LangFlow: Continuous Diffusion Rivals Discrete in Language Modeling

Paper discussion • 25-minute main presentation

Paper authors: Yuxin Chen, Chumeng Liang, Hangke Sui, Ruihan Guo, Chaoran Cheng, Jiaxuan You, and Ge Liu

University of Illinois Urbana-Champaign

arXiv:2604.11748v3 • April 15, 2026
Reading prepared October 14, 2026

arxiv.org/abs/2604.11748v3

Continuous diffusion becomes competitive at 130M

Central result

Token cross-entropy, an information-based noise schedule, and self-conditioning make embedding diffusion competitive with strong discrete DLMs.

Validation PPL ↓	AR	MDLM	LangFlow
LM1B	22.8	31.0	30.0
OpenWebText	17.5	23.2	24.6

DLM numbers are upper bounds. The claim is competitiveness with diffusion baselines at this scale; AR retains better in-domain PPL.

Paper, Sec. 5, Table 2, PDF pp. 9–10.

A dense state still needs a categorical learning signal

Representation choice

- One-hot / simplex state: vocabulary-sized and sparse.
- Embedding state: $\mathbf{z} \in \mathbb{R}^{L \times D}$, with $D = 768$.
- A continuous trajectory can support ODE solvers and future distillation.

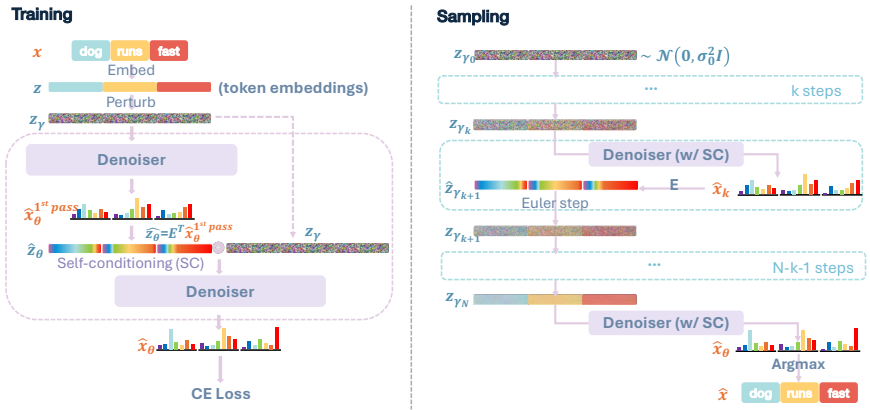
Three obstacles

1. What objective should learn the continuous flow?
2. How can its likelihood be evaluated?
3. Where should training and sampling spend their steps?

The network still predicts $L \times V$ token probabilities: the smaller state does not remove the vocabulary output layer.

Paper, Secs. 1 and 3.1, PDF pp. 1–5.

Predict tokens, then move through embedding space



Train on noisy embeddings with CE. During generation, turn probabilities into an expected embedding and update the entire sequence.

Paper, Fig. 1, PDF p. 6. Original author figure.

The noise coordinate makes denoising difficulty explicit

$$\gamma = \log \frac{\sigma_\gamma^2}{\alpha_\gamma^2}, \qquad \mathbf{z}_\gamma = \alpha_\gamma \mathbf{z} + \sigma_\gamma \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$$

$$\sigma_\gamma^2 = \mathrm{sigmoid}(\gamma), \qquad \alpha_\gamma^2 = \mathrm{sigmoid}(-\gamma).$$

Direction matters

Large γ: noisy. Small γ: clean.

Generation travels from b to a, with $a < b$.

- Condition the model on log noise-to-signal ratio, rather than an arbitrary time index.
- Noise sampling density $\pi(\gamma)$ determines where training effort is spent.

Paper, Sec. 3.1, Eqs. (7)–(8), PDF p. 4.

Cross-entropy matches the token posterior

For a one-hot target $p = \mathbf{1}_{x^{(i)}}$ and prediction q,

$$\mathbb{E}[D_f(p, q) \mid \mathbf{z}_\gamma] = D_f(\mathbb{E}[p \mid \mathbf{z}_\gamma], q) + C.$$

- The best prediction is the posterior probability vector.
- With $f(p) = \sum_k p_k \log p_k$, Bregman divergence becomes KL; a one-hot target gives cross-entropy.

$$\mathcal{L}_{\mathrm{CE}} = \mathbb{E}_{\gamma\sim\pi,\mathbf{x},\boldsymbol{\epsilon}}\left[-\frac{1}{L}\sum_{i=1}^{L}\log \hat{x}_\theta^{(i,x^{(i)})}(\mathbf{z}_\gamma, \gamma)\right].$$

A consistency argument for posterior prediction, for a fixed representation; not an identity between CE and exact model NLL.

Paper, Sec. 3.1, Eqs. (9)–(18), PDF pp. 4–5.

Posterior probabilities determine the continuous velocity

$$\hat{\mathbf{z}}_\theta^{(i)} = \sum_{k=1}^{V} \hat{x}_\theta^{(i,k)} \mathbf{e}_k = \mathbf{E}^\top \hat{\mathbf{x}}_\theta^{(i)}.$$

The bridge to flow matching

Correct token posterior $\Rightarrow$ correct expected embedding
$\Rightarrow$ correct marginal velocity for an affine Gaussian path.

$$\mathbf{v}_\theta(\mathbf{z}_\gamma, \gamma) = \frac{\dot{\sigma}_\gamma}{\sigma_\gamma} \mathbf{z}_\gamma - \frac{\alpha_\gamma}{2} \hat{\mathbf{z}}_\theta(\mathbf{z}_\gamma, \gamma).$$

Dots denote derivatives with respect to γ. The soft embedding preserves uncertainty during refinement; argmax is deferred to the end.

