Large Language Continuous Diffusion Models

Sigma: scalable training, trajectory steering, and few-step generation

Paper by Zhihan Yang, Wei Guo, Jean-Marie Lemercier,
Simon Welker, et al.

NVIDIA / Cornell University / Georgia Institute of Technology

Paper discussion · 14 October 2026
arXiv:2610.02665v1

Continuous token trajectories reach competitive 3B/8B reasoning quality.

Can continuous diffusion handle conditional reasoning?

The question

Scale likelihood-trained diffusion over learned token embeddings to 3B/8B models, with math and code evaluation against a matched masked dLM.

Design

- 16D continuous state per token.
- 32-token blocks; cached clean prefix.
- AR warm start and auxiliary AR loss.

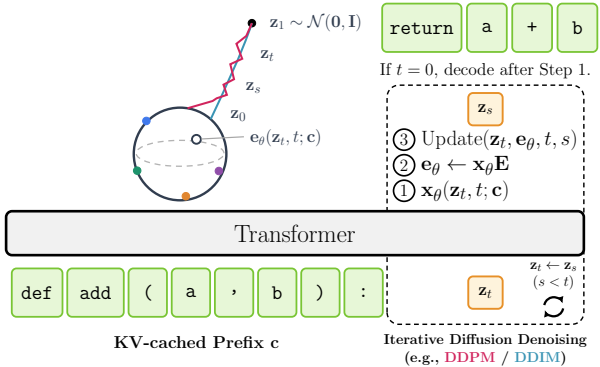
Six-task base-model average

	Sigma	NLD
3B	64.34	64.74
8B	69.67	70.22

Sigma: DDPM-256, diffusion mode.

Paper, Secs. 1, 3; Tables 1–2, PDF pp. 1–5, 8. Scores in percent; quality comparison.

Generate each block by moving continuous states

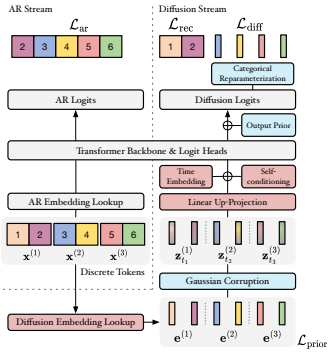


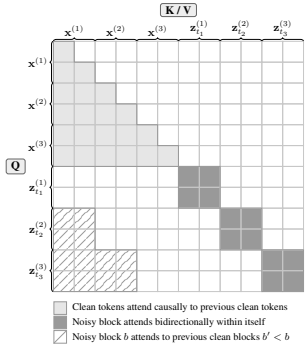
1. Cache the clean prefix.
2. Draw Gaussian noise for the new block.
3. Jointly denoise all positions.
4. Read out tokens and extend the prefix.

Autoregressive across blocks; bidirectional within a noisy block.

Paper, Fig. 1 (left), PDF p. 2; Sec. 4.1, PDF p. 6. Original author figure.

Two streams share a Transformer backbone





Clean tokens attend causally. A noisy block sees earlier clean blocks and itself.

Paper, Fig. 2, PDF p. 4. Original architecture and attention-mask panels.

The likelihood objective also learns the geometry

$$\mathbf{E} \in \mathbb{R}^{V\times 16}, \quad \|\mathbf{E}_v\|_2 = 1, \qquad \mathbf{z}_t = \alpha_t \mathbf{x}\mathbf{E} + \sigma_t \epsilon, \quad \alpha_t^2 + \sigma_t^2 = 1$$

$$\underbrace{\mathbf{x}_\theta(\mathbf{z}_t, t; \mathbf{c})}_{\text{vocabulary probabilities}} \qquad \mathbf{E} = \qquad \underbrace{\mathbf{e}_\theta(\mathbf{z}_t, t; \mathbf{c})}_{\text{predicted mean embedding}}$$

A per-block negative ELBO

$$\mathcal{L}_{\text{NELBO}} = \mathcal{L}_{\text{prior}} + \mathcal{L}_{\text{rec}} - \tfrac{1}{2}\mathbb{E}\Big[\text{SNR}'(t)\|\mathbf{e}_\theta - \mathbf{x}\mathbf{E}\|_F^2\Big]$$

