Which Tokens to Merge?

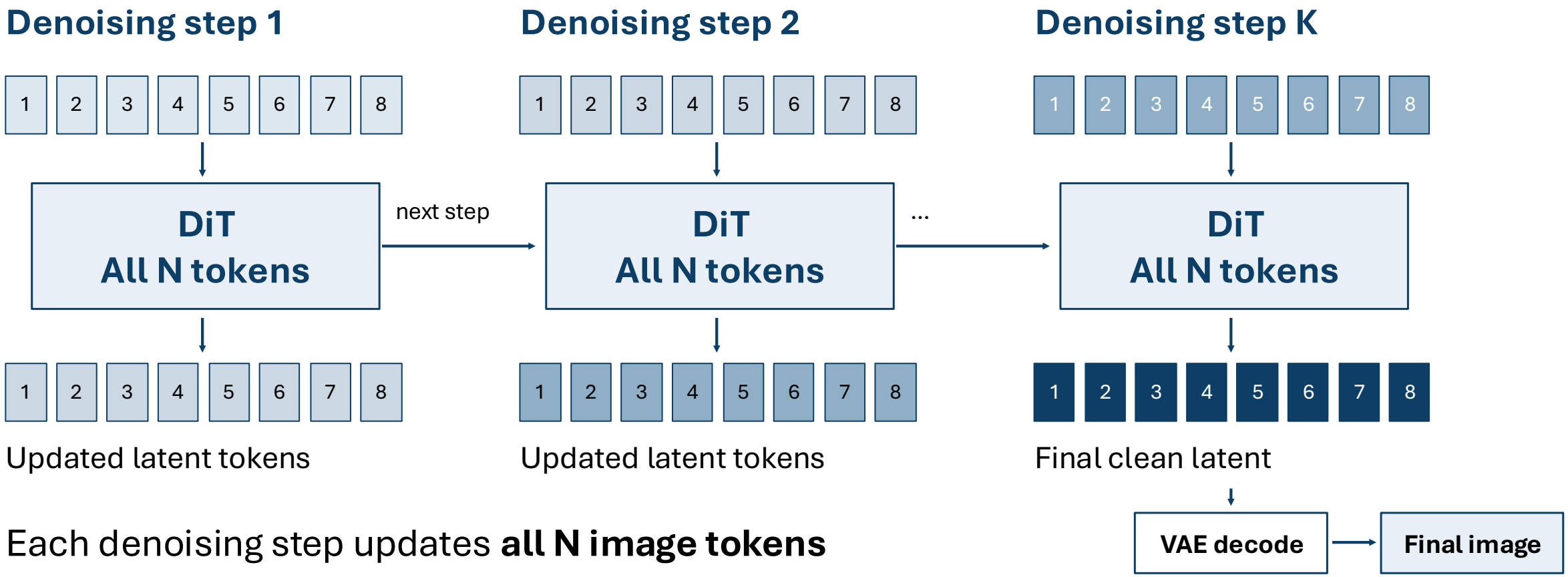
Diffusion Dynamics for Efficient Image Generation

SeungJu Cha, Ye-Chan Kim, HyunGee Kim, Sungho Koh, Dong-Jin Kim

Hanyang University, South Korea

github.com/SeungJuCha/DaTMe

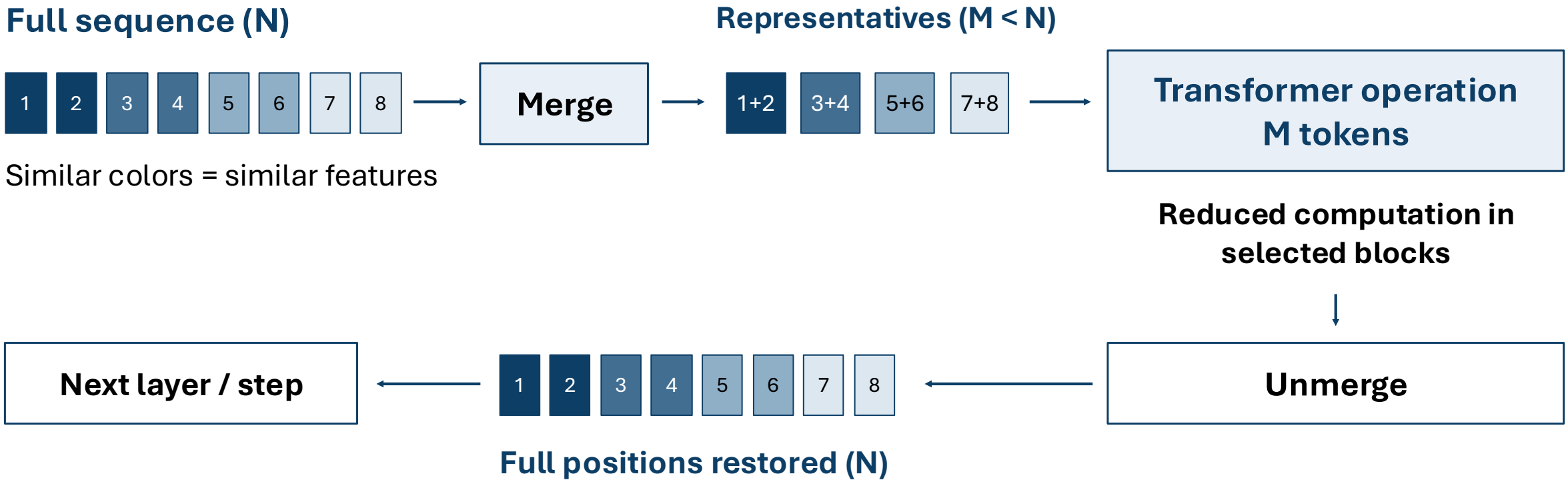
DiT repeatedly processes the full token sequence



Each denoising step updates **all N image tokens**

What if similar tokens are computed once?

Merge similar features, compute representatives, then **unmerge**



Which tokens should **merge**? Which should **remain distinct**?

