PAPER OVERVIEW · arXiv 2606.02544

SimSD

Simple Speculative Decoding in Diffusion Language Models

A plug-and-play attention mask that lets dLLMs verify drafted tokens in a single forward pass

Up to 7.46× throughput · Training-free · Plug-and-play

Junxia Cui*, Haotian Ye*, Runchu Tian*, Hongcan Guo, Jinya Jiang, Haoru Li, Chaojie Ren, Yiming Huang, Kaijie Zhu, Zhongkai Yu, Kun Zhou†, Jingbo Shang†

UC San Diego · UIUC · UC Santa Barbara

01 · PROBLEM

Fast dLLMs still can’t use speculative decoding

Autoregressive LLMs

- The causal mask keeps every token’s prefix fixed
- Concatenate γ drafted tokens: one forward pass gives valid logits for each, under its correct prefix

→ **Many drafted tokens verified in parallel**

Diffusion LLMs (dLLMs)

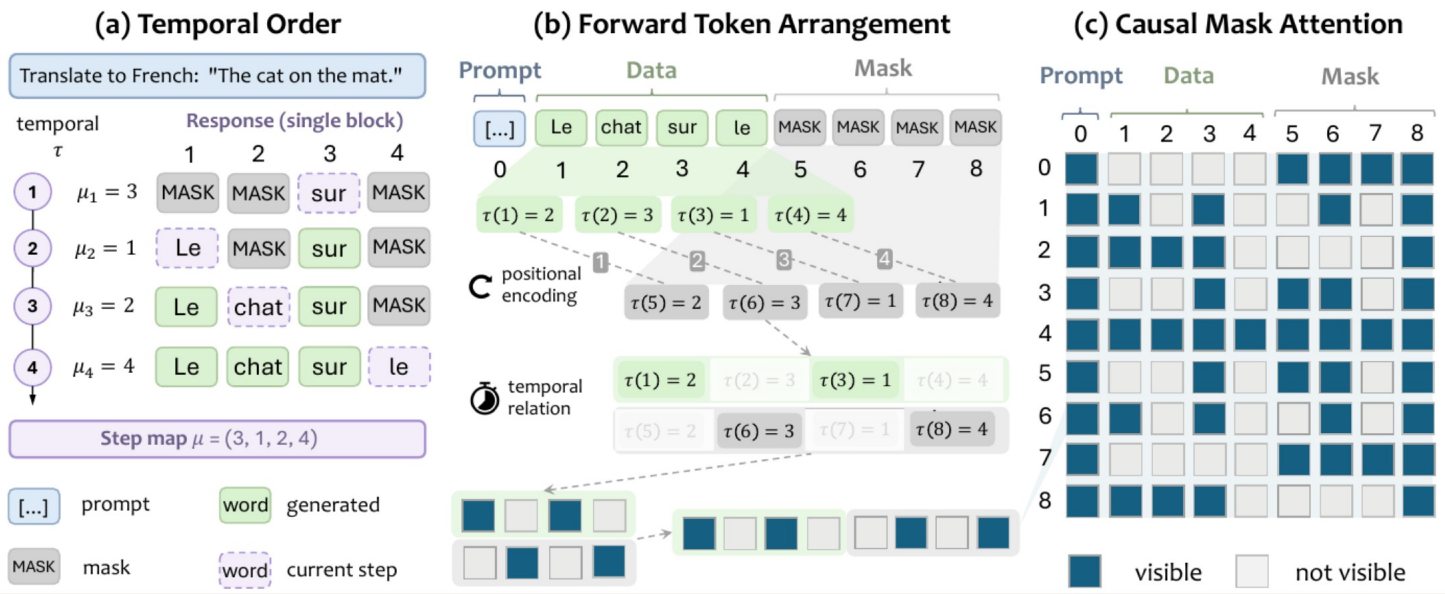
- [MASK] placeholders + bidirectional attention
- The effective context shifts across denoising steps, so there are no prefix-conditioned token logits

→ **Standard speculative verification does not apply**

What’s missing: temporally valid token-level context

02 · METHOD

Key idea: give dLLMs a token-level “prefix”



(a) Temporal order. A step map μ assigns each response position a decoding step τ, in any order

(b) Paired layout. Data slots hold reference tokens, Mask slots make predictions; each pair shares τ and its RoPE position

(c) Attention rule. A Mask slot sees only Data tokens from earlier steps, just like an AR prefix