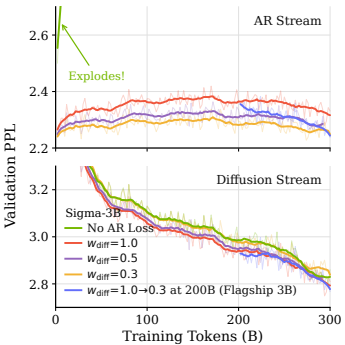
- Prior KL aligns the noisy endpoint; reconstruction CE reads out tokens.
- $\text{SNR}'(t) < 0$: the denoising MSE has positive weight.
- Sum the bound over $p_\theta(\mathbf{x}^{(b)} \mid \mathbf{x}^{(<b)})$ across blocks.

Paper, Eqs. (1)–(4), PDF pp. 2–3. 16D is the latent size; the output still spans all 131,072 tokens.

AR transfer and self-conditioning make scaling practical

$$\mathcal{L} = \mathcal{L}_{\text{ar}} + w_{\text{diff}}\mathcal{L}_{\text{BD}}$$

- Start from a 1T-token AR checkpoint; randomly initialize 16D embeddings.
- Continue for 150K steps, over 300B tokens.
- Switch $w_{\text{diff}} : 1.0 \rightarrow 0.3$ at 200B tokens.
- SC: detached first prediction on 25% of training samples.



Paper, Fig. 3c, Sec. 3.2, App. E; PDF pp. 5, 28. Validation PPL: lower is better.

DDIM reduces the cost of many-step generation

For $s < t$, predict a clean embedding $\hat{\mathbf{e}}_\theta$ and implied noise:

$$\hat{\epsilon}_\theta = \frac{\mathbf{z}_t - \alpha_t \hat{\mathbf{e}}_\theta}{\sigma_t}.$$

DDIM: deterministic given the initial noise

$$\mathbf{z}_s = \alpha_s \hat{\mathbf{e}}_\theta + \sigma_s \hat{\epsilon}_\theta.$$

DDPM instead samples a Gaussian reverse transition:

$$\mathbf{z}_s = (1-\kappa)\frac{\alpha_s}{\alpha_t}\mathbf{z}_t + \kappa\alpha_s\hat{\mathbf{e}}_\theta + \sqrt{\kappa(1-\alpha_s^2)}\,\epsilon, \quad \kappa = 1 - e^{\gamma(s)-\gamma(t)}.$$

One extra call at $t = 0$ performs the vocabulary argmax readout.

Paper, Sec. 4.1, PDF p. 6; Algorithm 2, PDF p. 27. Main-table step budgets exclude this readout.

CFG and score temperature steer different offsets

Let $\mathbf{e}_c$ use the clean prefix and $\mathbf{e}_\varnothing$ drop the prefix.

Classifier-free guidance

$$\mathbf{e}^w = \mathbf{e}_\varnothing + w(\mathbf{e}_c - \mathbf{e}_\varnothing)$$

$w \geq 1$: amplify the conditional change.

Score temperature

$$\mathbf{e}^\tau = \frac{\mathbf{z}_t}{\alpha_t} + \frac{1}{\tau}\left(\mathbf{e}_c - \frac{\mathbf{z}_t}{\alpha_t}\right)$$

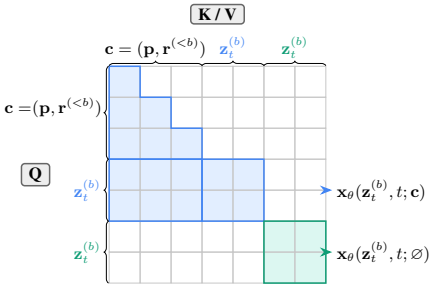
$0 < \tau \leq 1$: rescale the score by $1/\tau$.

