NeurIPS 2025

Don’t Let It Fade:

Preserving Edits in Diffusion Language Models via Token Timestep Allocation

Woojin Kim, Jaeyoung Do†

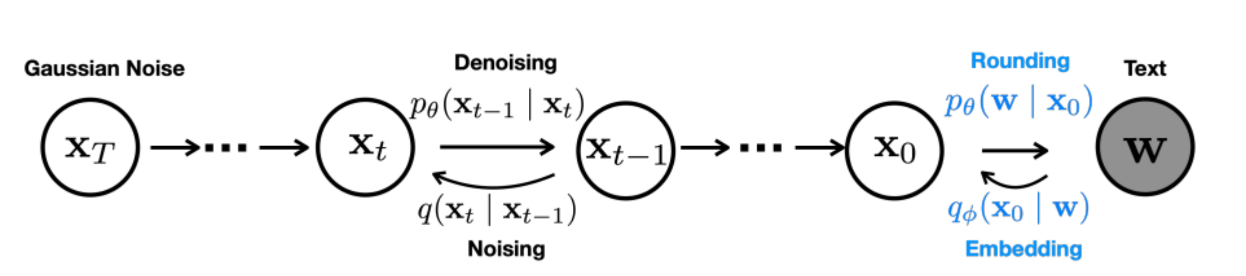
AIDAS LAB
ECE, Seoul National University
{wjk9904, jaeyoung.do}@snu.ac.kr

AIDAS Lab

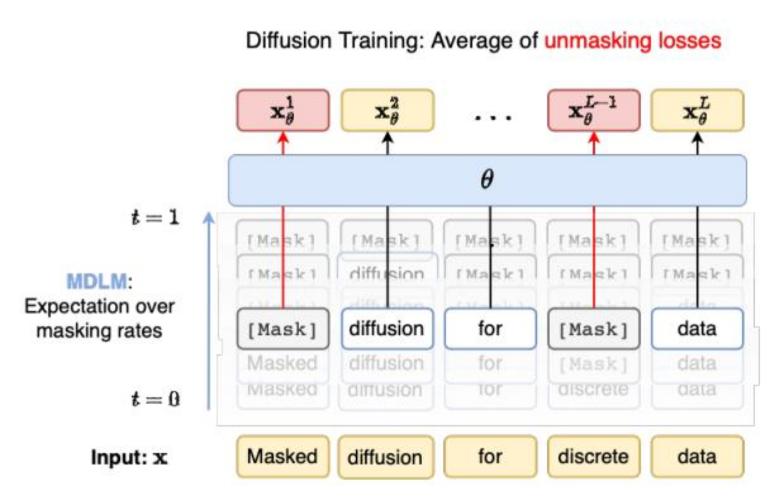
Diffusion Language Models

A **Diffusion Language Model** is a language model that generates text by iteratively denoising noise into coherent sequences, analogous to how diffusion models generate images from noise.

Diffusion-LM
(Li et al., 2022)