Representative token merging methods

ToMeSD[1]

Feature similarity identifies redundant tokens
Merge for selected computation, then restore token positions

Importance-Based Token Merging[2]

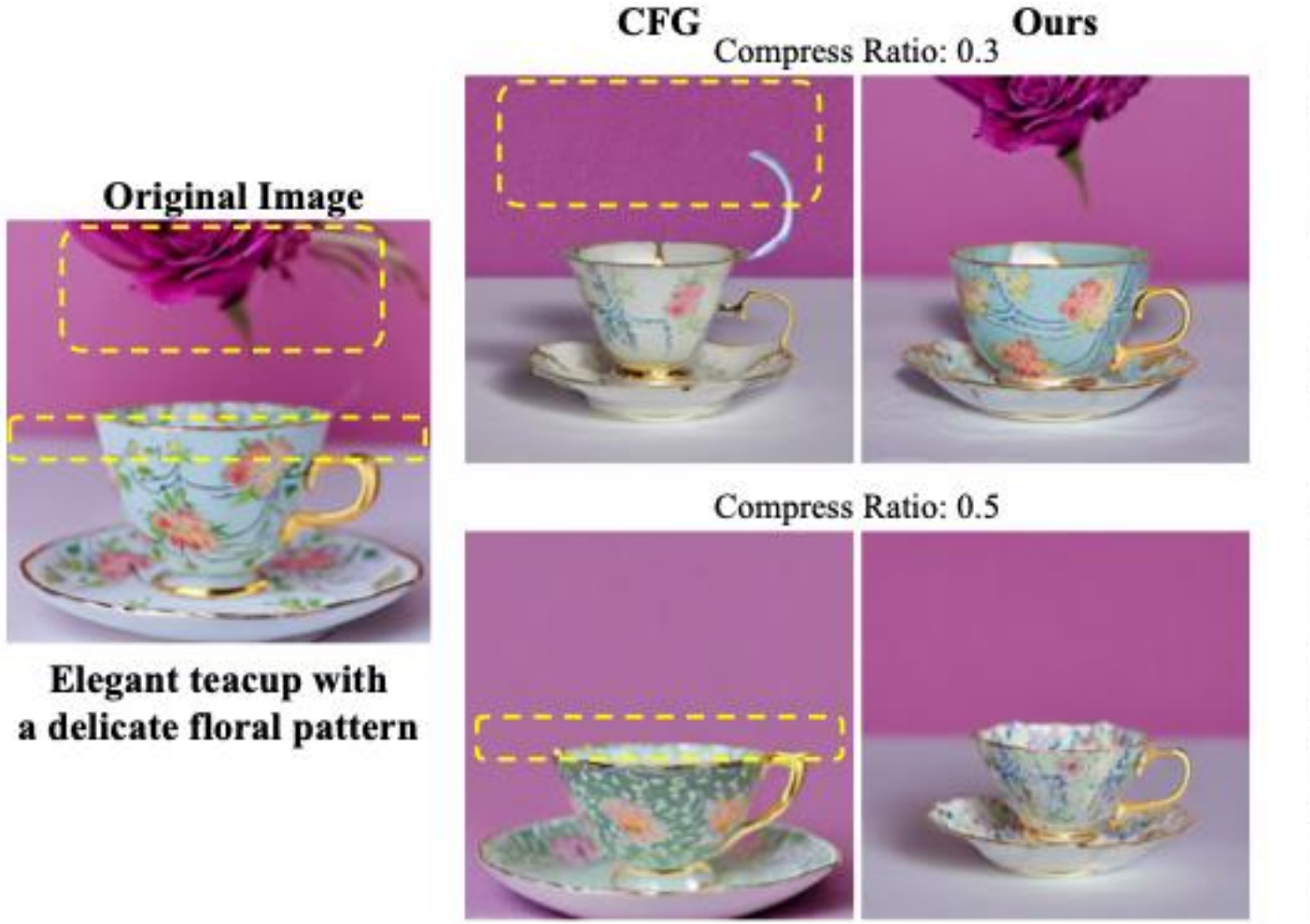
CFG discrepancy provides token importance
Important-token pool guides anchor and source selection

$$\epsilon_\theta\,(x_t, t) = \epsilon_\theta\,(x_t, t) + w[\epsilon_\theta\,(x_t, c, t) - \epsilon_\theta\,(x_t, t)]$$

Does **prompt relevance** capture all image structure?

[1] Bolya, Daniel, and Judy Hoffman. "Token merging for fast stable diffusion." *2023 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW)*. IEEE, 2023.
[2] Wu, Haoyu, et al. "Importance-based token merging for efficient image and video generation." *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2025.

