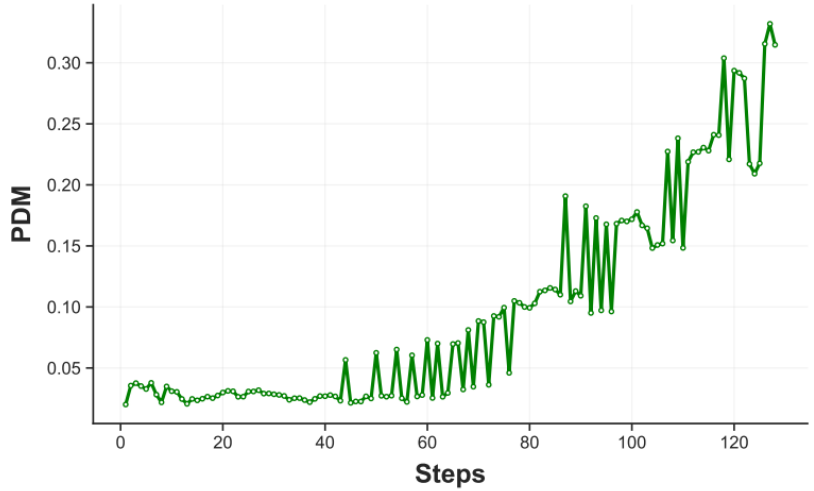
- **Measuring Answer Emergence:** Track the diffusion step where the final answer token first appears under different length and step settings.
- **Early Answer Generation:** dMLLMs often generate the final answer too early, especially when generation length and diffusion steps are limited.

# Analysis of Reasoning in dMLLMs

## Observation 2: Week Visual Grounding




(a) LLaVA-1.5




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

**Figure 3.** Comparison of PDM measurements on the M3CoT validation set between the autoregressive-based model

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

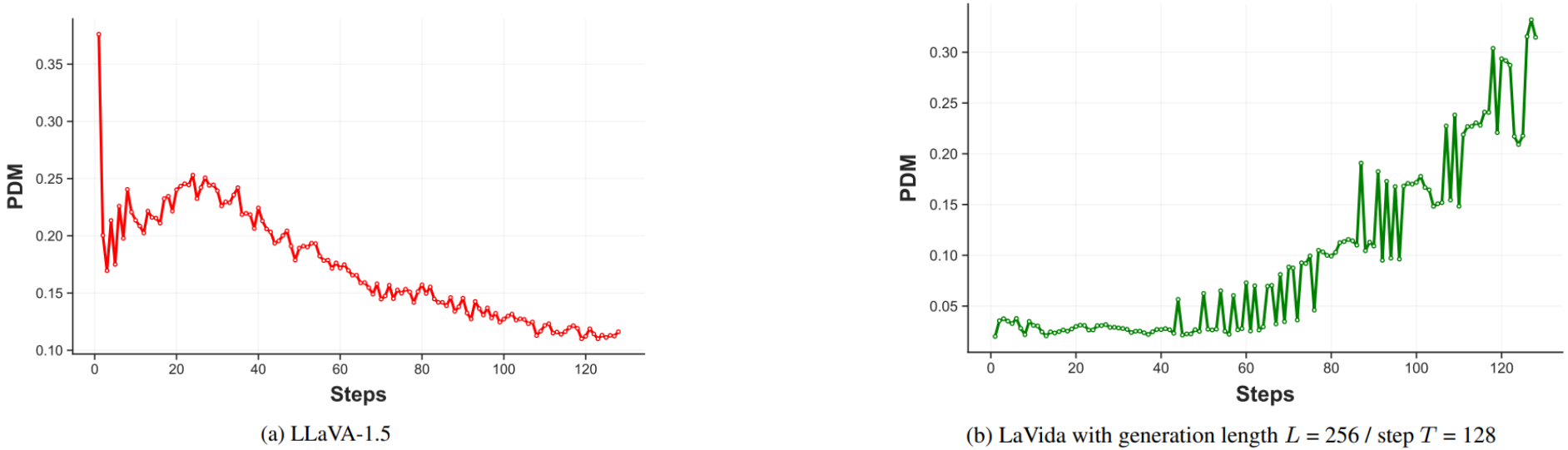
where $v$, $c$, and $X_t$ denote the visual input, the prompt, and the partially masked sequence, respectively.



[1] Multi-modal hallucination control by visual information grounding, CVPR 2024

# Analysis of Reasoning in dMLLMs

## Observation 2: Week Visual Grounding




(a) LLaVA-1.5

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

**Figure 3.** Comparison of PDM measurements on the M3CoT validation set between the autoregressive-based model

- **Measuring Visual Grounding with PDM:** Compare token distributions conditioned on both text and image inputs versus text-only inputs to quantify visual prompt dependency.
- **Delayed Visual Utilization in dMLLMs:** dMLLMs rely weakly on visual information at early diffusion steps, suggesting that answers can be generated before sufficient visual grounding.