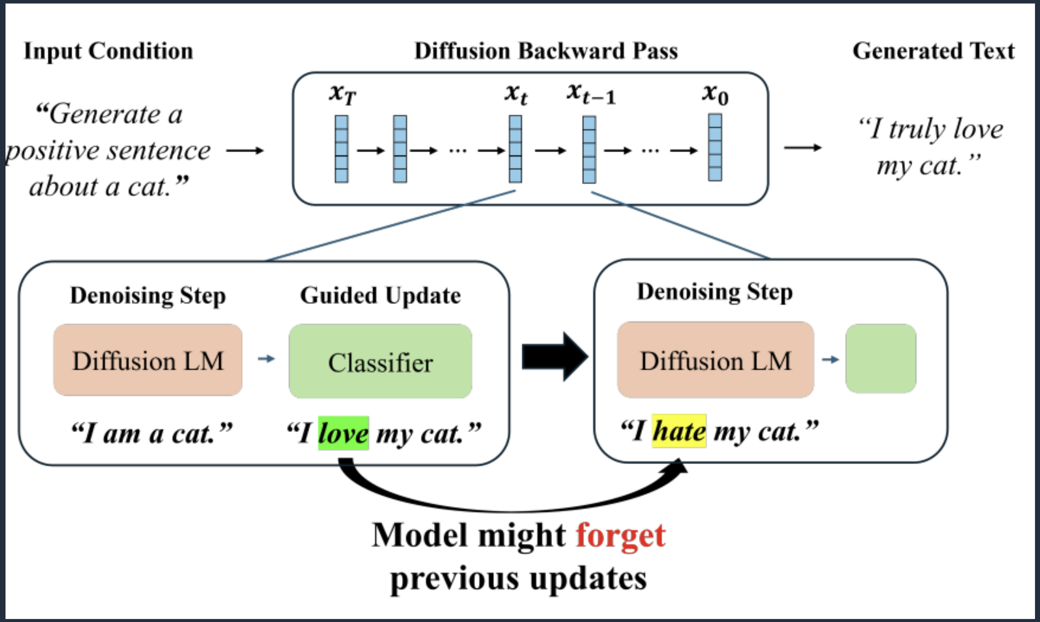


MDLM
(Sahoo et al., 2024)



Controllability Challenge in DLMs

With their iterative denoising and bidirectional context, diffusion language models (DLMs) enable fine-grained and flexible control over text generation.



However, major limitations remain:

(1) Low fluency — weak token dependency

(2) High computational cost — hundreds of steps

↓

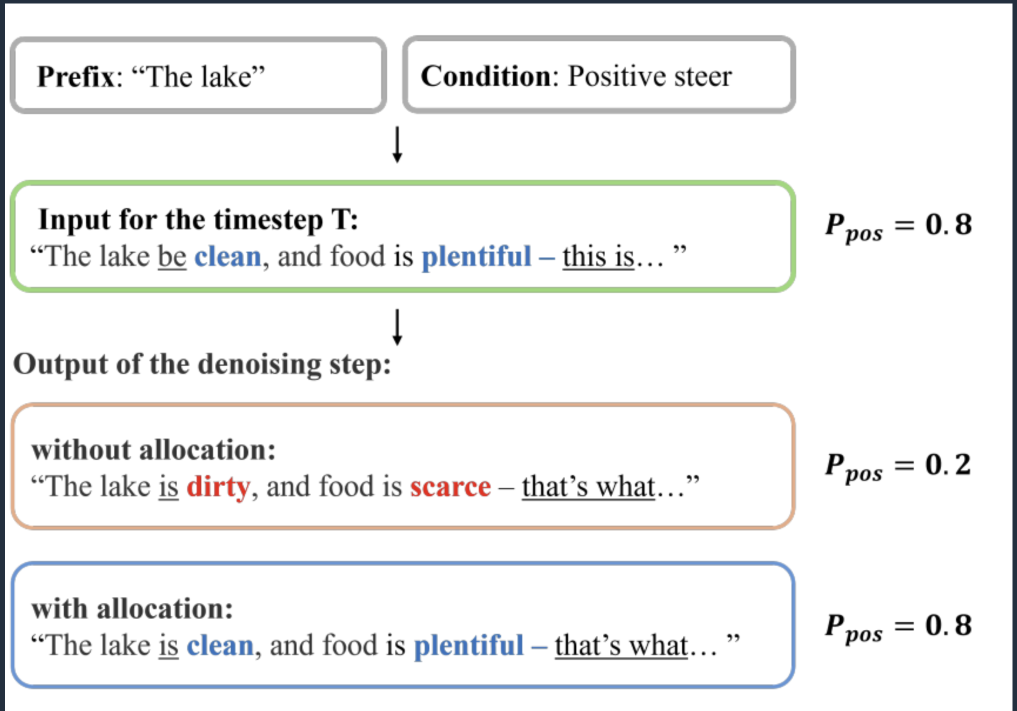
We argue that these issues stem from **uniform, context-agnostic** updates.

Our Contribution

Goal: Achieve **stable and controllable text generation** by preserving guided edits across timesteps.

Our Contributions:

1. Identify **update-forgetting** as the key bottleneck in controllable diffusion text generation.
2. Propose **TTA-Diffusion** — an inference-time method that allocates timesteps per token for stable control.
3. Demonstrate improved controllability, fluency, and efficiency across tasks and domains.



Diffusion Fluctuation

- Each diffusion step introduces small perturbations to tokens.
- When fluctuations grow large, sentences lose **coherence and fluency**.
- Strong correlation observed: **higher fluctuation → higher perplexity**.