CFG signal can miss background structure



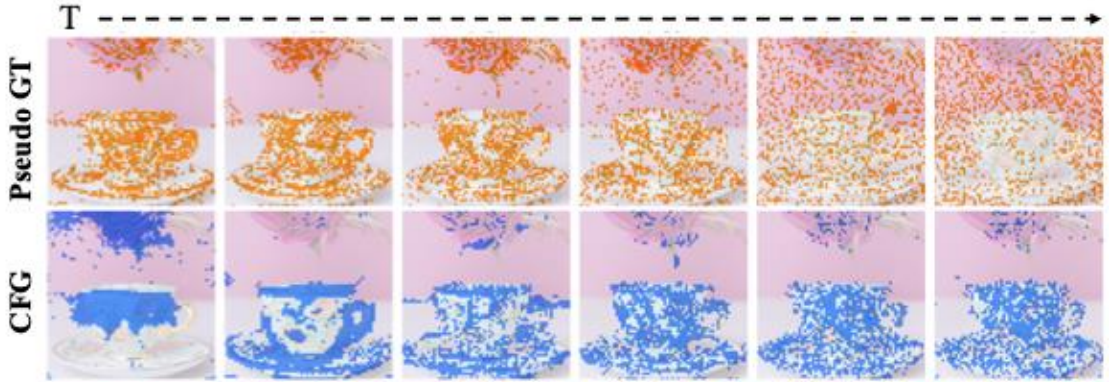
Prompt relevance

Strong guidance around prompt-aligned objects

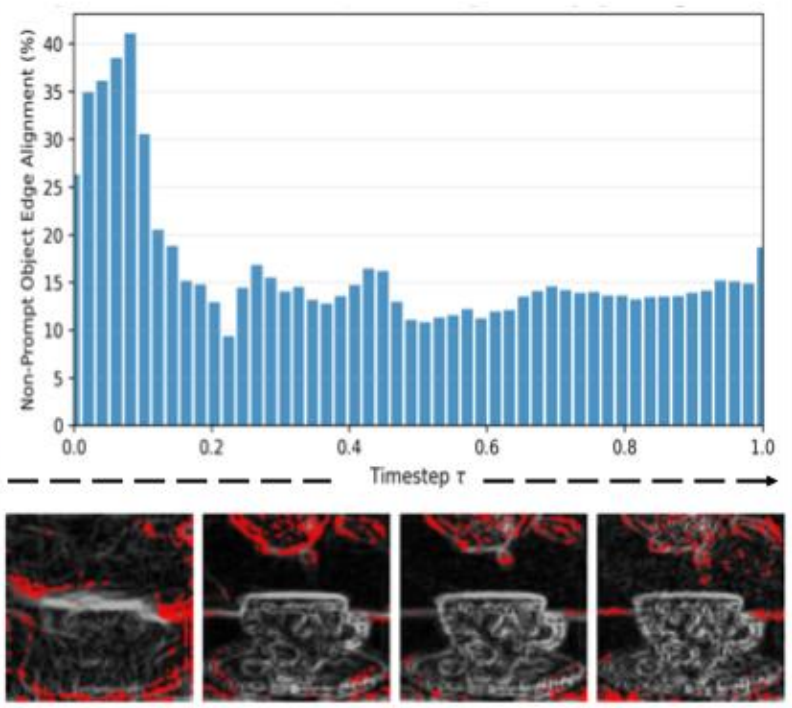
What about the background?

Unmentioned boundaries and fine details

Prompt relevance leaves structural blind spots



(b) Spatial distribution of top-ranked tokens



(c) Blind spots at non-prompt region

Time-varying importance

Refinement shifts beyond the foreground

Prompt-independent structure

Background tokens remain low-ranked in CFG

Prompt relevance leaves structural blind spots

(b) Spatial distribution of top-ranked tokens (c) Blind spots at non-prompt region

How to faithfully capture these diffusion dynamics?

Propose DaTMe

Two signals from the predicted noise

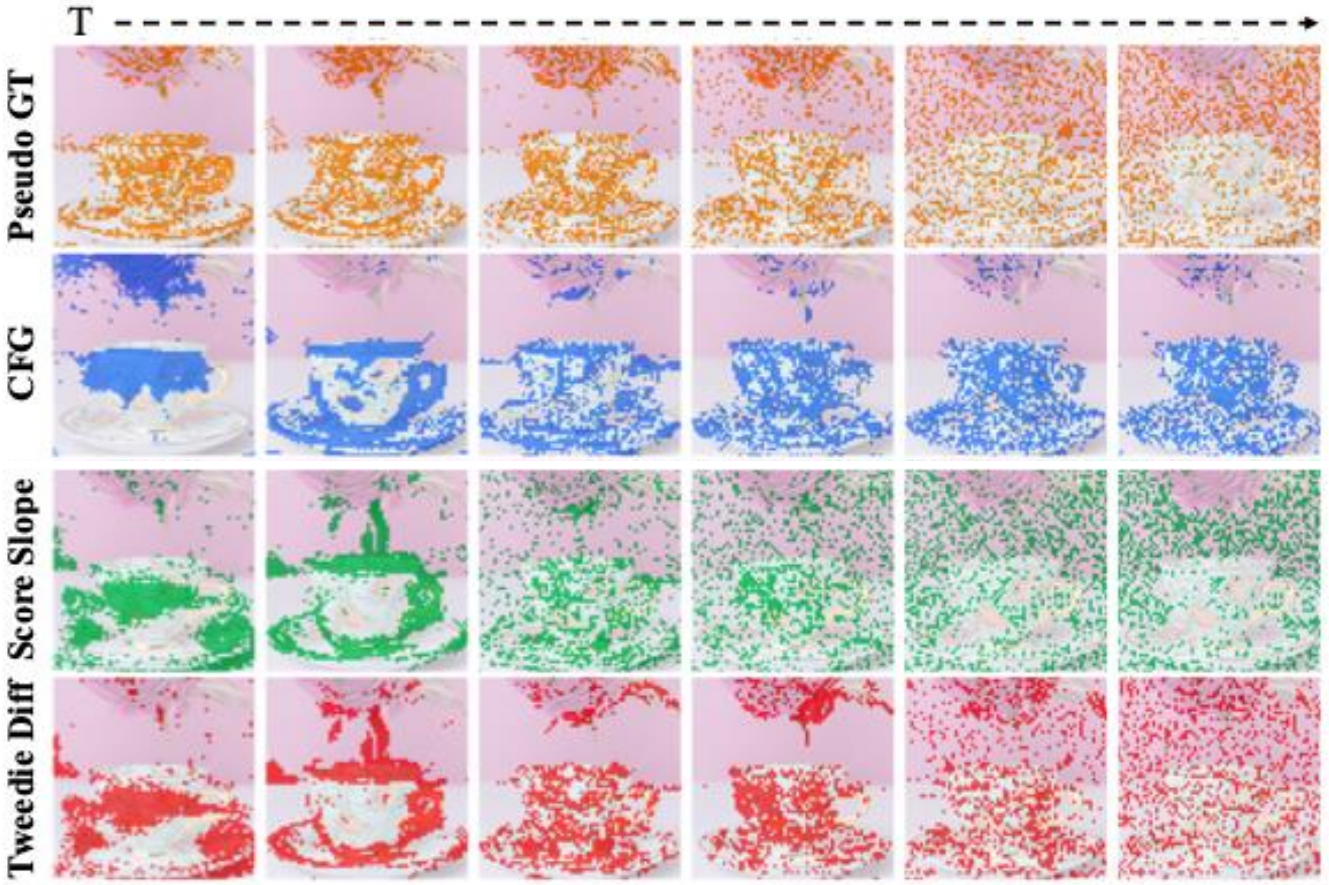
Score function

$$s(t) = \nabla_{x_t} log p_t(x_t) \approx -\frac{1}{\sigma_t}\epsilon_\theta(x_t, t)$$

Tweedie estimation

$$\hat{x}_0(t) = \mathbb{E}[x_0|x_t] \approx \frac{x_t - \sigma_t \epsilon_t(x_t, t)}{\alpha_t}$$