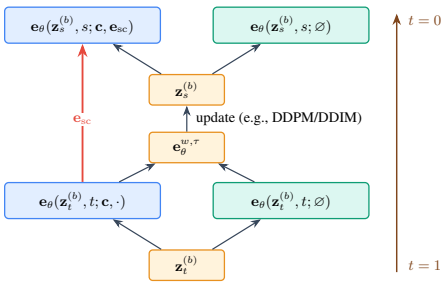
$$\mathbf{e}^{w,\tau} = \frac{w}{\tau}\mathbf{e}_c + \frac{1-w}{\tau}\mathbf{e}_\varnothing + \left(1 - \frac{1}{\tau}\right)\frac{\mathbf{z}_t}{\alpha_t}$$

- Both act in embedding space; ST differs from logit temperature.
- The two affine transformations commute.

Paper, Sec. 4.2, PDF pp. 6–7; Proposition 1 / Eq. (28), PDF pp. 45–46.

Packed guidance has a small marginal latency cost





- Conditional and unconditional queries share one packed forward pass.
- SC reuses the previous *unsteered* conditional mean; unconditional SC is zero.
- H100, 3B, batch 1, block 32: 19.08 → 19.89 ms per NFE (+4.2%).

Paper, Fig. 4, PDF p. 7; Table 20, PDF p. 45. Timing covers the first generated block.

Read the evaluation as several distinct protocols

Evaluation	What is scored?	Relevant conditions
Base models	2 math + 4 coding tasks; pass@1	Few-shot prompts; 256/512 output tokens
Instruct	10 tasks: math, code, knowledge, instructions	Non-thinking SFT; task-specific output caps
Multiple choice	Candidate ranking by negative ELBO	Monte Carlo likelihood surrogate
Pass@k	At least one correct candidate among k	Paired uncertainty across problems

- NLD is the main aligned masked-diffusion control.
- Sigma main-table errors: mean ± 2 SE over 10 sampling lanes.
- Quality tables do not match end-to-end inference latency.

Paper, Apps. F, G.1, G.8, H.2; PDF pp. 31–34, 42–44, 55.

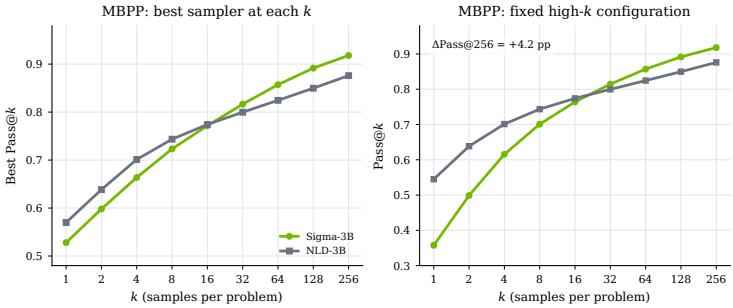
Base quality is close to NLD, with task-level tradeoffs

Diffusion mode	GSM8K	HumanEval	MBPP	Six-task avg.
NLD-3B	83.40	56.10	57.60	64.74
Sigma-3B, DDPM-256	84.50	59.45	52.32	64.34
Sigma-3B, DDIM-32	81.91	56.59	47.26	60.45
NLD-8B	87.26	62.80	67.00	70.22
Sigma-8B, DDPM-256	89.07	61.40	64.78	69.67
Sigma-8B, DDIM-32	87.78	58.54	59.54	66.72

- At 3B, stronger HumanEval accompanies weaker MBPP.
- At 8B, strong math scores coexist with a slightly lower overall mean.
- Reducing diffusion steps from 256 to 32 loses quality, especially in coding.

Paper, Tables 1–2, PDF p. 8. Percent, higher is better. Average also includes Minerva, HumanEval+, MBPP+.

MBPP shows a quality–coverage tradeoff



- Fixed configuration: pass@256 is 91.8% vs. 87.6%; paired CI for the gain: [+1.7,+6.7] pp.
- Exact-string repetition: 10.4% vs. 53.1% in the selected configurations.

Paper, Fig. 16 / App. G.8.1, PDF pp. 42, 51. Sigma: DDPM-256; NLD: 32 transfer steps.