  -> Indicates instability in token transitions harms generation quality.

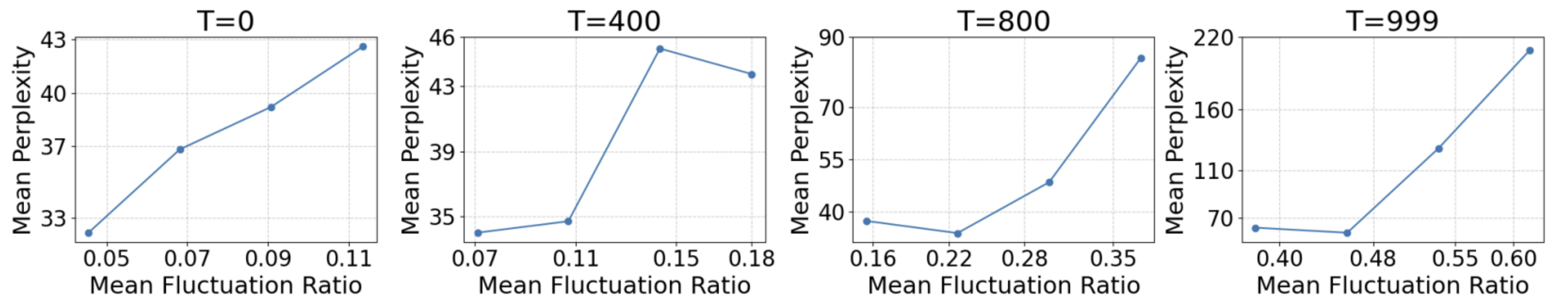


Figure 2: Fluctuation vs. perplexity across timesteps. At each timestep t, samples are grouped by fluctuation ratio, showing that higher fluctuation is consistently associated with higher perplexity.

Update Forgetting

- Guided edits made at one step often **fade in later steps**.
- Classifier confidence drops when key tokens are overwritten.

This causes semantic drift and loss of control accuracy.

Need for **preserving guided token updates** across timesteps

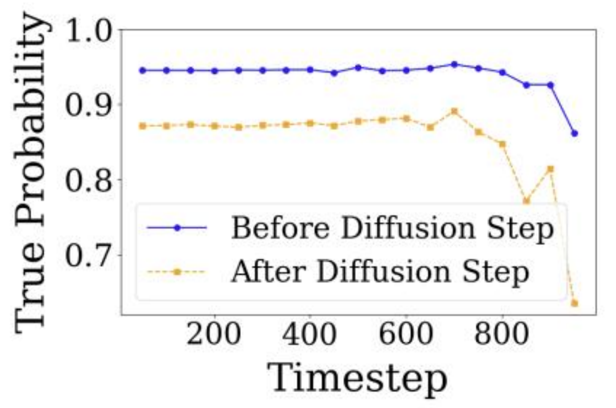


Figure 3: Classifier confidence drop due to update-forgetting.

Token Timestep Allocation (TTA-Diffusion)

- We propose a soft ordering based on timesteps, applied only during inference time.
- Each token has its own refinement rate, allowing flexible and continuous updates.