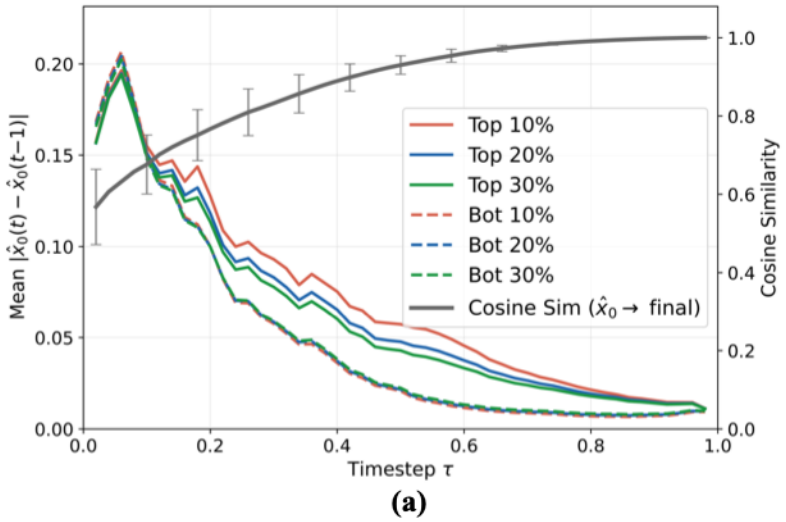
Which is better?

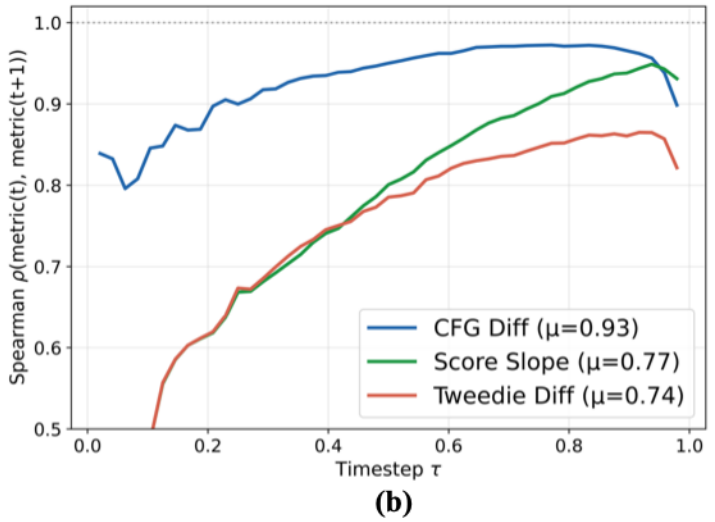


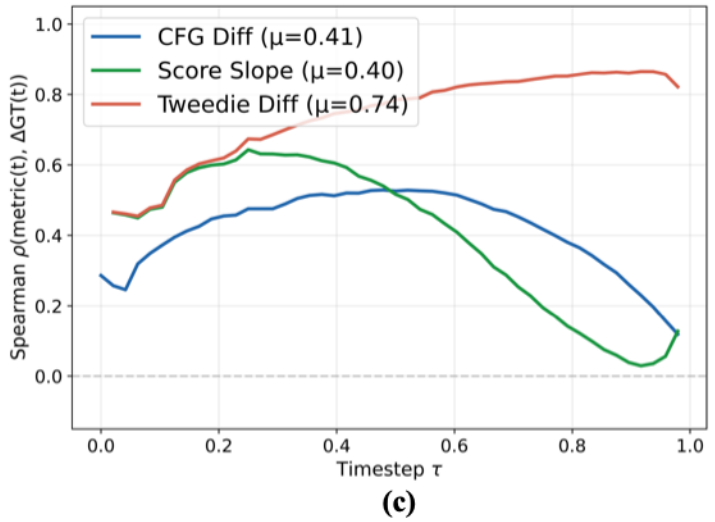
Tweedie difference instead of score difference

Pseudo-GT = $|x_0 - \hat{x}_0(t)|$

Reference: **remaining change** to the final generated latent







(a) Different refinement

High-gap tokens keep changing

(b) Dynamic importance

CFG remains relatively static
Focus on the same region

(c) Pseudo-GT alignment

Tweedie difference stays aligned

Tweedie difference instead of score difference

Raw score slope $|s(t) - s(t-1)|$

$$\left|\frac{\epsilon_\theta(t-1) - \epsilon_\theta(t)}{\sigma_{t-1}}\right| + \epsilon_\theta(t)\left|\left(\frac{1}{\sigma_{t-1}} - \frac{1}{\sigma_t}\right)\right|$$

