Consistent Diffusion Language Models

Exact posterior bridges for few-step discrete generation

Paper authors: Hasan Amin, Yuan Gao, Yaser Souri, Subhojit Som,
Ming Yin, Rajiv Khanna, and Xia Song

Purdue University and Microsoft

ICML 2026 | arXiv:2605.00161v2
Reading presentation | 13 October 2026

Train agreement across stochastic bridges,
then generate with fewer denoising evaluations.

Parallel updates need a model trained for large jumps

The opportunity

Diffusion can update many token positions in one network evaluation.

- High quality often needs hundreds of refinement steps.
- Fewer steps make each prediction harder.

The missing construction

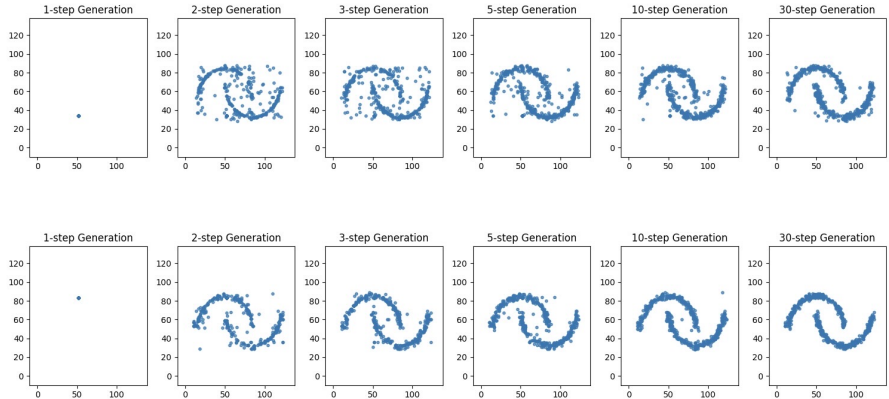
Continuous consistency uses a deterministic probability-flow ODE path.

- Categorical states evolve by stochastic jumps.
- CDLM uses exact posterior bridges to couple noise levels.

Question: can a single training stage build few-step denoising ability?

Paper, Secs. 1–2, PDF pp. 1–3. The ODE obstacle concerns the categorical sample space.

A discrete toy task shows structure emerging in fewer steps



Top: masked diffusion. Bottom: CDLM. Columns: 1, 2, 3, 5, 10, 30 steps.

Paper, Fig. 1, PDF p. 1. Quantized 2D moons; axes are quantized coordinates. Original figure crop.

The masked posterior bridge reveals a known fraction of tokens

Forward masking: a clean token survives at time t with probability $1 - t$.

$$q(x_s^i \mid x_t^i, x_0^i) = \begin{cases} \delta_{x_t^i}, & x_t^i \neq \texttt{[MASK]}, \\ \dfrac{t-s}{t}\delta_{x_0^i} + \dfrac{s}{t}\delta_{\texttt{[MASK]}}, & x_t^i = \texttt{[MASK]}. \end{cases}$$

Here δ_a denotes a point mass at token a.

Example: $t = 0.8 \to s = 0.4$

Each masked position reveals its true token with probability 0.5.
Already visible tokens stay fixed.

Training knows $\mathbf{x}_0$, so this intermediate state is sampled analytically.

Paper, Algorithm 2, PDF p. 14; general bridge in Eq. (2), p. 4. Numerical example is explanatory.

MPDC requires agreement after averaging over valid paths

$$p(\mathbf{x}_s \mid \mathbf{x}_t) = \mathbb{E}_{\mathbf{x}_0|\mathbf{x}_t}[q(\mathbf{x}_s \mid \mathbf{x}_t, \mathbf{x}_0)]$$

Multi-Path Discrete Consistency, per position i

$$f^*(\mathbf{x}_t, t)_i = \mathbb{E}_{\mathbf{x}_s \sim p(\cdot|\mathbf{x}_t)}[f^*(\mathbf{x}_s, s)_i] = [\mathcal{C}_{s \leftarrow t} f^*](\mathbf{x}_t, t)_i.$$

- Direct prediction agrees with the expected prediction after a cleaner hop.
- Bridges compose: marginalizing $t \to s \to u$ recovers $t \to u$.
- Individual sampled paths can have different predictions.

Paper, Lemma 3.1, Definitions 3.2–3.3, Eqs. (2)–(4), PDF p. 4.

