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

ICML International Conference On Machine Learning

Motivation

Masked diffusion LMs promise parallel generation, but existing confidence- and entropy-based planners pick neighbouring tokens, ignore dependencies between co-decoded positions, and effectively yield near-autoregressive, clustered decoding.

Background. Masked diffusion language models

- **Forward process:** for noise level $t \in (0, 1]$, each token is replaced by `[MASK]` with probability t, producing $\mathcal{M}_t$ from $\mathcal{X}$.
- **Training objective:** predict masked tokens with reweighted cross-entropy

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

- **Inference loop:** planner selects masked positions, denoiser predicts them, state $\mathcal{S}_t$ updates.
- **Key lever:** the planner controls the speed-quality trade-off; 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

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

Spacing keeps co-decoded tokens far apart, so the tractable sum-of-marginals entropy is a tight surrogate for the joint entropy under fast-mixing.

Example. B=16 $\Rightarrow$ 4 denoiser calls (s_t=8, 4, 2, 1):

Stage: coarse → fine (row label: t/s_t)

1/8	DUS	[M]	[M]	[M]	[M]	[M]	[M]	[M]	we	[M]	[M]	[M]	[M]	[M]	[M]	[M]
2/4	DUS	[M]	[M]	[M]	are	[M]	[M]	[M]	we	[M]	[M]	[M]	in	[M]	[M]	[M]
3/2	DUS	[M]	tokens	[M]	are	[M]	apart	[M]	we	[M]	fill	[M]	in	[M]	log-B	[M]
4/1	DUS	picks	tokens	that	are	far	apart	so	we	can	fill	them	in	just	log-B	steps

darker blue = newly unmasked, **lighter blue** = already unmasked, **white box** = still masked.

Novelties

- **Inference-only, model-agnostic.** Drop-in scheduler, with zero changes to architecture or training.
- **Logarithmic schedule.** Deterministic, coarse-to-fine dilation in $\mathcal{O}(\log B)$ denoiser calls per block.
- **Theory.** Under fast-mixing, the surrogate is tightest in early coarse iterations (large spacing); later iterations rely on richer revealed context.
- **Empirical breadth.** 8 benchmarks (math, code, reasoning, IFEval) × 5 models (LLaDA, Dream, DiffuCoder).
- **Hybrid extension.** Dilated spacing as a drop-in post-filter on EB/CB samplers (Sec. 4.4).

Theory. VLMC spacing bound (informal)

Assume a stationary, ergodic fast-mixing VLMC. For indices in the same DUS step with minimum spacing s_t:

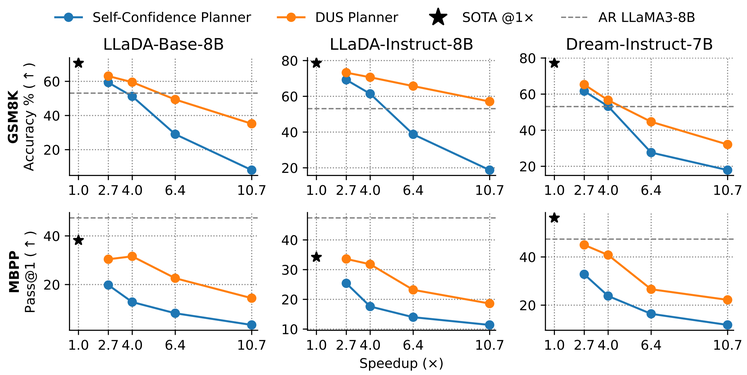
$$I\big(X_i;\, X_{i+s_t}, X_{i+2s_t}, \ldots\big) \;\le\; C\,\rho^{\,s_t}, \qquad \rho \in (0, 1).$$

MI decays with spacing, so within-step dependence weakens and the entropy gap between joint and sum-of-marginals objectives shrinks. Hence, for the same set $\mathcal{I}_t$,

$$H\big(X_{\mathcal{I}_t} \mid \mathcal{S}_t\big) \;\ge\; \sum_{i \in \mathcal{I}_t} H\big(X_i \mid \mathcal{S}_t\big) - \varepsilon(s_t),$$

with $\varepsilon(s_t)$ exponentially small in s_t.
Tightest at coarse spacing; later gains come from richer revealed context.

Main Results. Speed-Quality Frontier (4 Benchmarks, 5 Models)

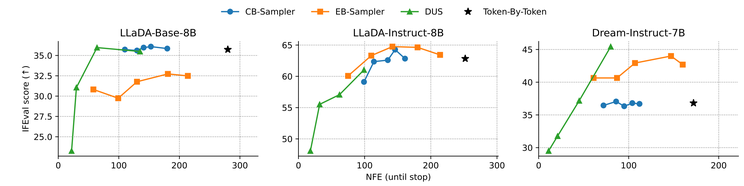


DUS (orange) vs self-confidence (blue) vs Llama-3-8B AR (gray dashed) on GSM8K, MATH500, HumanEval, MBPP.