Difference term
Prediction change

Scale term
Noise-schedule rescaling

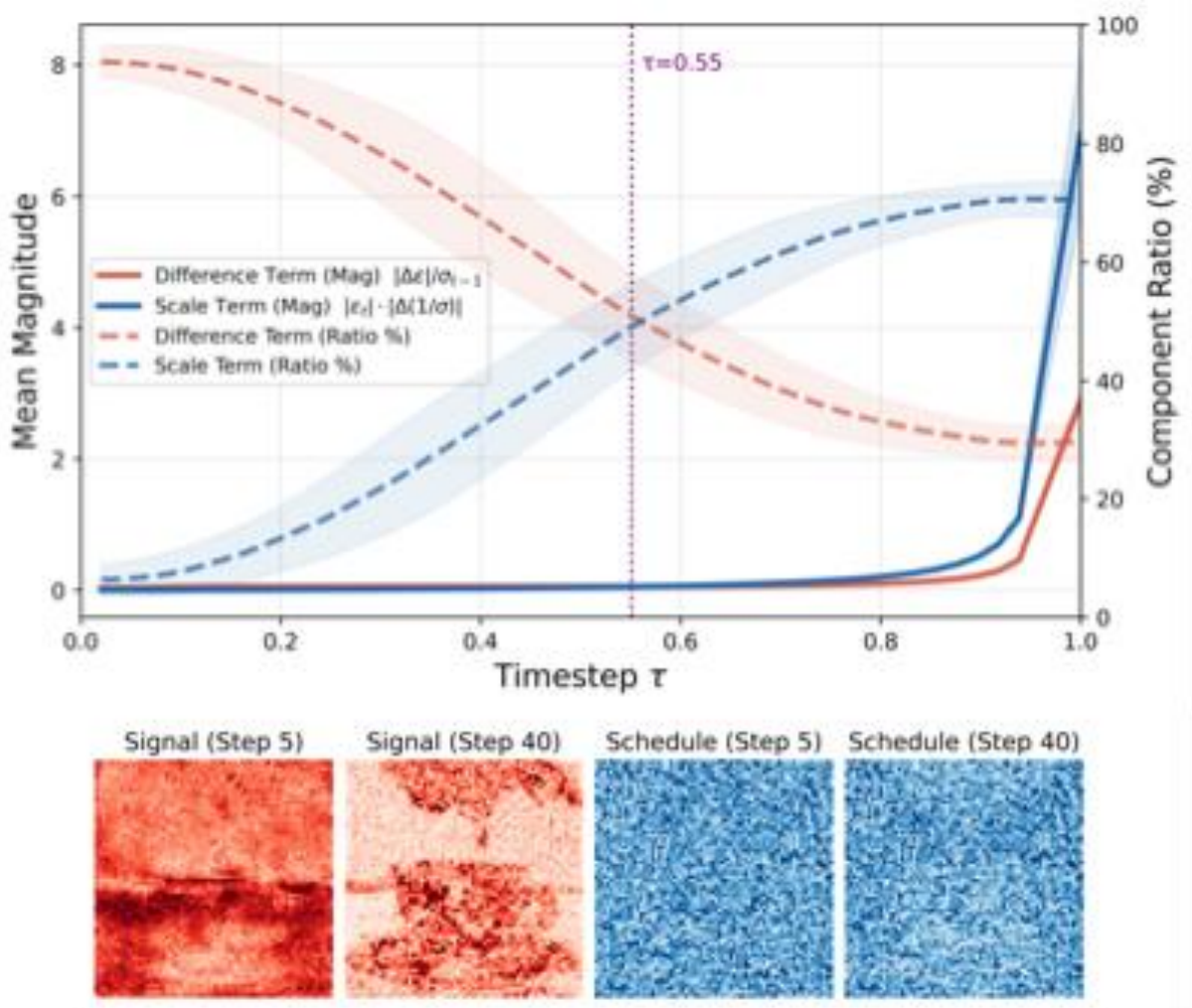
Late steps: **scale term can dominate**

$\sigma_t \to 0$: schedule effects can mask **token refinement**

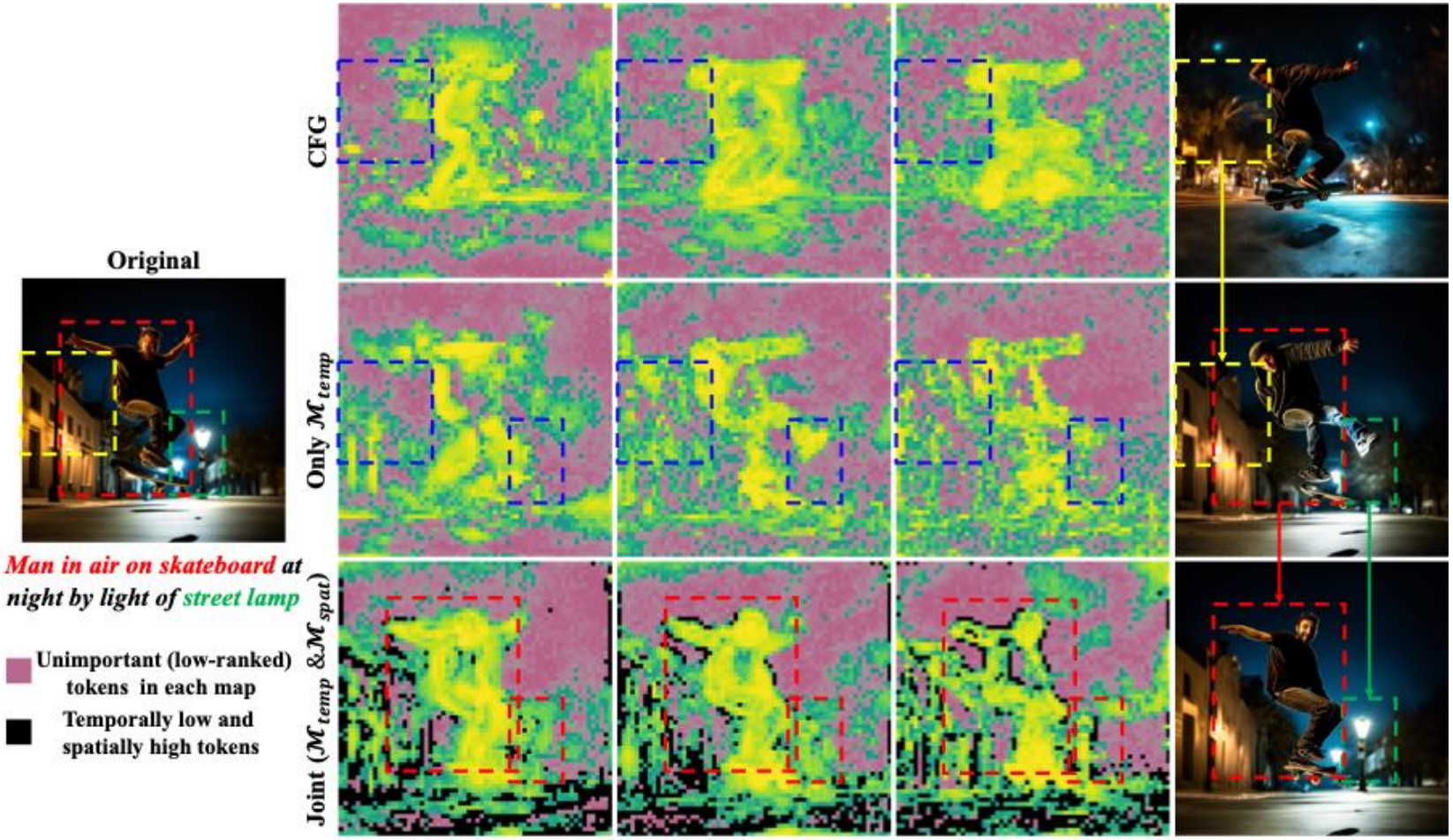
Tweedie difference $|\hat{x}_0(t) - \hat{x}_0(t-1)|$

$$\mathcal{M}_{temp}^{(t)} = \frac{\sigma_{t-1}}{\alpha_{t-1}}|\epsilon_\theta(t-1) - \epsilon_\theta(t)| \quad (6)$$

