HANYANG UNIVERSITY

ICML
International Conference
On Machine Learning

# ST-Veto:
## Spatio-Temporal Token Veto for Diffusion MLLMs via Taylor Prediction and Visual Grounding

**Keuntae Kim[1]*, Beomseok Lee[2]*, Hyunwoo Kim[3]*, Yong Suk Choi[1]†**

1)Hanyang University, 2)LG Electronics, 3)KT Corporation

(*: Equal contribution, †: Corresponding author)

HYU Artificial Intelligence Laboratory

Contact us: ktkpv94@hanyang.ac.kr

# Introduction

ICML
International Conference
On Machine Learning

## Background

- Diffusion MLLMs generate text through iterative masked-token denoising rather than left-to-right autoregressive decoding.
- By predicting multiple token positions in parallel, they provide **an efficient and flexible generation paradigm with potential for iterative refinement and self-correction.**

## Motivation

- Despite their efficiency, dMLLMs struggle with multimodal reasoning, where AR-style reasoning methods do not naturally transfer to parallel diffusion decoding.
- Moreover, **confidence-based unmasking can prematurely commit temporally unstable or weakly grounded tokens**, motivating **a diffusion-native strategy based on temporal stability and visual grounding**.

# Introduction

ICML
International Conference
On Machine Learning

## Overview

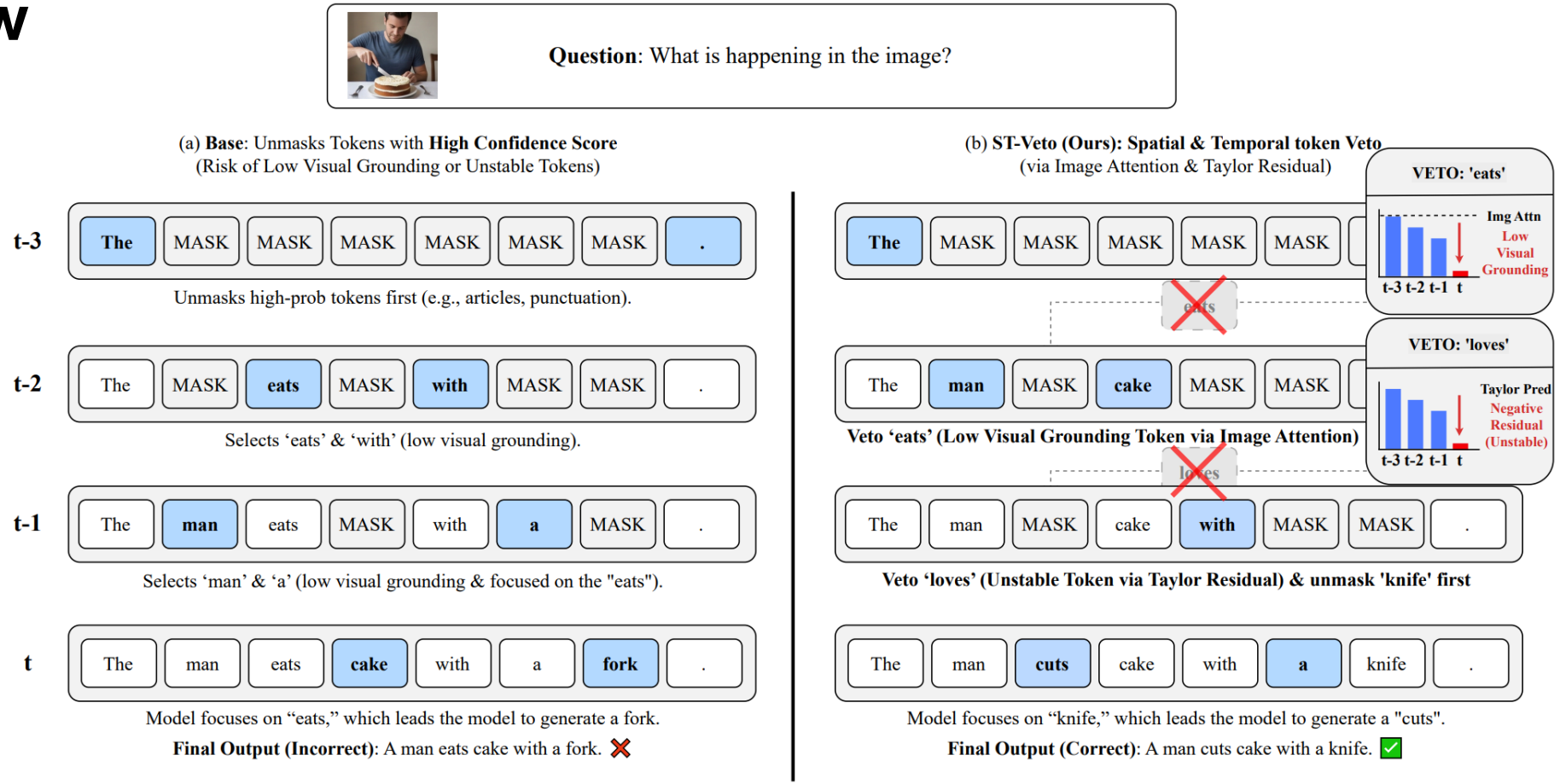



**Figure 1.** Overview of our method and comparison with the baseline

# Method – ST-Veto

ICML
International Conference
On Machine Learning

## Method Overview

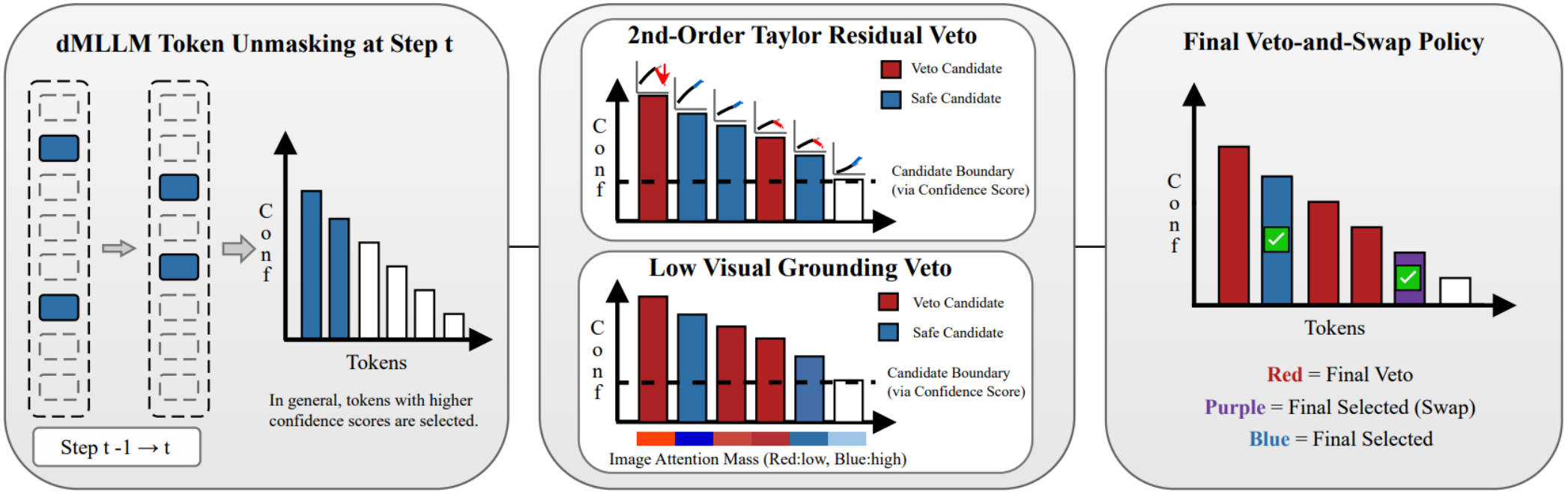
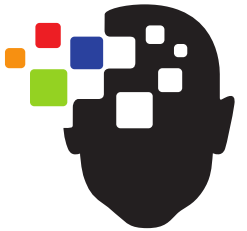



*Figure 2.* Overview of ST-Veto. Tokens are vetoed based on the Taylor residual and low visual grounding, and are replaced by safe tokens that are unmasked instead.

# Method – ST-Veto



ICML
International Conference
On Machine Learning

## Step 1: Top-k Token Proposal

$$T_k = \left\{ i \mid \text{rank}_{\downarrow} \left( c_i^{(t)} \right) \leq k \right\}$$

- At each diffusion step, **the model first proposes the top-k tokens with the highest current confidence**.