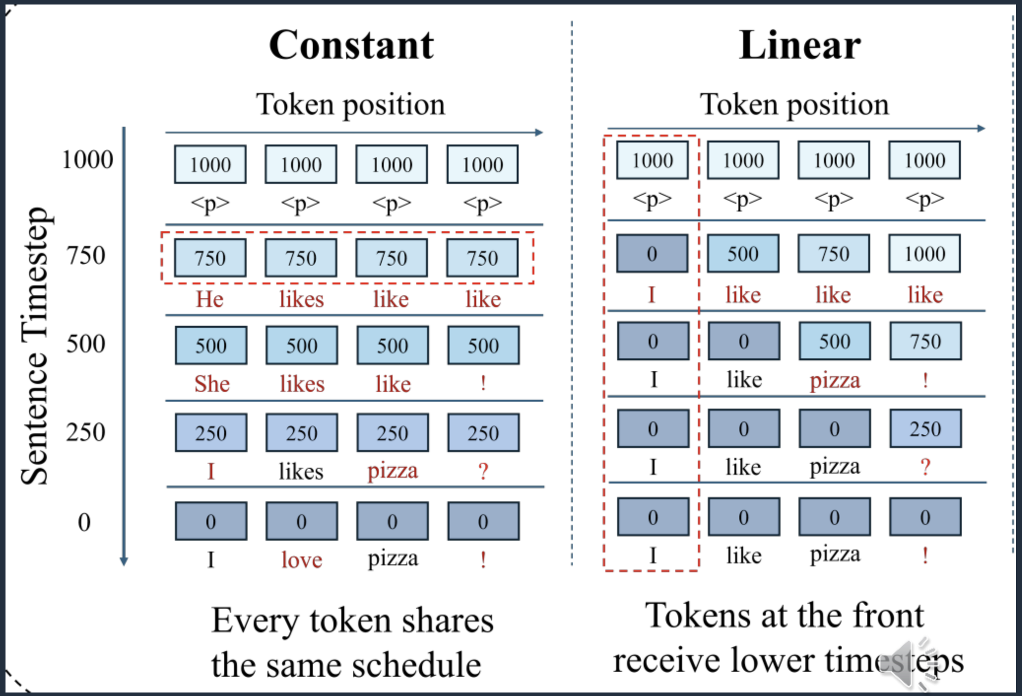
Core Idea:

- Assign a per-token timestep:

$$t_i = f(i, t)$$

Large t_i -> higher noise -> stronger denoising
Small t_i -> lower noise -> weak denoising

- This enables token-wise control in inference time



Semantic-based Adaptive Allocation

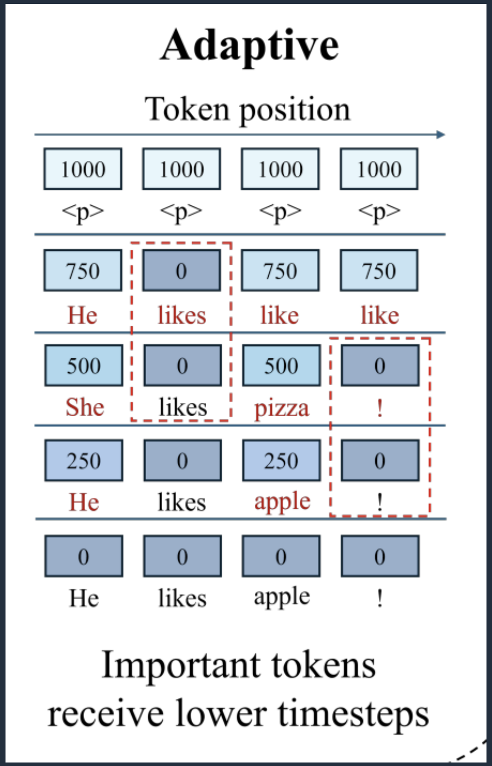
- Fixed schedules might ignore semantic importance
- Some tokens (e.g., sentiment words) should stay stable, others can change

Core Idea:

- Use classifier gradients to measure token importance.
- High gradient -> token has already been refined much -> assign smaller timestep

$$\hat{g}_i = \frac{g_i - \min_j g_j}{\max_j g_j - \min_j g_j}, \quad i = 1, \ldots, N.$$

$$t_i^{\text{adaptive}} = \alpha_{\text{smooth}} t + (1 - \alpha_{\text{smooth}})(1 - \hat{g}_i) t$$



Results: Controllable Text Generation

- We evaluate on **detoxification and sentiment control**, showing that **TTA-Diffusion improves both control accuracy and fluency** (lower perplexity).