Scale term removed, only focus on token refinement

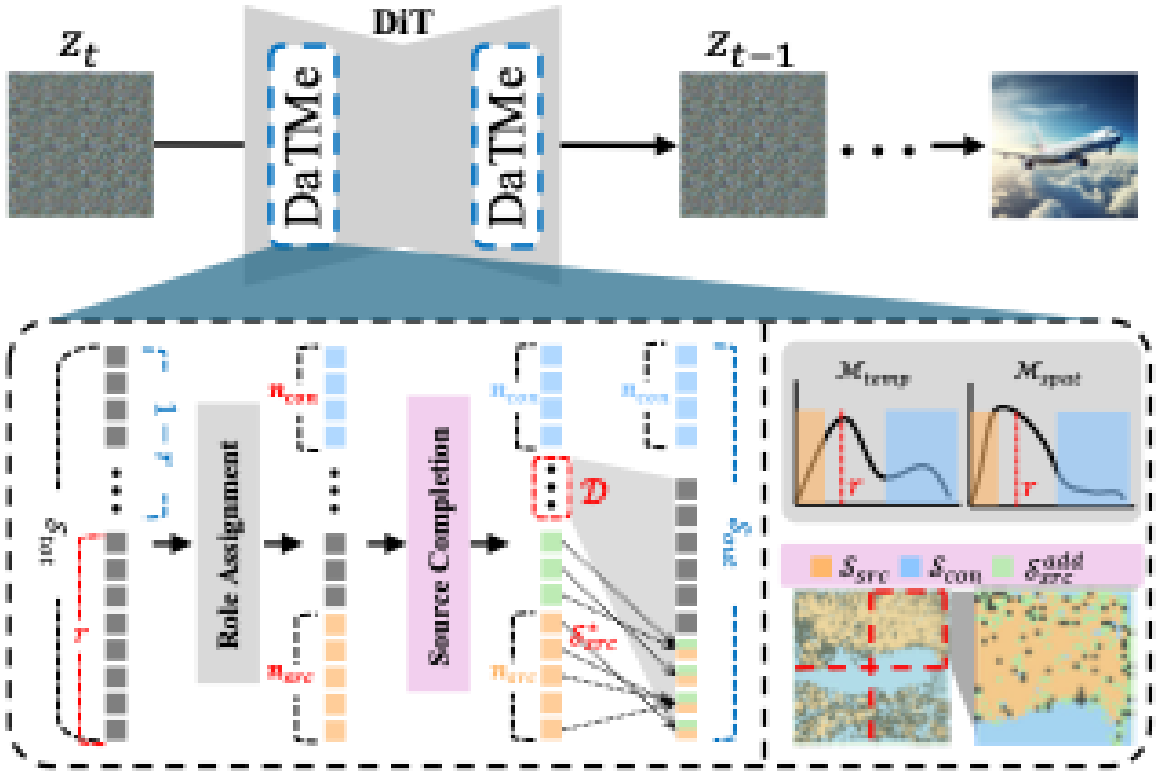


Stable tokens can still carry important structure



$$\mathcal{M}_{spat}^{(t)} = |\hat{x}_0(t) - UP_s(Down_s(\hat{x}_0(t)))|$$

DaTMe



Algorithm 1 Dynamics-Aware Token Merging (DaTMe) at timestep t

Require: token features $\mathbf{F}^{(t)}$, temporal map $\mathcal{M}_{\text{temp}}^{(t)}$, spatial map $\mathcal{M}_{\text{spat}}^{(t)}$, compression ratio r, number of grids P

Ensure: merged token features $\mathbf{F}'^{(t)}$

$N \leftarrow H \times W, \quad \mathcal{S}_{\text{tot}} \leftarrow \{1, \ldots, N\}, \quad n_{\text{tgt}} \leftarrow \lfloor rN \rfloor$

// 1. Dynamics-Aware Role Assignment

$\mathcal{S}_{\text{con}} \leftarrow \text{TopFrac}_{1-r}\left(\mathcal{M}_{\text{temp}}^{(t)}\right) \cap \text{TopFrac}_{1-r}\left(\mathcal{M}_{\text{spat}}^{(t)}\right)$ ▷ protected tokens

$\mathcal{S}_{\text{src}} \leftarrow \text{BotFrac}_{r}\left(\mathcal{M}_{\text{temp}}^{(t)}\right) \cap \text{BotFrac}_{r}\left(\mathcal{M}_{\text{spat}}^{(t)}\right)$ ▷ safe merge sources

// 2. Spatially Distributed Source Completion

$\mathcal{U} \leftarrow \mathcal{S}_{\text{tot}} \setminus \left(\mathcal{S}_{\text{con}} \cup \mathcal{S}_{\text{src}}\right)$

$\mathcal{S}_{\text{src}}^{\text{add}} \leftarrow \text{GridSample}\left(\mathcal{U}, n_{\text{tgt}} - |\mathcal{S}_{\text{src}}|, P\right)$

$\mathcal{S}_{\text{src}}^{*} \leftarrow \mathcal{S}_{\text{src}} \cup \mathcal{S}_{\text{src}}^{\text{add}}$

// 3. Locality-Constrained Spatial Matching

$\mathcal{D} \leftarrow \mathcal{S}_{\text{tot}} \setminus \left(\mathcal{S}_{\text{con}} \cup \mathcal{S}_{\text{src}}^{*}\right)$ ▷ destination tokens

for each $j \in \mathcal{S}_{\text{src}}^{*}$ **do**

  $d(j) \leftarrow \arg\min_{k \in \mathcal{D}} \|\mathbf{p}_j - \mathbf{p}_k\|_2$

end for

$\mathbf{F}'^{(t)} \leftarrow \text{Merge}\left(\mathbf{F}^{(t)}, \{j \rightarrow d(j)\}_{j \in \mathcal{S}_{\text{src}}^{*}}\right)$ ▷ $\mathcal{S}_{\text{con}}$ remains unchanged

return $\mathbf{F}'^{(t)}$

DaTMe merging in action

Conceptual example: N = 36, r = 0.5, P = 4

01 **Assign token roles**

Temporal ∩ spatial ranks

02 **Complete the budget**

Sample unassigned tokens

03 **Merge locally**

Keep constraints unchanged

Compute with 18 merged tokens



Protected + merged tokens

Constraints stay unchanged

Compute with fewer tokens

Restore positions after computation

Evaluation on PixArt-α and FLUX.1-dev

Setting	PixArt-α	FLUX.1-dev
Denoising steps	20	35
Guidance scale	4.5	3.5
Compression ratios	0.3 / 0.5 / 0.7	0.3 / 0.5 / 0.7
Merging blocks	Transformer layers 9–19	Dual: 9–19; single: 7–22