The Bayes fixed point gives token marginals

$$f^*(\mathbf{x}_t, t)_i = p(x_0^i \mid \mathbf{x}_t), \qquad f^*(\mathbf{x}_0, 0)_i = \delta_{x_0^i}.$$

- The tower property establishes the fixed point under the true reverse chain.
- A positive data anchor and a mean-eliciting rule support identification.
- Factorized outputs do not represent every joint posterior.

Why multiple passes help: an illustrative two-token example

Data: AA or BB, each with probability $1/2$.
From all masks, both token marginals are $A/B = 1/2$.
Independent one-step sampling assigns 50% probability to AB or BA.

Paper, Proposition 3.4, PDF p. 4; Sec. 5, p. 9. Two-token example is our explanation.

Exact training pairs remove the need for a pretrained teacher

$$\mathbf{x}_0 \sim p_{\text{data}} \quad \longrightarrow \quad \mathbf{x}_t \sim q(\cdot \mid \mathbf{x}_0) \quad \longrightarrow \quad \mathbf{x}_s \sim q(\cdot \mid \mathbf{x}_t, \mathbf{x}_0).$$

$$\mathcal{L}_{\text{CDLM}} = \mathbb{E}\left[\frac{1}{\delta} \sum_{i \in \mathcal{M}(\mathbf{x}_t)} \text{JSD}(f_\theta(\mathbf{x}_t, t)_i, \text{sg}[f_{\tilde{\theta}}(\mathbf{x}_s, s)_i])\right].$$

- $\delta = t - s$; masked models train on masked positions.
- Uniform models include all positions.
- The target starts with the online network and follows an EMA.

At inference: predict clean token probabilities, then use an ancestral sampler.

Paper, Eq. (6), PDF p. 4; Sampling, p. 5; Algorithms 1–2, pp. 13–14.

A clean-data anchor prevents trivial self-consistency

$$\mathcal{L}_{\text{final}} = (1 - \kappa_{\text{ms}})\mathcal{L}_{\text{CDLM}} + \kappa_{\text{ms}}\mathcal{L}_{s=0}.$$

Cross-time tasks

Match predictions across several step sizes.
Weight each task by $1/\delta$.

Max-step tasks

Set $s = 0$ and use clean tokens as the target.
Ground the model in observed data.

Stop-gradient, a slow target network, and bounded symmetric JSD stabilize learning.

Paper, Sec. 3.2 and Eq. (7), PDF p. 5. Exact MDLM recovery additionally requires forward KL/CE.

The practical JSD loss differs from the ideal mean target

Let Q be the online prediction and P a random cleaner-state target.

$$\mathbb{D}(Q, \mathbb{E}[P]) \quad \text{and} \quad \mathbb{E}[\mathbb{D}(Q, P)]$$

are generally different objectives.

Forward KL has an exact mean-target decomposition

$$\mathbb{E}[\mathrm{KL}(P\|Q)] = \underbrace{\mathbb{E}[\mathrm{KL}(P\|\bar{P})]}_{\text{target variance}} + \mathrm{KL}(\bar{P}\|Q), \qquad \bar{P} = \mathbb{E}[P].$$

- JSD was selected for empirical stability; it generally has a different optimum.
- The KL excess-risk proof does not directly certify the JSD implementation.

Paper, Eqs. (5)–(6), p. 4; Remark C.2, PDF pp. 14–15. Interpretation of the objective distinction.

The main comparison uses 110M models on OpenWebText

- 150K training updates; global batch size 2048.
- Single-stage CDLM: no separate pretrained teacher.
- Distilled models: 100K pretraining + 50K distillation updates.
- AR: 75K updates, to match seen tokens according to the paper.

Family	Base models	Distilled comparison
Masked	MDLM, MCDLM	MDLM / MCDLM + SDTT
Uniform	UDLM, DUO, UCDLM	DUO + DCD

Matching updates does not match training FLOPs: target forwards and rollouts differ.

Paper, Secs. 4.1–4.2, PDF pp. 6–7; Appendix F.1, p. 20.

Quality requires reading PPL together with diversity

Unconditional generation

32 samples per model, each 1024 tokens.
Generative PPL from `gpt2-large` ↓; reported entropy checks repetition.