# Method – Position & Step Penalty

CVPR JUNE 3-7, 2026 DENVER COLORADO

## Method 1: Position & Step Penalty (PSP)

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

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

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

- $C_{ij}$ : Original confidence of the j-th token at the i-th diffusion step
- $\tilde{C}_{ij}$ : Modified confidence after applying PSP
- $\gamma$ : Penalty strength coefficient
- $\tau_i$ : Current diffusion progress, i / K → $(1 - \tau_i)$ becomes larger at earlier diffusion steps
- rel(j): Relative position of the j-th token in the response → rel(j) becomes larger for later token positions

# Method – Visual Reasoning Guidance

CVPR JUNE 3-7, 2026 DENVER COLORADO

## Method 2: Visual Reasoning Guidance (VRG)

Similarly, in the context of dMLLM reasoning, we compute the conditional logits $\text{logits}_c$ (conditioned on the visual prompt $v$) and the unconditional logits $\text{logits}_u$ in parallel, and apply the following VRG formulation:

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

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

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

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

- logits_c : Conditional logits computed with the visual prompt v
- logits_u : Unconditional logits computed without the visual prompt
- s_vrg : Visual guidance scale
- logits_vrg : Final guided logits after applying VRG

# Experiment

CVPR JUNE 3-7, 2026 DENVER COLORADO

## Experimental Setup

- **Benchmarks:** M3CoT, ScienceQA, MMBench, and V* Bench → Multimodal Reasoning Benchmarks

- **Models and Baselines:** LaViDa-llada-reason, MMaDA-8B-MixCoT → Reasoning Models

- **Implementation Details:** Evaluate multiple (generation length L / diffusion step T) settings for speed-quality tradeoff. Set $\gamma = 0.5$ for PSP and $s_{vrg} = 0.5$ for VRG. Use greedy decoding without temperature scaling

# Results



## Main Results

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** |

# Results



## Ablation Study

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 |

# Analysis

## Analysis of PSP & VRG

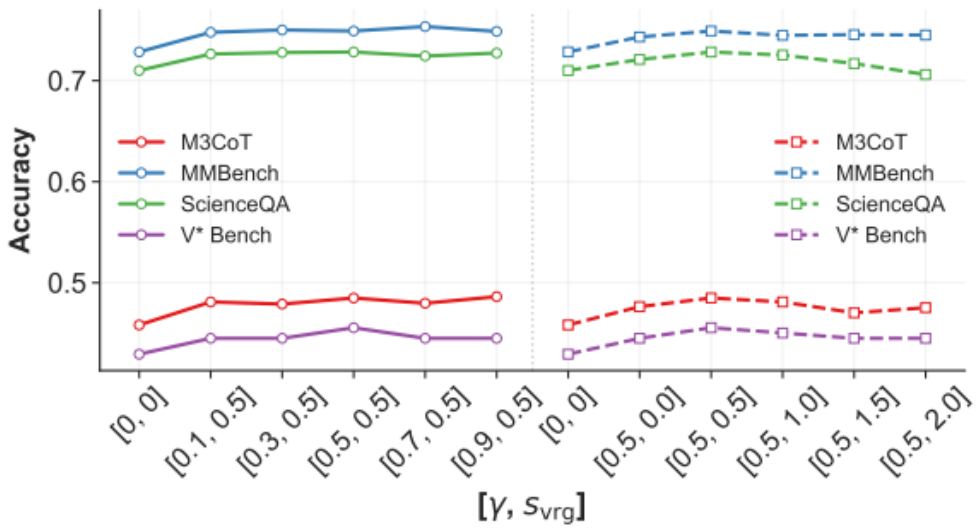



Figure 5. Performance comparison of LaViDa under different hyperparameter settings. On the horizontal axis, the notation $[x, y]$ represents the hyperparameter pair, where $x$ denotes the penalty strength coefficient $\gamma$ and $y$ denotes the visual guidance scale $s_{vrg}$.

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.

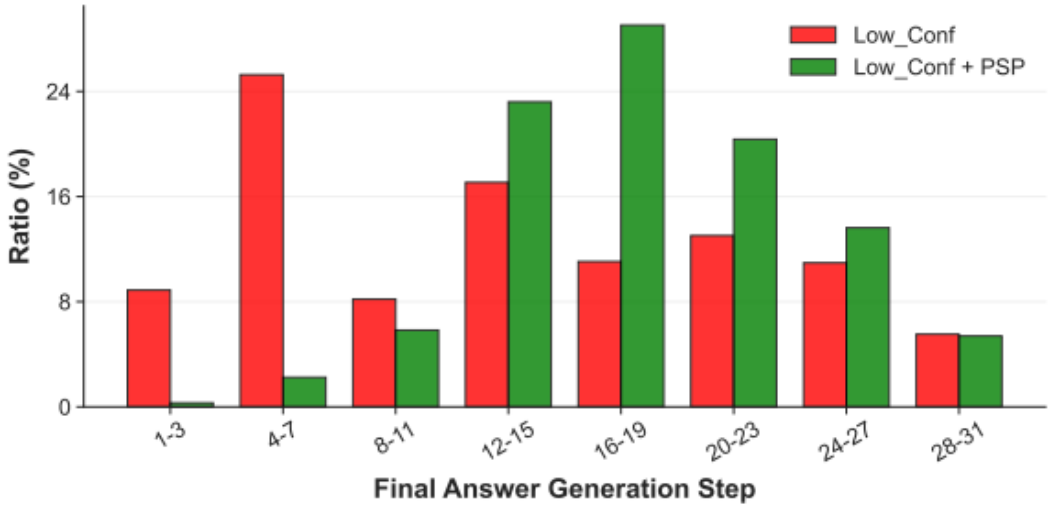
| Method | 64/32 | 128/64 | 256/128 |
|---|---|---|---|
| Low-conf | 4.01s | 6.07s | 14.48s |
| Low-conf w/ DDCoT | 6.03s | 9.99s | 25.95s |
| Low-conf w/ CCoT | 4.05s | 6.09s | 14.59s |
| Low-conf w/ PSP | 4.03s | 6.06s | 14.51s |
| Low-conf w/ VRG | 4.73s | 8.46s | 22.05s |
| Low-conf w/ PSP & VRG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