High-pass@k gains depend on the task and control

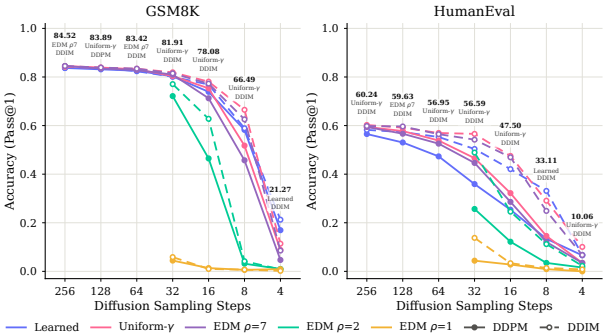
Task	High-budget result	Interpretation
Geometry	About 95.4% vs. 94.2% at $k = 512$	CI spans zero. At matched pass@1, Sigma ties best-tuned NLD.
GSM8K	1314 vs. 1305 of 1319 solved at least once	Small hard-tail gain; pass@256 difference CI excludes zero.
AIME 2025	63.3% vs. 53.3% at $k = 256$	Only 30 problems. Best-pair CI spans zero.

What pass@k tells us

Additional samples can expand correct-solution coverage.
Coverage depends on sampler tuning and the per-sample compute budget.

Paper, Tables 17–19, Figs. 17–19; PDF pp. 43–44, 51–52. 3B base / 3B instruct as appropriate.

Low-step sampling favors DDIM and a log-SNR grid



Uniform-γ grid

$$\gamma_i = \gamma_{\max} + \frac{i}{N}(\gamma_{\min} - \gamma_{\max})$$

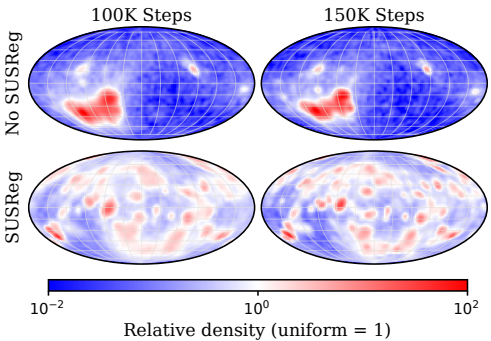
Training learns a variance-efficient schedule.

Inference chooses a finite integration grid.

These are different optimization problems.

Paper, Fig. 13; Sec. 5.1 / App. G.6; PDF pp. 8–9, 38–39, 49. CFG/ST tuned per configuration.

SUSReg reduces clustering of the token embeddings



- Unregularized embeddings use about 4 effective principal dimensions.
- SUSReg matches random 1D projections to a uniform sphere distribution.
- Less clustering may reduce token ambiguity under coarse updates.
- Validation PPL changes little; low-step generation improves.

Paper, Fig. 8 / Eq. (16), Sec. 5.2 / App. E.1; PDF pp. 9, 28–29. Random sphere projection; density=1 is uniform.

Embedding regularization helps most below 32 steps

DDIM steps per block	Sigma	+SUSReg	Gain (pp)
256	63.88	64.34	+0.46
32	60.41	61.85	+1.44
16	55.17	58.50	+3.33
8	37.06	42.76	+5.70
4	7.74	11.80	+4.06

- Six-task mean accuracy (%), 3B, block size 32, self-conditioning on.
- Guidance settings are selected for each model and step budget.
- Four-step decoding still needs trajectory distillation.

Paper, Table 8, PDF p. 33. Table 1 reports 60.45 for the ordinary DDIM-32 row; Table 8 reports 60.41.

PDD shares one backbone call across fine-grid updates

A segment of the teacher's integration grid

One student Transformer call $\longrightarrow$ Δ parallel prediction heads
$\longrightarrow$ Δ inexpensive DDIM updates $\longrightarrow$ next segment

$$\Delta = G/M, \qquad \hat{\mathbf{e}}_j^{S,\mathrm{raw}} = \exp(2\tanh \mathbf{s}_j) \odot \hat{\mathbf{e}}^{S,\mathrm{anchor}} + \mathbf{r}_j$$

- G: fine intervals; M: expensive student evaluations per block.
- Zero-initialized residual heads start from the pretrained prediction.
- Distill CFG into one conditional student call; apply ST analytically.
- Disable SC in teacher and student: local heads are predicted in parallel.