Conditional reconstruction

32 sequences per dataset; perturb 50% of tokens.
PPL ↓, BLEU2 against the original text ↑, and MAUVE ↑.

- OpenWebText in-domain; Lambada, WikiText-103, PTB out-of-domain.
- Keep FP32 and FP64 sampling results separate.
- NFE counts network evaluations; it does not measure wall-clock latency.

Paper, Sec. 4.3, PDF pp. 7–8; Tables 7–8, pp. 20–21.

Masked CDLM improves quality at every reported budget

Model	16 steps	64 steps	1024 steps
MDLM	297.1 / 5.6	124.4 / 5.6	97.7 / 5.6
MCDLM	125.6 / 5.4	67.7 / 5.6	53.4 / 5.5
MCDLM–PPL Opt.	**71.6 / 5.3**	**40.1 / 5.3**	**33.8 / 5.4**

FP64 ancestral sampling; entries: generative PPL ↓ / entropy.

- At 64 steps, base MCDLM improves PPL with the same reported entropy.
- At 16 steps, PPL Opt. reduces PPL by 75.9% relative to MDLM.
- PPL optimization introduces a quality–diversity tradeoff.

Paper, Table 1, PDF p. 7. All three models: 150K single-stage training updates.

CDLM competes with distillation and is a stronger starting point

Model	Training	16 steps	64 steps	1024 steps
MCDLM–PPL Opt.	150K	71.6 / 5.3	40.1 / 5.3	33.8 / 5.4
MDLM + SDTT	100K + 50K	76.0 / 5.4	40.1 / 5.3	31.2 / 5.4
MCDLM + SDTT	100K + 50K	**57.5 / 5.1**	**35.3 / 5.3**	**27.1 / 5.2**

FP64 ancestral sampling; entries: generative PPL $\downarrow$ / entropy.

- Single-stage PPL Opt. is competitive in the few-step regime.
- At 1024 steps, standard SDTT has lower PPL than single-stage PPL Opt.
- Applying SDTT to MCDLM improves on applying it to MDLM.

Paper, Table 1, PDF p. 7. Selected columns; entropy varies across configurations.

The acceleration claim is a reduction in network evaluations

A directly traceable FP64 comparison

AR: 1024 evaluations, PPL 40.2, entropy 5.6.
MCDLM–PPL Opt.: 64 evaluations, PPL 40.1, entropy 5.3.
16× fewer evaluations at similar generative PPL.

- MDLM at 1024 steps: 97.7 / 5.6; PPL Opt. at 16 steps: 71.6 / 5.3.
- This pair shows a 64× NFE reduction with lower PPL and entropy.
- Per-evaluation cost differs between diffusion and cached AR decoding.

Paper, Table 1 and Sec. 4.3, PDF p. 7. Authors report 64–128x vs MDLM and 16–32x NFE vs AR; no latency measurements.

Uniform CDLM also helps, with sampler-dependent tradeoffs

Model	Sampler	4 steps	64 steps
DUO	Ancestral	514.4 / 5.6	85.1 / 5.4
UCDLM	Ancestral	377.3 / 5.2	81.0 / 5.5
DUO + DCD	Ancestral	408.3 / 5.6	80.2 / 5.5
UCDLM	Greedy	110.4 / 4.8	64.7 / 5.5
DUO + DCD	Greedy	118.4 / 3.9	65.5 / 5.4

FP64; entries: generative PPL $\downarrow$ / entropy.

UCDLM beats the base uniform models, but the 64-step ancestral result is an exception to the text's claim of strictly beating DCD at all budgets.

Paper, Table 2, PDF p. 8; accompanying claim in Sec. 4.3, p. 7. Entropy below 5 is flagged by the authors.

Conditional reconstruction improves across four datasets

Model	OpenWebText	Lambada	WikiText-103	PTB
MDLM	40.4 / 61.1	62.9 / 57.7	39.6 / 62.2	220.0 / 50.7
SDTT	30.8 / 62.5	46.4 / 59.3	30.5 / 63.7	125.4 / 52.4
PPL Opt.	**28.6 / 62.9**	**38.8 / 60.1**	**27.3 / 64.4**	**107.5 / 53.3**

64 steps, FP64; entries: PPL $\downarrow$ / BLEU2 $\uparrow$.

