Plan for Speed

Dilated Scheduling for Masked Diffusion Language Models

Omer Luxembourg ▪ Haim Permuter ▪ Eliya Nachmani

School of Electrical and Computer Engineering, Ben-Gurion University of the Negev

Motivation

- Masked diffusion LMs promise **parallel generation**.
- But existing confidence and entropy based planners, in high probability, **pick neighbouring tokens**.
- Due to the model architecture, they ignore dependencies between co-decoded positions.
- Effectively yielding **near-autoregressive, clustered decoding**.

Inference planners can reach different quality and speedups; DUS is an inference planner.

Background: masked diffusion language models

- **Forward process:** each token is replaced by `[MASK]` with probability t.
- **Training:** predict masked tokens with reweighted cross-entropy

$$\mathcal{L}(\theta) = -\,\mathbb{E}_{t,\mathcal{X},\mathcal{M}_t}\Big[\frac{1}{t}\sum_{i:\,M_{t,i}=\texttt{[MASK]}} \log p_\theta(X_i \mid \mathcal{S}_t)\Big].$$

- **Inference loop:** *planner* selects masked positions $\rightarrow$ denoiser predicts them $\rightarrow$ state updates.
- **Key lever:** planners control the quality and speed of generation. **DUS is a planner.**

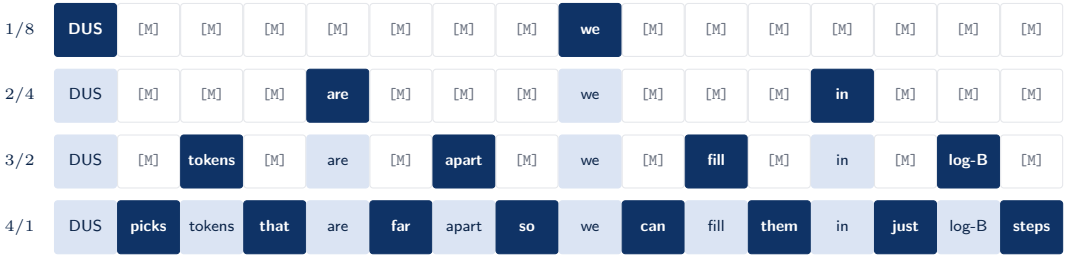
Dilated Unmasking Scheduler (DUS)

Within each block of B tokens, at iteration $t = 1, \ldots, R$, DUS unmasks positions

$$\mathcal{I}_t = \Big\{ k \in \{1, \ldots, B\} \setminus \mathcal{U}_{t-1} \;\Big|\; (k-1) \bmod s_t = 0 \Big\},$$

with

$$s_t = \lfloor B/2^t \rfloor, \qquad R = \lceil \log_2 B \rceil, \qquad \mathcal{U}_t = \mathcal{U}_{t-1} \cup \mathcal{I}_t,\ \mathcal{U}_0 = \emptyset.$$



$B{=}16 \Rightarrow 4$ denoiser calls ($s_t{=}8, 4, 2, 1$). **dark**=newly unmasked, **light**=already unmasked, white=masked.

Novelties

- **Inference-only, model-agnostic.** Drop-in scheduler; zero changes to architecture or training.
- **Logarithmic schedule.** Deterministic coarse-to-fine dilation in $\mathcal{O}(\log B)$ denoiser calls per block.
- **Theory.** The cheap sum-of-marginals objective is tightest exactly when spacing is largest.
- **Empirical breadth.** 8 benchmarks $\times$ 5 models (LLaDA, Dream, DiffuCoder).
- **Hybrid extension.** Dilated spacing as a post-filter on top of adaptive EB/CB samplers.