# Method – ST-Veto

ICML
International Conference
On Machine Learning

## Step 2: Taylor Residual (Temporal Instability Signal)

$$\hat{c}_i^{(t)} = c_i^{(t-1)} + \left(c_i^{(t-1)} - c_i^{(t-2)}\right) + \frac{1}{2}\left[\left(c_i^{(t-1)} - c_i^{(t-2)}\right) - \left(c_i^{(t-2)} - c_i^{(t-3)}\right)\right]$$

$$e_i = c_i^{(t)} - \hat{c}_i^{(t)}$$

- ST-Veto **predicts the expected confidence at the current step using a second-order Taylor approximation**.

  A large negative residual indicates that the token is temporally unstable.

# Method – ST-Veto

ICML
International Conference
On Machine Learning

## Step 3: Image Attention (Grounding Signal)

$$\mathcal{L}_{\text{attn}} = \left\{ \frac{N_{\text{lay}}}{3}, \frac{N_{\text{lay}}}{2}, \frac{2N_{\text{lay}}}{3} \right\} \qquad \rho_i^{(\ell)} = \frac{\sum_{j \in \mathcal{I}} A_{i,j}^{(\ell)}}{\sum_{p=1}^{L_{\text{seq}}} A_{i,p}^{(\ell)}} \qquad m_i = \frac{1}{|\mathcal{L}_{\text{attn}}|} \sum_{\ell \in \mathcal{L}_{\text{attn}}} \rho_i^{(\ell)}$$

- This measures **how much attention token i assigns to image tokens**.
- **A low image-attention score indicates weak visual grounding.**

# Method – ST-Veto

ICML
International Conference
On Machine Learning

## Step 4: Veto-and-Swap Policy

$$\mathcal{C}_{\text{veto}}(i) \iff (e_i < \mu_e - \sigma_e) \vee (m_i < \mu_m - \sigma_m)$$

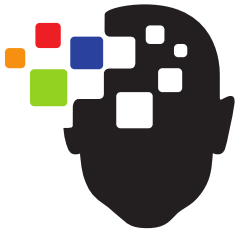
$$\mathcal{C}_{\text{safe}}(i) \iff (e_i \geq \mu_e + \sigma_e) \wedge (m_i \geq \mu_m + \sigma_m)$$

$$\mu_X = \text{median}(X)$$

$$\sigma_X = \text{MAD}(X)$$

- ST-Veto vetoes tokens that are **either temporally unstable or weakly grounded**, then swaps them with safe near-boundary candidates.
- ST-Veto uses **median and MAD to define adaptive thresholds for each signal**. This provides robust **token selection by reducing sensitivity to outliers and avoiding hand-crafted fixed thresholds**.