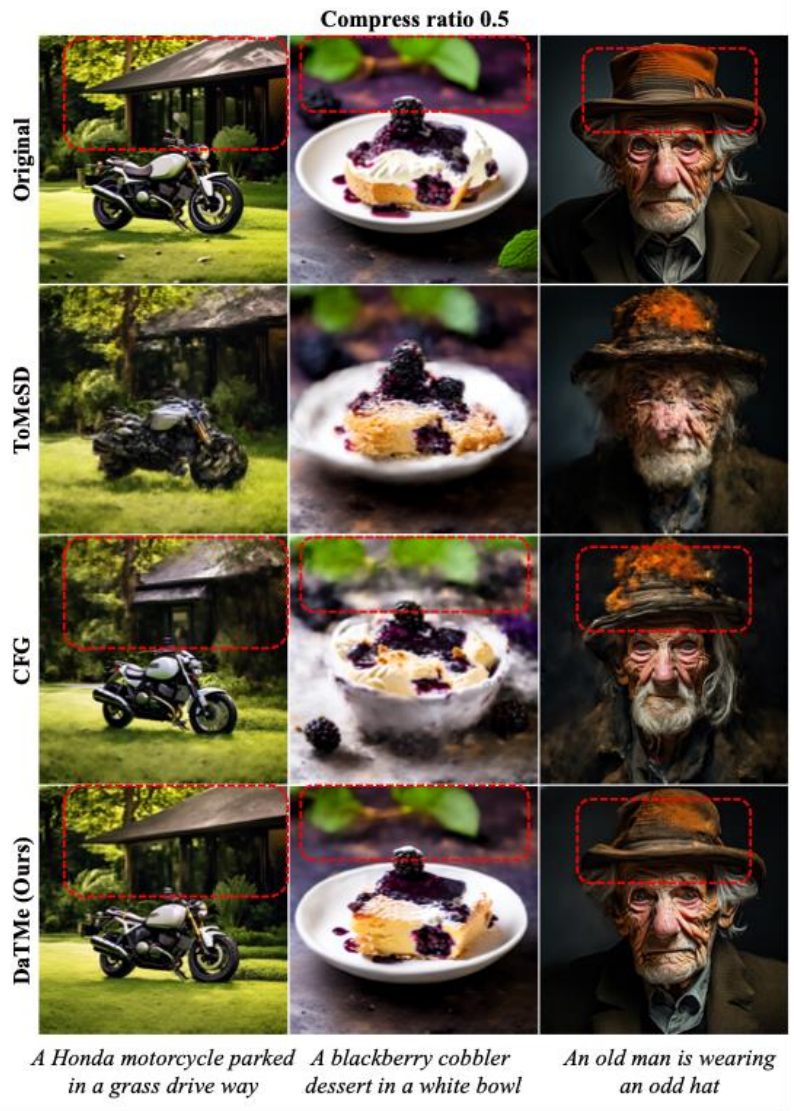
Data COCO 2014 validation · 10K pairs

Quality CLIP-T, HPSv2, FID, SSIM, PSNR, LPIPS, aesthetics

Efficiency Latency and TFLOPs on one RTX PRO 6000 Blackwell

Preserving structure beyond prompt-aligned objects

PixArt-α



FLUX.1-dev



Higher fidelity at a comparable compression budget

Table 1: **Text alignment and efficiency scores on PixArt-α.** DaTMe achieves the best overall quality-efficiency trade-off, preserving strong text alignment with minimal degradation under token compression.

	CLIP-T↑	HPSv2↑	FID↓	Latency(s)↓	TFLOPs↓
PixArt-α (Base) [5]	31.2485	29.2780	42.1001	1.9417	183.45
Compression ratio 0.3					
ToMeSD [10]	30.5735	24.1699	61.5229	1.6917	172.514
CFG [9]	30.8801	27.7375	44.6070	1.7458	172.514
DaTMe (Ours)	**31.1375**	**28.5591**	**42.4474**	1.8030	175.631
Compression ratio 0.5					
ToMeSD	29.5178	21.3334	94.8901	1.5918	165.201
CFG	30.4645	25.8345	52.2912	1.6445	165.201
DaTMe (Ours)	**31.0142**	**27.5174**	**43.4155**	1.6884	170.410
Compression ratio 0.7					
ToMeSD	28.2805	18.9797	123.0139	1.5356	157.897
CFG	29.4543	22.6156	70.7062	1.5761	157.897
DaTMe (Ours)	**30.6259**	**25.2393**	**47.0519**	1.6426	165.193

Table 2: **Image fidelity score on PixArt-α.** DaTMe achieves stronger structural preservation and image-level fidelity under token compression.

	SSIM↑	PSNR↑	LPIPS↓	Aes.↑
Compression ratio 0.3				
ToMeSD [10]	0.7306	18.9748	0.2433	5.9906
CFG [9]	0.7137	17.8157	0.2369	6.1521
DaTMe (Ours)	**0.8112**	**20.5136**	**0.1501**	**6.2014**
Compression ratio 0.5				
ToMeSD	0.6737	17.7161	0.3151	5.7739
CFG	0.6331	16.2768	0.3254	6.1135
DaTMe (Ours)	**0.7300**	**18.4277**	**0.2258**	**6.1695**
Compression ratio 0.7				
ToMeSD	0.6301	16.8537	0.3682	5.5809
CFG	0.5695	15.2003	0.4129	5.9559
DaTMe (Ours)	**0.6581**	**16.9048**	**0.3078**	**6.0544**

Higher fidelity at a comparable compression budget

Table 3: **Text alignment and efficiency score on FLUX.** DaTMe achieves stronger text alignment with comparable efficiency under token compression. † denotes our reproduced results.