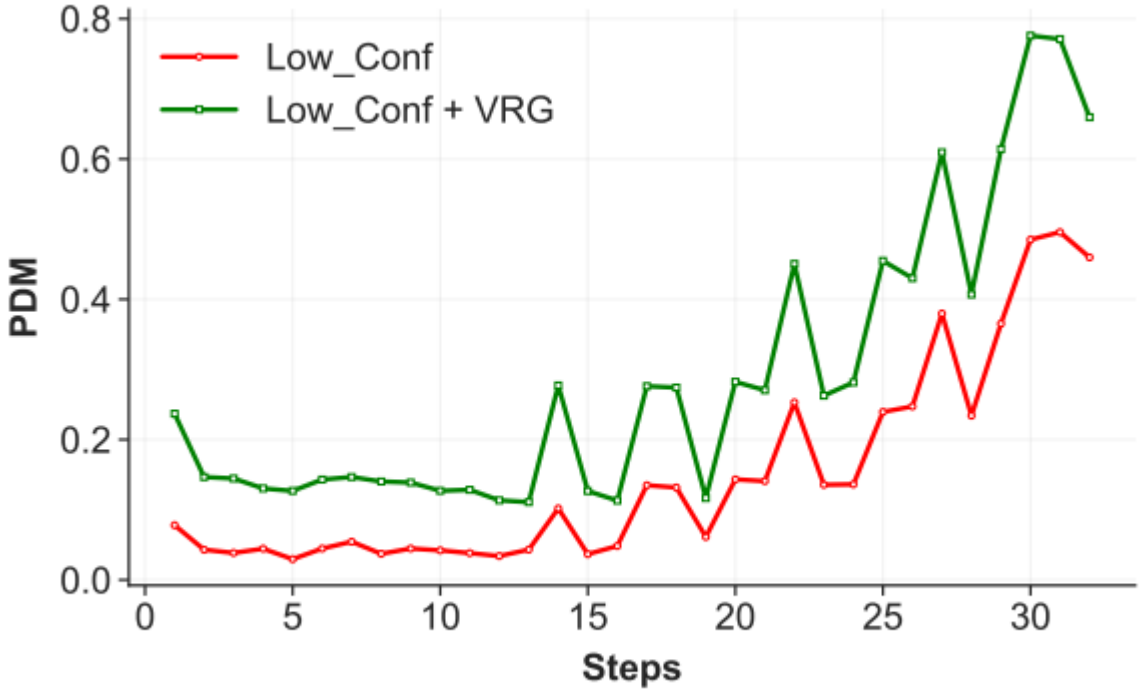
| Method | $M^3$CoT | MMBench | SQA-IMG | V* Bench |
|---|---|---|---|---|
| Low-conf | 45.8 | 72.8 | 71.0 | 42.9 |
| Low-conf w/ L2R | 47.2 | 74.0 | 71.1 | 43.4 |
| Low-conf w/ PSP | 47.6 | 74.3 | 72.0 | 44.5 |
| Low-conf w/ L2R & VRG | 47.6 | 74.6 | 71.3 | 44.5 |
| Low-conf w/ PSP & VRG | **48.4**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

# Analysis

## Analysis of PSP & VRG




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




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

# Conclusion

CVPR JUNE 3-7, 2026 DENVER COLORADO

## Contribution

I. We provide an **empirical analysis** showing that **dMLLMs often generate final answers too early and rely weakly on visual prompts in early diffusion steps**.

II. We introduce **PSP**, which controls the diffusion reasoning process by delaying early answer generation through **position- and timestep-aware confidence penalties**.

III. We introduce **VRG**, which **strengthens visual-grounded reasoning** by amplifying the difference between visual-conditioned and text-only predictions during inference.

CVPR JUNE 3-7, 2026 DENVER COLORADO

# Thank you for watching :D

If you have any questions, feel free to contact us at

ktkpv94@hanyang.ac.kr

# Appendix

**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.

# Appendix



**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

# Appendix



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** |