# Experiment



ICML
International Conference
On Machine Learning

## 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. Use greedy decoding without temperature scaling

# Results

ICML
International Conference
On Machine Learning

## Main Results

| Model | Method | $M^3$CoT | | MMBench | | SQA-IMG | | V* Bench | |
|---|---|---|---|---|---|---|---|---|---|
| | | 128/64 | 256/128 | 128/64 | 256/128 | 128/64 | 256/128 | 128/64 | 256/128 |
| *LaViDa* | Base-Confidence | 27.96 | 29.55 | 33.96 | 31.55 | 45.81 | 40.06 | 35.60 | 37.17 |
| | Base-Margin | 27.88 | 28.22 | 32.99 | 31.68 | 46.01 | 38.99 | 35.08 | 37.70 |
| | Base-Entropy | 27.91 | 29.23 | 32.67 | 31.92 | 45.89 | 39.17 | 34.55 | 36.65 |
| | DDCoT | 26.99 | 29.11 | 33.02 | 30.98 | 45.99 | 38.81 | 35.08 | 35.60 |
| | CCoT | 26.57 | 28.98 | 33.19 | 31.04 | 45.21 | 38.90 | 36.65 | 36.65 |
| | SCAFFOLD | 26.61 | 28.81 | 33.24 | 31.22 | 44.98 | 39.25 | 35.60 | 35.08 |
| | ICoT | 27.82 | 29.29 | 33.22 | 31.43 | 45.68 | 39.17 | 37.17 | 37.17 |
| | **ST-Veto (Ours)** | **30.21** | **31.87** | **38.05** | **38.25** | **47.80** | **48.54** | **42.94** | **41.36** |
| *MMaDA* | Base-Confidence | 26.32 | 25.11 | 33.75 | 32.04 | 44.72 | 41.70 | 34.03 | 31.41 |
| | Base-Margin | 26.22 | 25.54 | 33.22 | 31.25 | 44.99 | 41.55 | 32.98 | 32.46 |
| | Base-Entropy | 26.36 | 25.32 | 33.53 | 31.87 | 44.23 | 41.98 | 33.51 | 31.41 |
| | DDCoT | 25.44 | 24.45 | 33.21 | 30.04 | 43.99 | 39.98 | 33.51 | 29.84 |
| | CCoT | 26.10 | 24.88 | 34.01 | 31.55 | 44.81 | 40.12 | 34.55 | 29.32 |
| | SCAFFOLD | 25.88 | 24.91 | 33.88 | 30.98 | 44.92 | 40.33 | 31.41 | 28.80 |
| | ICoT | 26.01 | 24.78 | 33.93 | 32.44 | 44.52 | 40.47 | 34.55 | 31.41 |
| | **ST-Veto (Ours)** | **27.52** | **27.74** | **36.55** | **36.74** | **47.25** | **45.31** | **41.36** | **37.17** |

# Results

ICML
International Conference
On Machine Learning

## Ablation Study

| Model | Setting | $M^3$CoT | | MMBench | | SQA-IMG | | V* Bench | |
|---|---|---|---|---|---|---|---|---|---|
| | | 128/64 | 256/128 | 128/64 | 256/128 | 128/64 | 256/128 | 128/64 | 256/128 |
| *LaViDa* | Base | 27.96 | 29.55 | 33.96 | 31.55 | 45.81 | 40.06 | 35.60 | 37.17 |
| | w/o Taylor | 28.77 | 29.64 | 36.04 | 35.94 | 45.86 | 46.52 | 38.22 | 39.27 |
| | w/o Img | 29.68 | 29.11 | 36.22 | 35.22 | 47.47 | 47.89 | 41.36 | 38.22 |
| | **ST-Veto** | **30.21** | **31.87** | **38.05** | **38.25** | **47.80** | **48.54** | **42.94** | **41.36** |
| *MMaDA* | Base | 26.32 | 25.11 | 33.75 | 32.04 | 44.72 | 41.70 | 34.03 | 31.41 |
| | w/o Taylor | 26.66 | 25.99 | 35.12 | 35.52 | 47.15 | 43.99 | 39.79 | 32.84 |
| | w/o Img | 26.88 | 26.24 | 34.98 | 35.10 | 45.56 | 43.87 | 37.17 | 33.51 |
| | **ST-Veto** | **27.52** | **27.74** | **35.55** | **35.74** | **47.25** | **45.31** | **41.36** | **37.17** |

# Results

ICML
International Conference
On Machine Learning

## Ablation Study

*Table 5.* Layer ablation for the image-attention grounding signal on LaViDa. Combining intermediate depths is more robust than using a single layer.

| Dataset | $L/T$ | Base | $L/3$ | $L/2$ | $2L/3$ | All |
|---|---|---|---|---|---|---|
| MMBench | 128/64 | 33.96 | 35.83 | 37.12 | 37.68 | **38.05** |
| | 256/128 | 31.55 | 34.21 | 36.08 | 35.74 | **38.25** |
| SQA-IMG | 128/64 | 45.81 | 46.95 | 47.40 | 46.85 | **47.80** |
| | 256/128 | 40.06 | 44.52 | 46.71 | 47.85 | **48.54** |
| $M^3$CoT | 128/64 | 27.96 | 29.12 | 29.85 | **30.37** | 30.21 |
| | 256/128 | 29.55 | 30.67 | 31.23 | 30.85 | **31.87** |
| V* Bench | 128/64 | 35.60 | 38.74 | 40.31 | 41.36 | **42.94** |
| | 256/128 | 37.17 | 39.27 | 40.31 | **41.88** | 41.36 |