		TBT	B=8 (×2.7)		B=16 (×4)		B=32 (×6.4)		B=64 (×10.7)	
Dataset	**Model**	Conf.	Conf.	DUS	Conf.	DUS	Conf.	DUS	Conf.	DUS
GSM8K	LLaDA-Base	72.63	59.29	**63.08**	51.23	**59.51**	29.04	**49.36**	8.04	**35.18**
	LLaDA-Inst	80.29	69.22	**73.24**	61.41	**70.66**	38.74	**65.73**	18.73	**57.09**
	Dream-Inst	77.10	61.64	**65.28**	53.22	**56.63**	27.60	**44.66**	17.89	**32.07**
MATH500	LLaDA-Base	24.00	16.6	**21.4**	11.2	**19.2**	6.0	**13.6**	2.6	**10.2**
	LLaDA-Inst	28.80	21.4	**23.8**	15.4	**22.8**	10.8	**19.2**	8.0	**14.8**
	Dream-Inst	37.00	22.4	**27.0**	15.4	**19.8**	7.2	**13.2**	4.0	**11.6**
HumanEval	LLaDA-Base	34.76	15.85	**25.61**	12.80	**19.51**	4.88	**14.02**	4.88	**6.71**
	LLaDA-Inst	39.02	21.95	**28.05**	14.02	**23.17**	9.76	**10.37**	10.98	**11.59**
	Dream-Inst	57.90	8.54	**14.63**	5.49	**11.59**	**6.71**	**6.71**	6.10	**9.15**
	DiffuCoder-Base	67.10	17.07	**28.66**	6.71	**38.41**	2.44	**21.95**	0.61	**6.10**
	DiffuCoder-Inst	72.00	7.93	**22.56**	14.02	**20.12**	**13.41**	12.80	**11.59**	8.54
MBPP	LLaDA-Base	38.0	19.8	**30.4**	12.8	**31.6**	8.2	**22.6**	3.4	**14.4**
	LLaDA-Inst	39.4	25.4	**33.6**	17.6	**31.8**	14.0	**23.2**	11.4	**18.6**
	Dream-Inst	56.2	32.8	**45.0**	23.8	**40.8**	16.4	**26.6**	11.8	**22.2**
	DiffuCoder-Base	74.2	29.2	**48.6**	17.4	**43.0**	10.2	**27.4**	3.4	**17.2**
	DiffuCoder-Inst	75.1	31.8	**46.4**	25.6	**43.6**	21.0	**26.6**	13.0	**18.2**

Accuracy (%) for math and pass@1 for code. **Bold** = better planner. TBT = token-by-token MDLM baseline. Headline: up to **+27 points** vs self-confidence at matched NFE, and up to **+13 points** at matched block size.

Vs. adaptive samplers (EB/CB)



On IFEval (G=1024), DUS (green) sits left of EB (orange) and CB (blue) at every accuracy level: **near-token-by-token accuracy at** 3-4× **fewer NFE**. On math/code, DUS at B=16 stays within ∼2 accuracy points of EB at fewer denoiser calls (LLaDA-Inst).

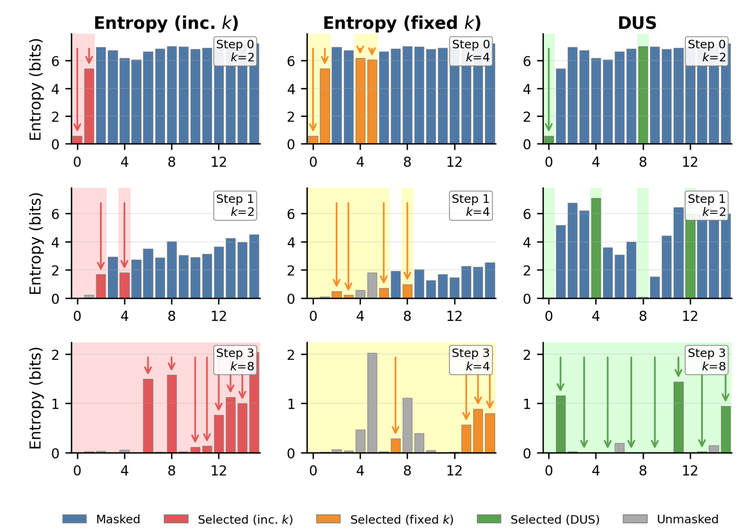
Hybrid: Dilated Spacing as a Post-Filter

Applied on top of entropy-bounded (EB, γ=2) and confidence-bounded (CB, τ=0.5) samplers with initial gap g_0=8, dilated spacing improves accuracy at modestly higher NFE, without modifying the score function.

	LLaDA-Inst				Dream-Inst			
	EB (γ=2)		CB (τ=0.5)		EB (γ=2)		CB (τ=0.5)	
Dataset	off	+ spacing	off	+ spacing	off	+ spacing	off	+ spacing
HumanEval	24.4 / 35	**37.8** / 59	22.6 / 22	**34.8** / 42	7.9 / 55	**9.8** / 101	12.8 / 42	**14.6** / 70
MATH500	24.8 / 47	**27.8** / 81	18.4 / 37	**28.2** / 62	26.8 / 47	**33.6** / 84	23.8 / 39	**31.4** / 69
IFEval	63.7 / 91	**65.2** / 132	59.7 / 83	**63.8** / 120	46.6 / 87	**49.0** / 121	43.8 / 80	**47.1** / 108

Cells: Accuracy / NFE. Headline gain: **+13.4 accuracy points** on HumanEval (LLaDA-Inst, EB).

Mechanism. DUS Spaces Co-Decoded Tokens Further Apart



Per-position entropy + each planner's picks (arrows) at B=16. DUS (right) keeps co-decoded tokens ∼2-3× further apart on average than entropy-based planners (Sec. 3.4).