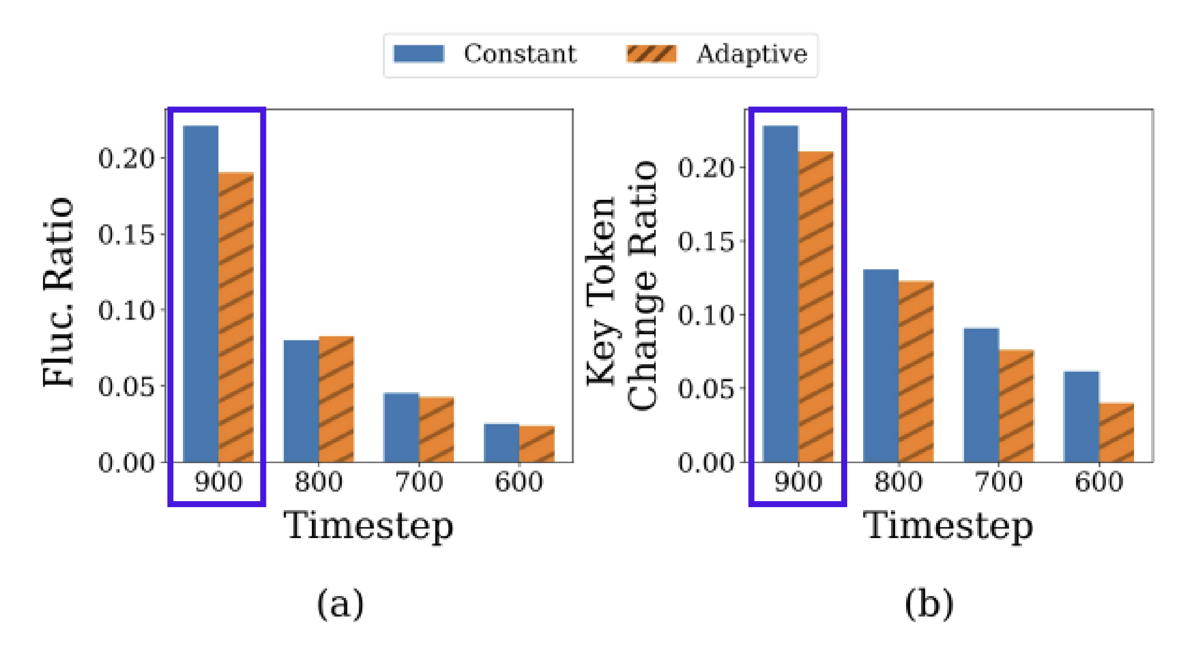
Model	Toxicity				Sentiment Control		
	Avg. tox↓	Max. tox↓	PPL↓	Dist-3↑	Acc↑	PPL↓	Dist-3↑
Auto-regressive Baselines							
PPLM	30.6	59.7	107.4	0.95	42.6	201.1	0.94
GeDi	22.0	36.1	98.8	**0.94**	79.9	98.6	0.91
DExperts	15.1	32.0	48.0	0.87	83.2	31.8	0.93
Air-decoding	18.5	40.4	49.0	0.93	82.6	27.1	0.94
LM-Steer	19.1	47.0	44.4	0.91	85.4	78.8	0.86
Diffusion Baselines							
Diffusion-LM$_{T=2000}$	21.8	-	131.2	0.94	72.8	89.3	0.94
SSD-LM$_{T=1000}$	24.6	50.3	58.3	0.94	76.2	51.1	**0.94**
LD4LG$_{T=250}$	14.5	-	296.4	0.90	59.9	70.7	0.95
TESS$_{T=1000}$	14.6	32.3	58.8	0.92	71.1	31.7	0.85
Ours							
TTA (50) $_{T=200}$	**12.2**	**26.0**	**40.6**	0.92	**94.7**	**20.5**	0.86
TTA (50) $_{T=100}$	12.2	26.7	46.3	0.93	92.7	28.7	0.86
TTA (50) $_{T=50}$	12.5	27.3	59.5	**0.94**	88.7	47.3	0.87

Results: Effect of TTA & Transferability

(a) Detoxification and sentiment control.

Model	T	Detoxification Tox. ↓	Detoxification PPL ↓	Sentiment Acc. ↑	Sentiment PPL ↓
TTA (5000)	200	13.2	630.4	80.8	47.3
+ with schedule		**12.8**	**70.8**	**82.1**	**35.5**
TTA (50)	50	14.0	68.0	83.5	44.0
+ with schedule		**12.5**	**59.5**	**85.9**	**40.2**

γ	Method	Valid (%)	Mean Property
1	D-CBG	989	0.474
	+ Adaptive	**998**	**0.494**
10	D-CBG	721	0.585
	+ Adaptive	**756**	**0.591**



For more detailed and interesting results, please check out our paper!

Thank you!

Woojin Kim, Jaeyoung Do[†]

AIDAS LAB
ECE, Seoul National University
{wjk9904, jaeyoung.do}@snu.ac.kr

AIDAS Lab