Paper, App. I.1–I.2, Eqs. (32)–(37), PDF pp. 57–58. Top flow is an explanatory schematic.

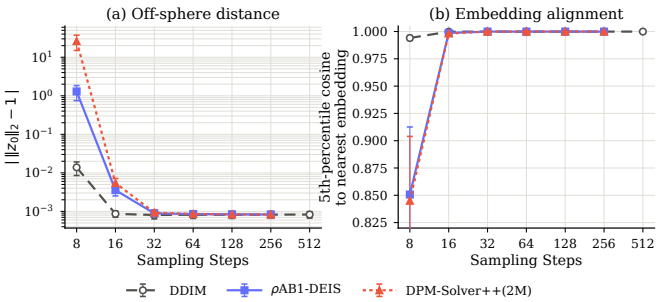
PDD recovers quality at four denoising calls

3B + SUSReg	GSM8K	HumanEval	MBPP	Six-task avg.
DDIM-16, SC	80.85	53.60	45.22	58.50
DDIM-8, SC	70.88	34.45	27.98	42.76
PDD-8, no SC	72.48	43.90	29.00	46.79
DDIM-4, SC	26.04	11.59	4.32	11.80
PDD-4, no SC	67.60	31.10	16.20	35.45

- The largest recovery occurs at four diffusion evaluations.
- Code remains sensitive to aggressive step reduction.
- PDD-4 uses distilled $(w, \tau) = (4, 0.4)$; PDD-8 uses $(3, 0.3)$.

Paper, Table 8, PDF p. 33; App. I, PDF pp. 57–60. Percent; each row uses its reported sampling recipe.

Higher-order integration can damage endpoint geometry



GSM8K, 16 calls

DDIM	78.08
DPM++(2M)	33.18
ρAB1-DEIS	48.96

Coarse extrapolation can place terminal states far from the vocabulary sphere.

Reconstruction is trained near the clean endpoint; off-manifold states are harder to decode.

Paper, Table 16 / Fig. 15, PDF pp. 40–41, 50. Uniform-γ, SC on; guidance selected for DDIM.

SFT preserves average parity, with different strengths

Ten-task mean (%)

Diffusion mode	3B	8B
NLD	52.89	59.92
Sigma DDPM-256	52.15	59.27
Sigma DDIM-32	49.86	57.01

Assistant-only supervision, joint diffusion + AR objective; geometry frozen by default.

8B DDPM-256 vs. NLD

Task	Sigma	NLD
GPQA	44.70	33.33
IFEval	79.14	73.02
MATH-500	78.88	84.60
AIME 2025	16.33	23.33
HumanEval	79.76	83.54

Paper, Tables 3–4, PDF p. 10; App. H, PDF pp. 54–56. Non-thinking protocol; IFEval uses instruction-level loose accuracy.

Count the readout before interpreting speed

$$\text{Forward calls per committed token} \approx \frac{N+1}{S}$$

N: diffusion updates per window; S: tokens committed from each 32-token window.

Default full-block decoding	Calls/block	Calls/token
DDPM-256	257	8.03
DDIM-32	33	1.03
DDIM-4	5	0.156

- One call processes a whole block and may contain two CFG branches.
- Small CFG overhead does not establish end-to-end speed over AR.
- The paper identifies high-step decoding speed and low-step code quality as limitations.

Paper, Algorithm 2 / Apps. B, G.5, G.10; PDF pp. 22, 27, 38–40, 44–45. Ratios computed from the reported formula.

Three connected contributions

1. **A scalable continuous language-model recipe**

   AR transfer, learned 16D embeddings, and a blockwise likelihood objective.