Theory: VLMC spacing bound (informal)

Assume a stationary, ergodic **fast-mixing** chain. For positions in the same DUS step with spacing s_t:

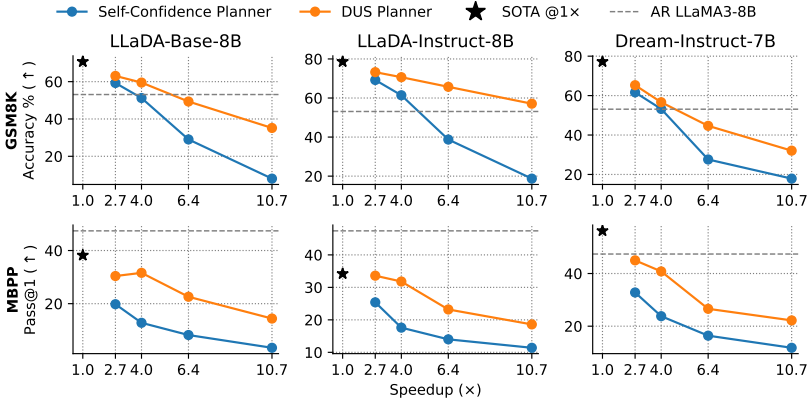
$$I(X_i;\, X_{i+s_t}, X_{i+2s_t}, \dots) \;\le\; C\,\rho^{\,s_t}, \qquad \rho \in (0,1).$$

Mutual information decays *exponentially* with spacing, so the gap between the true joint entropy and the cheap sum-of-marginals shrinks:

$$H(X_{\mathcal{I}_t} \mid \mathcal{S}_t) \;\ge\; \sum_{i \in \mathcal{I}_t} H(X_i \mid \mathcal{S}_t) - \varepsilon(s_t), \qquad \varepsilon(s_t) \text{ exp. small in } s_t.$$

Tightest at coarse spacing; later steps lean on the richer revealed context.

Main result: the speed-quality frontier



DUS (**orange**) vs self-confidence (**blue**) vs Llama-3-8B AR (gray dashed), on GSM8K, MATH500, HumanEval, MBPP. Up to **+27 points** over self-confidence at matched compute.

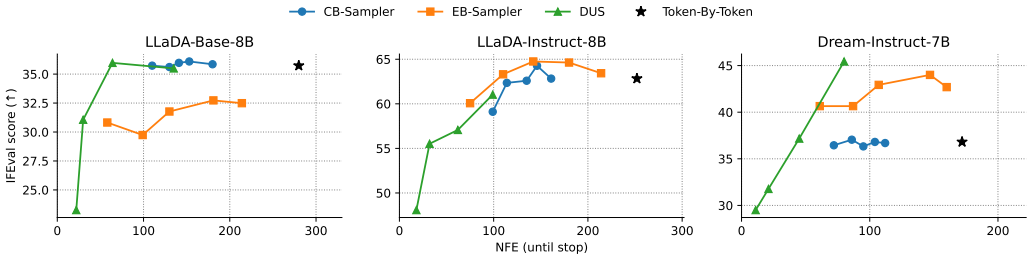
Consistent gains across models and tasks

		TBT	B=16 ($\times$4)		B=32 ($\times$6.4)		B=64 ($\times$10.7)	
Data	Model	Conf.	Conf.	DUS	Conf.	DUS	Conf.	DUS
GSM8K	LLaDA-Base	72.6	51.2	**59.5**	29.0	**49.4**	8.0	**35.2**
	LLaDA-Inst	80.3	61.4	**70.7**	38.7	**65.7**	18.7	**57.1**
	Dream-Inst	77.1	53.2	**56.6**	27.6	**44.7**	17.9	**32.1**
MBPP	LLaDA-Base	38.0	12.8	**31.6**	8.2	**22.6**	3.4	**14.4**
	LLaDA-Inst	39.4	17.6	**31.8**	14.0	**23.2**	11.4	**18.6**
	Dream-Inst	56.2	23.8	**40.8**	16.4	**26.6**	11.8	**22.2**

GSM8K accuracy / MBPP pass@1 (%). **Bold** = better planner. TBT = token-by-token baseline. The gap widens with speed-up: at $\times 10.7$, DUS more than **quadruples** self-confidence on GSM8K and **quadruples** it on MBPP (LLaDA-Base).