	CLIP-T↑	HPSv2↑	FID↓	Latency(s)↓	TFLOPs↓
FLUX.1-dev (Base) [22]	31.2274	29.6927	30.6141	14.601	2098
Compression ratio 0.3					
ToMeSD† [10]	31.3254	28.9339	**30.4576**	13.866	2046
ToMA [12]	31.1787	28.5153	30.6520	13.790	1971
DaTMe (Ours)	**31.3524**	**29.5042**	30.9258	13.772	2033
Compression ratio 0.5					
ToMeSD†	31.1852	28.3595	**30.8491**	13.278	2000
ToMA	31.2036	27.7566	30.9883	13.288	1895
DaTMe (Ours)	**31.2621**	**28.9266**	31.3423	13.271	1990
Compression ratio 0.7					
ToMeSD†	31.2045	27.6455	**31.0033**	12.728	1954
ToMA	31.2191	26.4796	35.0541	12.775	1793
DaTMe (Ours)	**31.2415**	**27.9093**	31.1585	12.737	1946

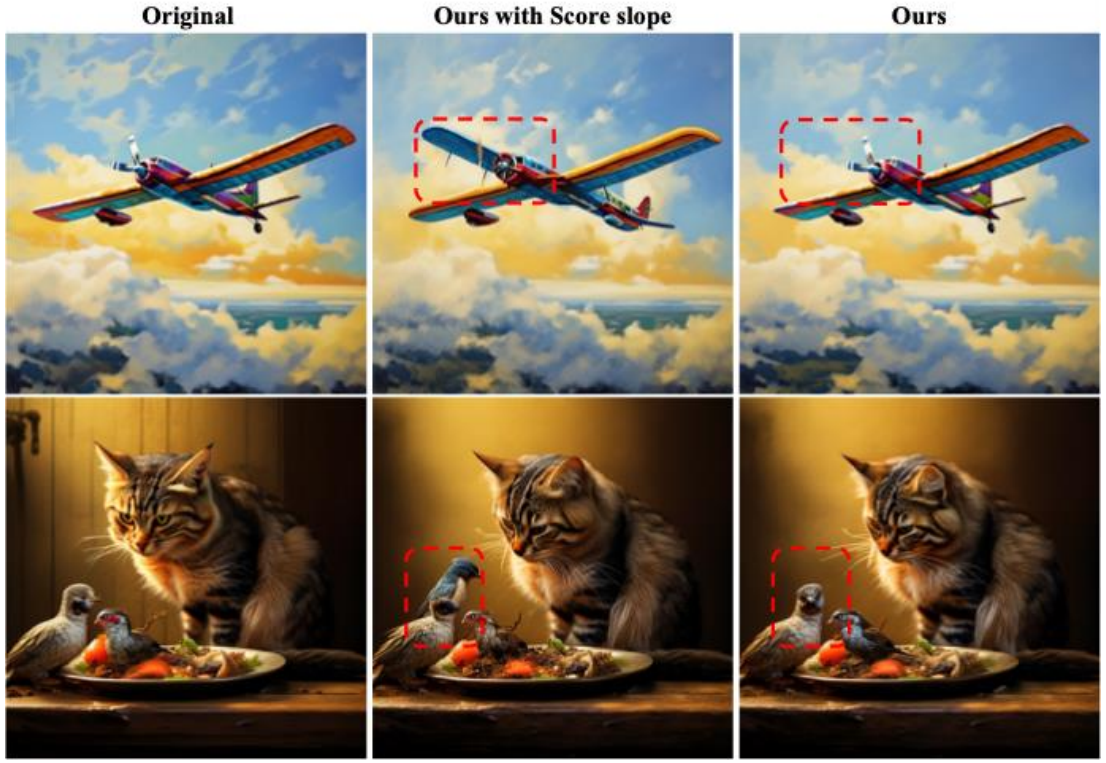
Table 4: **Image fidelity score on FLUX.** DaTMe achieves stronger structural preservation and image fidelity.

	SSIM↑	PSNR↑	LPIPS↓	Aes.↑
Compression ratio 0.3				
ToMeSD† [10]	0.6402	17.5070	0.2616	5.7626
ToMA [12]	0.7415	20.6450	0.1844	6.0210
DaTMe (Ours)	**0.7737**	**21.6352**	**0.1694**	**6.0318**
Compression ratio 0.5				
ToMeSD†	0.6166	16.9059	0.2932	5.6665
ToMA	0.7009	20.0375	0.2305	5.9030
DaTMe (Ours)	**0.7087**	**20.2021**	**0.2176**	**6.0191**
Compression ratio 0.7				
ToMeSD†	0.6032	16.6127	0.3181	5.5739
ToMA	0.6280	17.6972	**0.2835**	5.6971
DaTMe (Ours)	**0.6378**	**17.7655**	0.2877	**5.9147**

Temporal and structural cues are complementary

Table 5: **Ablation on temporal and spatial axes.** We compare using only the temporal axis, only the spatial axis, both axes, and both axes with score slope instead of Tweedie difference.

Setting	CLIP-T↑	SSIM↑	PSNR↑	FID↓
CFG [9]	30.8801	0.7137	17.8157	44.6070
Only $\mathcal{M}_{\text{temp}}$	31.0355	0.8109	20.5021	<u>42.5269</u>
Only $\mathcal{M}_{\text{spat}}$	<u>31.1318</u>	<u>0.8110</u>	<u>20.5029</u>	42.5736
Ours (score slope)	31.1166	0.8028	20.3944	42.9554
Ours	**31.1375**	**0.8112**	**20.5136**	**42.4474**



Replacing Tweedie difference with **score slope** reduces fidelity

The signal and merging design both contribute

Table 7: **Effectiveness of the proposed importance signal on PixArt-α at $r = 0.3$.** "Merging" indicates the merging strategy adopted from either [9] or ours, while "Signal" indicates whether token importance is determined by CFG [9] or our proposed importance map.

Merging	Signal	CLIP-T↑	HPSv2↑	SSIM↑	PSNR↑	LPIPS↓
CFG [9]	CFG [9]	30.8801	27.7375	0.7137	17.8157	0.2369
CFG [9]	Ours	31.0545	27.9366	0.7644	20.0917	0.1915
Ours	Ours	**31.1375**	**28.5591**	**0.8112**	**20.5136**	**0.1501**

Signal replacement improves the existing pipeline

Role assignment and spatial routing add further gains

TAKEAWAY

Token importance should follow generation dynamics

01 **Temporal evolution** reveals unfinished refinement

02 **Structural saliency** protects image-wide detail

03 **Dynamics-aware roles** guide faithful token merging

Training-free integration into existing diffusion models

github.com/SeungJuCha/DaTMe

Thank You

Poster Session 5 | Hall C1, Atlanta