2. **A useful inference control space**

   CFG and ST trade single-sample quality for broader correct-solution coverage.

3. **A route toward few-step generation**

   Embedding geometry and trajectory distillation jointly improve low-NFE behavior.

The next bottleneck: robust latent-to-token decoding at low compute.

Synthesis of Secs. 3–7 and App. B, PDF pp. 3–10, 22.

Reference and discussion material

Zhihan Yang, Wei Guo, Jean-Marie Lemercier, Simon Welker, et al.

Large Language Continuous Diffusion Models.

arXiv:2610.02665v1, submitted 2 October 2026.

https://arxiv.org/abs/2610.02665v1

Companion material

Chinese reading notes, editable Beamer source, original figure assets, and the versioned source paper are archived together in the paper library.

All experimental results and original figures in this deck come from this paper. Backup slides follow.

Backup: unbiased training with one loss branch per block

$$\widehat{\mathcal{L}} = \mathcal{L}_{\text{prior}} + \frac{R}{p_{\text{rec}}}\mathcal{L}_{\text{rec}} + \frac{1-R}{1-p_{\text{rec}}}\mathcal{L}_{\text{diff}}, \quad R \sim \text{Bern}(0.1)$$

- Select reconstruction at $t = 0$ on 10% of blocks; otherwise sample a diffusion time.
- Inverse-probability weighting preserves the expected NELBO.
- All blocks fit into one forward with independent times and branch choices.

Two noise-schedule objectives

Endpoints γ_0, γ_1: optimize the likelihood bound.

Interior shape $\tilde{\gamma}$: reduce Monte Carlo variance through the second moment.

Paper, Eq. (5), Algorithm 1 / App. D; PDF pp. 3, 26–27.

Backup: why guidance and temperature commute

Set $\mu_t = \mathbf{z}_t/\alpha_t$. The Gaussian denoising identity gives

$$\mathbf{s} = \frac{\alpha_t}{\sigma_t^2}(\mathbf{e} - \mu_t).$$

Define $T(\mathbf{e}) = \mu_t + \tau^{-1}(\mathbf{e} - \mu_t)$ and $G(\mathbf{e}_c, \mathbf{e}_\varnothing) = w\mathbf{e}_c + (1-w)\mathbf{e}_\varnothing$. Then

$$T\big(G(\mathbf{e}_c, \mathbf{e}_\varnothing)\big) = wT(\mathbf{e}_c) + (1-w)T(\mathbf{e}_\varnothing).$$

The resulting score is

$$\frac{1}{\tau}(w\mathbf{s}_c + (1-w)\mathbf{s}_\varnothing) = \nabla_{\mathbf{z}_t} \log\Big[p_t(\mathbf{z}_t \mid \mathbf{c})^{w/\tau} p_t(\mathbf{z}_t \mid \varnothing)^{(1-w)/\tau}\Big].$$

This is a per-noise-level score identity; finite-step sampling remains approximate.

Paper, Proposition 1, Eqs. (28)–(29), PDF pp. 45–46.

Backup: SUSReg compares projected distributions

$$\phi_{\mathbf{a}_m}(u) = \frac{1}{V}\sum_{v=1}^{V} e^{iu\langle \mathbf{a}_m, \mathbf{e}_v\rangle}, \qquad \mathbf{a}_m \sim \mathrm{Unif}(\mathbb{S}^{15})$$

$$s(\mathbf{E}) = \frac{V}{M}\sum_{m=1}^{M}\int_{-u_{\max}}^{u_{\max}} |\phi_{\mathbf{a}_m}(u) - \phi_{d_e}(u)|^2 e^{-u^2/2}\, du$$

- ϕ_{d_e}: characteristic function of a uniform-sphere projection.
- 512 random directions; 17 quadrature points over $[-5, 5]$.
- Add $\lambda_{\mathrm{susreg}} s(\mathbf{E})$ per token, $\lambda_{\mathrm{susreg}} = 0.005$.
- Here u denotes Fourier frequency, distinct from diffusion time.