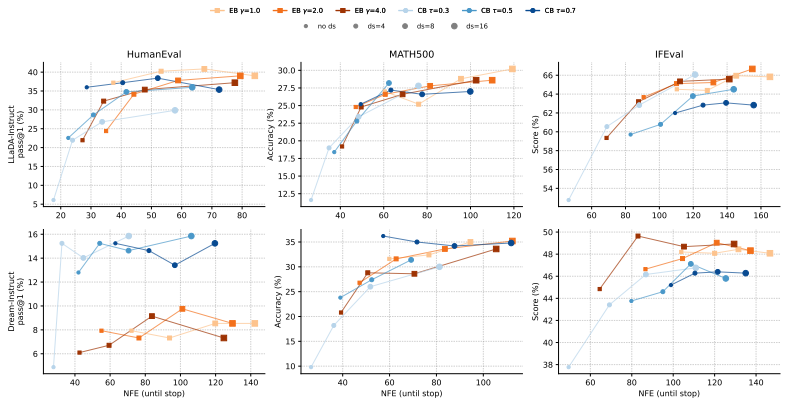
Versus adaptive samplers (EB / CB)

On IFEval, DUS (**green**) sits **left** of entropy-bounded (EB) and confidence-bounded (CB) at every accuracy level: it reaches near token-by-token accuracy at 3–$4\times$ **fewer** denoiser calls (NFE), and on math/code DUS at $B{=}16$ stays within $\sim$2 points of EB at fewer calls.



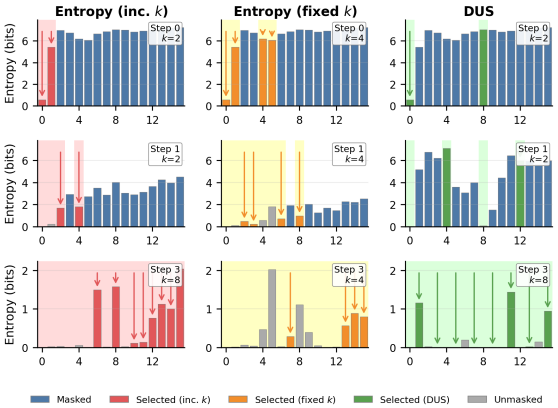
Hybrid: dilated spacing as a post-filter

Apply dilated spacing *on top of* EB ($\gamma{=}2$) and CB ($\tau{=}0.5$) samplers, with no change to the score function: keep their scores, but enforce a minimum gap between co-decoded positions. Accuracy rises on **every** dataset and sampler for a modest NFE increase; biggest jump $+13.4$ **pts** on HumanEval (EB).



LLaDA-Inst (top) and Dream-Inst (bottom). Accuracy vs. NFE; marker size = spacing block ($\texttt{ds}{=}4, 8, 16$). +**spacing** pushes every curve up-and-right.

Why it works: DUS spaces, others cluster



Per-position entropy and each planner's picks at $B{=}16$. DUS (right) keeps co-decoded tokens $\sim$2–3$\times$ further apart.

The same spacing shows up in the statistics (LLaDA-Base):

		Avg. pairwise distance			
Data	B	**DUS**	SC	EB	CB
MATH500	16	**6.2**	3.0	3.5	4.5
MATH500	32	**9.6**	4.1	4.7	7.1
HumanEval	16	**6.2**	2.8	4.4	4.7
HumanEval	32	**11.6**	3.7	6.4	8.6

- DUS spreads co-decoded tokens 2–3$\times$ **wider** than SC / EB / CB.
- 100% of DUS picks are *isolated* (no adjacent co-decoded token) vs $\sim 50\%$ for SC.
- Exactly the near-independent regime our theory calls safe.

Takeaways

- *Space, don't cluster*: separate co-decoded tokens.
- Training-free, deterministic, $\mathcal{O}(\log B)$ planner.
- Up to **+27** pts at matched compute; 3–$4\times$ fewer NFE than adaptive samplers.
- Composes as a post-filter on top of EB / CB.

omerlux.github.io/DUS

Code & paper