- CDLM also leads SDTT at 8 and 512 steps on both metrics.
- PTB at 64 steps: 14.3% lower PPL, $+0.9$ BLEU2 points.
- Greedy DUO-DCD on PTB: PPL 45.0 but BLEU2 only 19.3.

Paper, Table 3 and Sec. 4.3.1, PDF p. 8. Protocol: reconstruct sequences after 50% token corruption.

Removing the data anchor creates misleadingly low PPL

Max-step weight	8 steps	64 steps	1024 steps
$\kappa_{\mathrm{ms}} = 0$	16.7 / 3.2	7.1 / 3.4	5.7 / 2.8
$\kappa_{\mathrm{ms}} = 0.4$	110.6 / 5.3	32.8 / 5.3	19.7 / 5.2
$\kappa_{\mathrm{ms}} = 1.0$	274.0 / 5.5	75.1 / 5.5	38.4 / 5.3

Entries: generative PPL $\downarrow$ / entropy.

A concrete failure mode

Self-consistency without grounding can collapse to repetitive outputs.
The smallest PPL in this table is not the desired model.

Paper, Table 5, PDF p. 9. Exact equivalence to MDLM at the max-step boundary requires forward KL/CE.

JSD gives the most useful stability–diversity balance here

Divergence	64-step PPL ↓	Entropy
JSD	32.8	5.3
Forward KL: $\mathrm{KL}(P_{\text{target}} \| Q_{\text{online}})$	8×10^4	6.9
Reverse KL: $\mathrm{KL}(Q_{\text{online}} \| P_{\text{target}})$	64.0	4.4

- Forward KL: the paper attributes failure to uniform drift.
- Reverse KL: low entropy is consistent with mode seeking.
- A favorable population target can still be hard to learn from scratch.

Paper, Table 6 and accompanying discussion, PDF p. 9. This supports the empirical choice of JSD.

Step-size schedules favor different inference budgets

Training schedule	8 steps	64 steps	1024 steps
Random	110.6 / 5.3	32.8 / 5.3	19.7 / 5.2
Staged increasing	160.8 / 5.2	45.1 / 5.3	25.3 / 5.2
Linear increasing	117.6 / 5.5	37.9 / 5.4	17.9 / 4.9
Linear decreasing	112.3 / 5.1	32.1 / 5.2	20.5 / 5.2

Entries: generative PPL ↓ / entropy.

Decreasing beats increasing at 8 and 64 steps; increasing has lower PPL at 1024 steps with lower entropy. Random remains slightly better at 8 steps by PPL.

Paper, Table 4, PDF p. 9. Training step size and inference step count are distinct quantities.

Additional results qualify the headline metrics

- **Precision matters.** At 1024 steps, MDLM PPL is 97.7 with FP64 versus 42.4 with FP32. Compare methods within the same precision.
- **MAUVE is not uniformly saturated.** At 8 steps on PTB, MDLM / SDTT / CDLM score 0.11 / 0.07 / 0.02.
  CDLM's PPL and BLEU2 gains do not imply a PTB MAUVE gain.
- **Few-step samples remain imperfect.** The 4-step CDLM example is fragmented; longer samples can still repeat themes and lose coherence.

Paper, Table 1, p. 7; Table 7, p. 20; Table 8, p. 21; qualitative examples, PDF pp. 22–24.

The contribution is a useful stochastic training principle

1. Exact posterior bridges provide cross-time supervision directly from data.
2. A data anchor, target network, and divergence choice make training usable.
3. Few-step gains are strong in this 110M setting and complement distillation.

Questions that determine the next step

Can mean-calibrated objectives be as stable as JSD?
Do gains persist with larger evaluations and models?
What is the latency and total compute at matched quality and diversity?

Synthesis of Secs. 3–5. Evaluation uses 32 generated samples per model; reported efficiency is NFE.

Reference and discussion material

Hasan Amin, Yuan Gao, Yaser Souri, Subhojit Som, Ming Yin, Rajiv Khanna, and Xia Song.

Consistent Diffusion Language Models.
ICML 2026. arXiv:2605.00161v2, revised 30 May 2026.

https://arxiv.org/abs/2605.00161v2

- Theory: Sec. 3 and Appendices C–E.
- Main results and ablations: Tables 1–6.
- FP32, MAUVE, and samples: Appendix F.

Paper-reading presentation; figure reproduced from the paper, tables selectively retypeset. Backup slides follow.