Paper, App. E.1, Eq. (16), PDF pp. 28–29. Integration variable renamed for clarity.

Backup: PDD trains on locally visited student states

$$\mathcal{L}_{\mathrm{PDD}} = \frac{1}{P} \sum_{p=1}^{P} \mathbb{E} \| \hat{\mathbf{e}}_{j_p}^{S,\mathrm{raw}} - \hat{\mathbf{e}}_{j_p}^{T,\mathrm{raw}} \|^2$$

$$\mathcal{L}_{\mathrm{distill}} = 5\mathcal{L}_{\mathrm{PDD}} + 0.1\mathcal{L}_{\mathrm{teacher-rec}} + 0.3\mathcal{L}_{\mathrm{teacher-ar}}$$

- Segment starts: forward marginal. Later nodes: student rollout.
- Frozen teacher predicts CFG targets before ST at those visited states.
- Forward KL terms preserve reconstruction and AR distributions.
- LoRA rank 128; 32 residual MLP heads; 10K steps; 512-token prefix.

Implementation ambiguity: App. I.5 reports $P = 8$ nodes per segment, but a 32-head grid at eight evaluations gives only four local nodes.

Paper, App. I.3–I.5, Eqs. (38)–(43), PDF pp. 58–60.

Backup: SFT geometry freezing is a simple default

3B instruct, ten-task mean	Frozen	Unfrozen
DDPM-256	52.15	51.65
DDIM-32	49.86	50.84

- Freeze embeddings, noise endpoints, and the schedule network.
- Train the Transformer, conditioning projections, and both output heads.
- Unfrozen geometry uses a $1000\times$ learning-rate multiplier.
- Ordering reverses across samplers; freezing is a simpler recipe.

SFT: 8K steps, batch 256, LR 2.5×10^{-6}; left-pad prompts during training so answers start on the 32-token grid.

Paper, App. H, Tables 21–22, PDF pp. 54–56.

Backup: guidance ablations support embedding-space CFG

Best score within each reported sweep

Setting	Embedding CFG	Logit CFG
GSM8K, DDIM-64, $\tau = 0.3$	83.19	83.49
HumanEval, DDIM-64, $\tau = 0.3$	56.22	53.78
GSM8K, DDPM-256, $\tau = 1$	82.43	76.76
HumanEval, DDPM-256, $\tau = 1$	46.95	29.70

- Logit CFG peaks early and degrades at stronger guidance.
- Retaining context but dropping SC is a weaker bad-prediction choice.
- Plain logit temperature helps, but adds no gain over tuned CFG+ST here.

Paper, Tables 9–13 / App. G.3–G.4, PDF pp. 34–37. Percent, higher is better.

Backup: partial commitment trades context for step count

Denoise 32 positions; commit only the first S; discard and re-noise the tail.

$$\text{Nominal NFE per 32 tokens} = \frac{32N}{S}, \qquad \text{actual calls} = \frac{32(N+1)}{S}.$$

3B sampling	N	S	Actual calls	Six-task avg.
DDPM, nominal 256	256	32	257	64.34
DDIM, nominal 256	64	8	260	67.00
DDIM, nominal 32	32	32	33	60.45
DDIM, nominal 32	16	16	34	61.82
DDIM, nominal 32	8	8	36	57.76

Smaller stride helps until it forces too few denoising steps per window.

Paper, Table 14 / App. G.5, PDF pp. 38–39. Actual calls calculated using the paper's counting rule.

Backup: initialization and likelihood scoring

MDLM versus AR warm start

3B base avg.	AR init	MDLM init
DDPM-256	64.34	64.67
DDIM-32	60.45	61.64

Same 300B-token continuation.

MBPP improves; HumanEval weakens.

Likelihood-based multiple choice

Diffusion avg.	Sigma	NLD
3B	68.40	68.03
8B	72.28	73.02

Sigma ranks with negative ELBO:

128 stratified time samples
plus 8 reconstruction samples.

Paper, Tables 5–7 / Apps. E.2, F, PDF pp. 29–32. These MMLU scores use candidate ranking, not instruct generation.