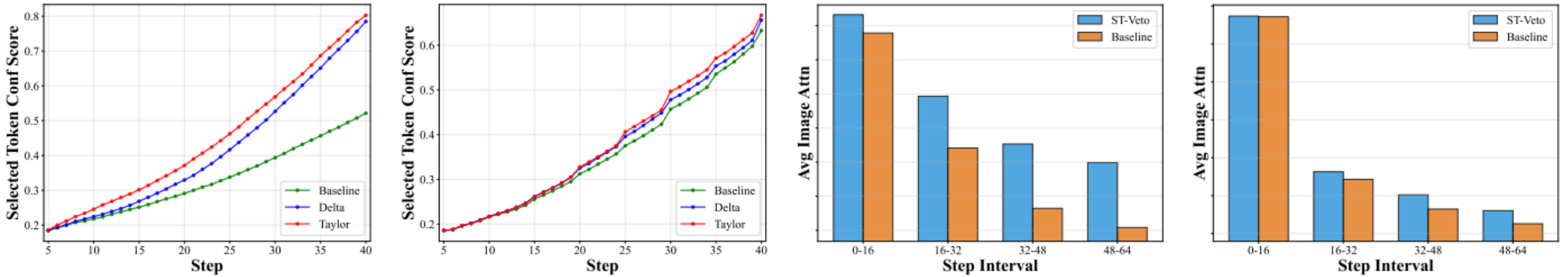
*Table 6.* Sensitivity to the robust threshold on LaViDa. The default MAD cutoff is consistently strong across benchmarks and generation settings.

| Dataset | $L/T$ | Base | 0.5MAD | MAD | 1.5MAD |
|---|---|---|---|---|---|
| MMBench | 128/64 | 33.96 | 36.62 | **38.05** | 37.24 |
| | 256/128 | 31.55 | 36.45 | **38.25** | 36.12 |
| SQA-IMG | 128/64 | 45.81 | 47.12 | **47.80** | **47.80** |
| | 256/128 | 40.06 | 47.45 | **48.54** | 47.22 |
| $M^3$CoT | 128/64 | 27.96 | 29.96 | **30.21** | 28.88 |
| | 256/128 | 29.55 | 31.22 | **31.87** | 31.22 |
| V* Bench | 128/64 | 35.60 | 39.27 | **42.94** | 41.36 |
| | 256/128 | 37.17 | 40.61 | **41.36** | 38.22 |

# Analysis

ICML
International Conference
On Machine Learning

## Analysis of ST-Veto




(a) LaViDa on MMbench (b) MMaDA on MMbench (c) LaViDa on MMbench (d) MMaDA on MMbench

*Figure 3.* (a) and (b) show the confidence scores of tokens selected at each step for different methods. (c) and (d) present the relative values of the Average Image Attention.

# Analysis

ICML
International Conference
On Machine Learning

## Analysis of ST-Veto

*Table 2.* Time cost table. The unit is inference per second, and the values are averaged over runs on MMbench.

| Method | MMaDA | | LaViDa | |
|---|---|---|---|---|
| | 128/64 | 256/128 | 128/64 | 256/128 |
| Base-Confidence | 5.68 | 11.32 | 4.87 | 12.88 |
| Base-Margin | 5.65 | 11.28 | 4.85 | 12.84 |
| Base-Entropy | 5.68 | 11.33 | 4.88 | 12.89 |
| DDCoT | 10.87 | 19.75 | 9.89 | 22.95 |
| CCoT | 10.93 | 19.48 | 9.92 | 22.94 |
| SCAFFOLD | 10.99 | 19.43 | 9.85 | 22.91 |
| ICoT | 10.96 | 19.89 | 9.87 | 22.87 |
| ST-Veto | 5.73 | 11.39 | 4.89 | 12.92 |

*Table 4.* Sensitivity to history length on LaViDa. $H = 4$ provides the best cost-performance trade-off; $H = 5$ occasionally matches or exceeds it but requires a higher-order finite-difference estimate.