Backup: the general bridge is Bayes' rule in matrix form

$$q(x_s^i \mid x_t^i, x_0^i) = \mathrm{Cat}\left(x_s^i; \frac{(x_0^i Q_{1:s}) \odot (x_t^i Q_{s+1:t}^\top)}{\langle x_0^i Q_{1:t}, x_t^i \rangle}\right).$$

- $x_0^i Q_{1:s}$: prior probability of an intermediate token.
- $x_t^i Q_{s+1:t}^\top$: likelihood of the observed noisy token.
- Denominator: probability of the observation under the forward process.

$$q(\mathbf{x}_u \mid \mathbf{x}_t, \mathbf{x}_0) = \sum_{\mathbf{x}_s} q(\mathbf{x}_u \mid \mathbf{x}_s, \mathbf{x}_0) q(\mathbf{x}_s \mid \mathbf{x}_t, \mathbf{x}_0).$$

Paper, Lemma 3.1, Eq. (2), PDF p. 4; proof in Appendix E.1, pp. 18–19.

Backup: local errors accumulate along the true reverse chain

Let $E_{j|k}$ average over the true reverse transition on a time grid.

$$\mathbb{E}\, D_{\mathrm{TV}}(f_k, E_{k-1|k} f_{k-1}) \le \varepsilon \quad \Longrightarrow \quad \mathcal{E}(K, m) \le (K - m)\varepsilon.$$

- Semigroup composition, the triangle inequality, and convexity yield the bound.
- Under forward KL, tracking error is controlled by excess risk:

$$\varepsilon \le \sqrt{\tfrac{1}{2}\mathbb{E}[\mathcal{E}_{\mathrm{opt}}]}, \qquad \mathcal{E}(K, 0) \le K\sqrt{\tfrac{1}{2}\mathbb{E}[\mathcal{E}_{\mathrm{opt}}]}.$$

This controls predictor tracking under the true chain; it is not yet an error bound for the full joint distribution produced by the learned sampler.

Paper, Lemma C.1 and Remark C.2, PDF pp. 12–15. Uniform control is required across local edges.

Backup: the PPL-optimized variant changes three schedules

Base MCDLM

Random step size in $[1/8, 5/8]$; $\kappa_{\mathrm{ms}} = 0.4$; EMA rate 0.999.

MCDLM–PPLOptimized

Same configuration for the first 100K updates. Then:
step-size upper limit shrinks toward $3/8$; anchor weight declines to 0.2;
the target is hard-updated every 10K steps instead of using EMA.

Uniform CDLM

Step size increases linearly from 0.125 to 0.375.

Paper, Sec. 4.2, PDF pp. 6–7. All valid time pairs must satisfy $0 \leq s < t \leq 1$.

Backup: expected JSD need not recover the arithmetic mean

Our illustrative example:

$$P = \begin{cases} (1,0), & \text{probability } 0.9, \\ (0,1), & \text{probability } 0.1, \end{cases} \qquad Q = (q, 1-q).$$

- Minimizing $\mathbb{E}[\mathrm{KL}(P\|Q)]$ gives $q = 0.9$.
- Minimizing $\mathbb{E}[\mathrm{JSD}(P,Q)]$ gives $q \approx 0.998$.

$$\mathbb{E}[\mathrm{JSD}(P,Q)] = 0.9\, h\left(\frac{1+q}{2}\right) + 0.1\, h\left(\frac{q}{2}\right) - \frac{1}{2}h(q),$$

where h is binary entropy using natural logarithms.

Our scalar derivation and numerical minimization, not a paper experiment. Illustrates the distinction in Remark C.2.

Backup: the unifying view contains different kinds of relations

- **MDLM: exact objective recovery.** Set $s = 0$, use forward KL/CE, and weight by $-\alpha_t'/(1-\alpha_t)$.
- **Continuous consistency: a stochastic counterpart.** Replace the deterministic ODE coupling with the exact reverse-chain expectation.
- **Progressive distillation / shortcuts: composition.** Analytic bridges marginalize intermediate states instead of using explicit model rollouts.
- **Discrete distillation: alternative couplings.** SDTT uses teacher predictions; DUO-DCD shares continuous latent noise before projection.

Paper, Sec. 3.3, PDF pp. 5–6; Appendix D, pp. 15–18. Coupling differences alone do not establish the cause of low entropy.