03 · SPECULATIVE DECODING

Draft small, verify once with the large model

STEP 1

Draft

A small dLLM (SDAR-1.7B) proposes γ tokens and records the order it decoded them in

STEP 2

Verify

The target dLLM (SDAR-8B) builds the temporal mask from that order and scores all γ tokens in one pass

STEP 3

Accept

Rejection sampling with $\alpha = \min(1, p_{target} / p_{draft})$ keeps the longest accepted prefix

Training-free

Only the input layout, attention mask and RoPE indices change

Composable

Works with KV cache, blockwise decoding and CUDA Graph

Low overhead

Extra compute is $O(nB + B^2)$, negligible for $B = 4$ or 8

04 · EFFICIENCY

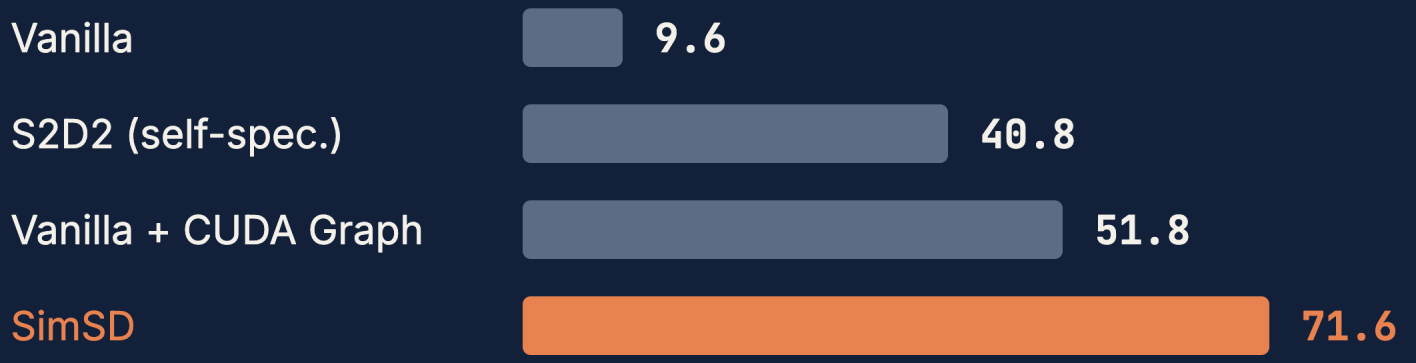
Up to 7.46× higher decoding throughput

7.46×

Block length B = 4, vs. Vanilla (TP = 2)

5.40×

Block length B = 8, vs. Vanilla (TP = 2)



Mean decoded tokens/s at B = 4 across GSM8K, MBPP, TriviaQA and MMLU. 1.75× faster than S2D2 (1.97× at B = 8).

05 · QUALITY

Faster, and average accuracy goes up

Benchmark (B = 4)	Vanilla	SimSD	Accept α
GSM8K	0.922	0.896 ↓	0.915
MBPP	0.669	0.662 ↓	0.806
TriviaQA	0.486	0.537 ↑	0.576
MMLU	0.706	0.758 ↑	0.973
Average	0.696	0.713 ↑	–

+1.7

points on average

B = 4: 0.696 → 0.713

B = 8: 0.681 → 0.698

Small drops on math and code, clear gains on knowledge tasks; draft acceptance reaches 0.973 on MMLU

06 · WHY IT WORKS

Small approximation error; RoPE alignment is key

Small approximation error

≥ 97.5%

One pass only approximates step-by-step denoising, yet top-1 agreement stays at or above 97.5% with self-drafting, and at or above 94% across models (1.7B → 8B)

RoPE alignment is essential

84% → 0%

Using physical positions instead of RoPE copy drops GSM8K accuracy to zero; acceptance falls from 0.902 to 0.718 and KL reaches 7.12

Speed–quality trade-off in γ

γ = B

Raising γ from 3 to 8 lifts acceptance from 0.834 to 0.964 and accuracy from 50% to 82%; throughput is not monotonic, so γ equals the block length

TAKEAWAYS

One mask brings speculative decoding to dLLMs

01	**Problem**	MASK tokens and bidirectional attention break token-level temporal context, so standard speculative verification fails in dLLMs
02	**Method**	Paired Data/Mask layout + temporal causal mask + RoPE alignment verify all drafts in one pass; training-free and plug-and-play
03	**Results**	Up to 7.46× decoding throughput on SDAR models across 4 benchmarks, with average accuracy up 1.7 points

Code: github.com/airevo2/SimSD-release