| Dataset | $L/T$ | Base | Delta | 1st | 2nd | 3rd |
|---|---|---|---|---|---|---|
| MMBench | 128/64 | 33.96 | 36.98 | 37.38 | 38.05 | **38.11** |
| | 256/128 | 31.55 | 36.89 | 37.98 | **38.25** | 38.12 |
| SQA-IMG | 128/64 | 45.81 | 46.69 | 47.25 | 47.80 | **47.91** |
| | 256/128 | 40.06 | 46.36 | 47.60 | **48.54** | 47.55 |
| $M^3$CoT | 128/64 | 27.96 | 28.69 | 29.21 | **30.21** | 30.18 |
| | 256/128 | 29.55 | 30.03 | 30.50 | **31.87** | 31.55 |
| V* Bench | 128/64 | 35.60 | 37.17 | 41.36 | **42.94** | 39.27 |
| | 256/128 | 37.17 | 39.27 | 39.27 | **41.36** | **41.36** |

# Analysis

ICML
International Conference
On Machine Learning

## Analysis of ST-Veto

*Table 7.* Analysis Table. Metrics are reported for *top-k* (Veto & Swap) and *candidate* (Veto / Safe; in percent with *Veto*+*Safe*=100) under each configuration (128/64 and 256/128). Candidate ratios for each model/setting are adjusted to be close to the *V* Bench* ratio.

| Model | Benchmark | 128/64 | | | 256/128 | | |
|---|---|---|---|---|---|---|---|
| | | top-k | candidate | | top-k | candidate | |
| | | Veto & Swap | Veto | Safe | Veto & Swap | Veto | Safe |
| *LaViDa* | $M^3$CoT | 4.8±1.6 | 18.2±5.2 | 81.8±5.2 | 2.6±0.9 | 19.3±3.9 | 80.7±3.9 |
| | MMBench | 5.3±2.2 | 19.0±6.4 | 81.0±6.4 | 3.1±1.3 | 20.1±3.2 | 79.9±3.2 |
| | SQA-IMG | 5.1±2.1 | 18.7±5.6 | 81.3±5.6 | 2.9±1.2 | 19.8±3.1 | 80.2±3.1 |
| | V* Bench | 4.8±6.1 | 18.4±5.9 | 81.6±5.9 | 2.2±3.0 | 19.5±2.9 | 80.5±2.9 |
| *MMaDA* | $M^3$CoT | 5.8±1.3 | 14.5±3.0 | 85.5±3.0 | 7.4±1.8 | 11.8±3.5 | 88.2±3.5 |
| | MMBench | 5.2±1.0 | 14.2±2.4 | 85.8±2.4 | 6.9±1.5 | 11.5±2.7 | 88.5±2.7 |
| | SQA-IMG | 3.9±0.8 | 13.6±2.3 | 86.4±2.3 | 5.5±1.2 | 10.9±3.1 | 89.1±3.1 |
| | V* Bench | 6.1±2.9 | 14.7±2.0 | 85.3±2.0 | 6.3±3.7 | 11.2±2.8 | 88.8±2.8 |

# Analysis

ICML
International Conference
On Machine Learning

## Analysis of ST-Veto

*Table 14.* Representative fine-grained MMBench category results.

| Category | *N* | Base | ST-Veto | Diff. |
|---|---|---|---|---|
| Structuralized image-text understanding | 282 | 36.88 | **57.09** | +20.21 |
| Physical relation | 94 | 34.04 | **48.94** | +14.89 |
| Image topic | 140 | 31.43 | **44.29** | +12.86 |
| Action recognition | 215 | 28.37 | **40.47** | +12.09 |
| Attribute comparison | 141 | 32.62 | **44.68** | +12.06 |
| Celebrity recognition | 396 | 25.51 | **27.02** | +1.52 |
| Spatial relationship | 177 | 28.25 | **28.81** | +0.56 |
| Attribute recognition | 264 | **35.61** | 35.23 | -0.38 |
| Image emotion | 200 | **40.50** | 40.00 | -0.50 |
| Identity reasoning | 176 | **26.14** | 25.00 | -1.14 |

# Conclusion

ICML
International Conference
On Machine Learning

## Contribution

I. We introduce ST-Veto, a training-free decoding strategy that uses **temporal confidence dynamics** to identify and veto unstable token commitments during diffusion generation.

II. We further incorporate **visual grounding through image-attention signals**, allowing dMLLMs to prioritize tokens that are both temporally stable and visually supported.

ICML
International Conference
On Machine Learning

# Thank you for watching :D

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

ktkpv94@hanyang.ac.kr

HYU Artificial Intelligence Laboratory