Paper, Eq. (6), p. 3; Eq. (19), p. 5; Eqs. (37)–(38), p. 17.

Sampling integrates a soft prediction over each interval

Hold $\hat{\mathbf{z}}_k$ fixed within a solver interval. Then

$$d\left(\frac{\mathbf{z}_\gamma}{\sigma_\gamma}\right) = \hat{\mathbf{z}}_\theta \, d\left(\frac{\alpha_\gamma}{\sigma_\gamma}\right),$$

$$\boxed{\mathbf{z}_{k+1} = \sigma_{k+1}\left[\frac{\mathbf{z}_k}{\sigma_k} + \left(\frac{\alpha_{k+1}}{\sigma_{k+1}} - \frac{\alpha_k}{\sigma_k}\right)\hat{\mathbf{z}}_k\right].}$$

1. Initialize $\mathbf{z}_0 \sim \mathcal{N}(0, \sigma_b^2 I)$.
2. Predict token probabilities, embed them, and integrate toward a.
3. Run a final prediction and decode each position by argmax.

Paper, Appendix A.1–A.2, Algorithm 2 and Eq. (22), PDF pp. 15–16.

Likelihood needs a decoder and a volume correction

Use a noisy data endpoint $\mathbf{z}_a \sim \mathcal{N}(\alpha_a \mathbf{z}, \sigma_a^2 I)$ and integrate $a \to b$ for evaluation:

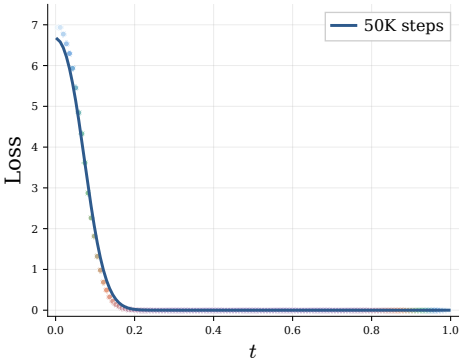
$$\log p(\mathbf{x}) \geq \mathbb{E}\left[\frac{LD}{2} - \frac{\|\mathbf{z}_b\|^2}{2\sigma_b^2} + \sum_i \log \hat{x}_\theta^{(i,x^{(i)})}(\mathbf{z}_a, a) - \int_a^b \frac{\alpha_\gamma}{2} \nabla_{\mathbf{z}_\gamma} \cdot \hat{\mathbf{z}}_\theta(\mathbf{z}_\gamma, \gamma)\, d\gamma\right].$$

- Gaussian energy + token reconstruction + continuous volume change.
- Negative ELBO gives an NLL upper bound, hence a PPL upper bound.
- Evaluation: 128-step Heun-2 and Hutchinson trace estimation.

The bound uses a probabilistic token decoder; sample generation uses argmax.

Paper, Theorem 3.1, Eq. (20), p. 6; Appendix C.4, p. 19.

Uniform time spends most steps on easy token recovery

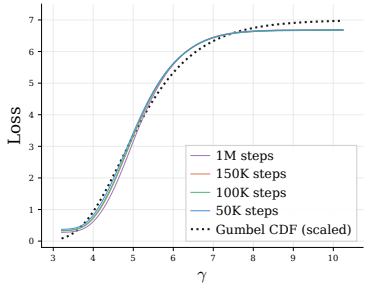


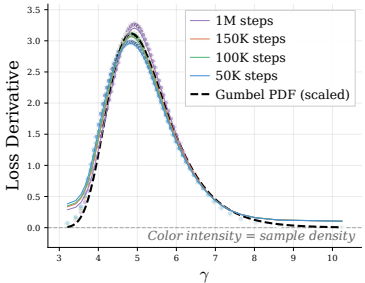
- CE is almost zero for $t > 0.2$ in the plotted convention.
- Finite, separated token endpoints can be identified while noise remains.
- Concentrate effort around the transition in information.

Paper, Sec. 4.1 and Fig. 2 left, PDF pp. 7–8. Original panel.

The caption's time-to-gamma formula has a sign inconsistency; generation here follows Algorithm 2: decreasing gamma.

Schedule density should follow the change in entropy





$$\ell_{\mathrm{CE}}(\gamma) = H_\gamma + \text{posterior KL} \;\approx\; H_\gamma, \qquad \pi(\gamma) \propto H'_\gamma.$$

Stable loss profiles motivate the entropy proxy. The derivative is right-skewed and fits a Gumbel density; this fit is empirical.

Paper, Sec. 4.1, Fig. 2 middle/right, PDF pp. 7–8. Original panels.

A learned Gumbel fit allocates both training and steps

$$F(\gamma) = \exp\left[-\exp\left(-\frac{\gamma - P_\mu}{P_\beta}\right)\right], \qquad H_\gamma \approx H_{+\infty} F(\gamma).$$

Training

Sample $\gamma \sim \mathrm{Gumbel}(P_\mu, P_\beta)$.

Fit entropy using squared error to detached CE. Detach sampled γ from the main model loss.

Sampling

Take descending quantiles:

$q_k = \mathrm{clip}(1 - k/N, \delta, 1 - \delta)$,

$\gamma_k = P_\mu - P_\beta \log(-\log q_k)$,

with $\delta = 10^{-5}$.

Equal CDF intervals approximate equal entropy reduction. The paper reports Gen. PPL falling from ~ 1000 to 154.2 after schedule improvements.

Paper, Sec. 4.1, Eq. (21), p. 8; Algorithms 1–2, p. 15.

Self-conditioning carries an earlier interpretation forward

Training: enabled with probability 0.25

1. Predict once from the current noisy state.
2. Embed and detach that prediction.
3. Predict again using it as an auxiliary input.

Otherwise, use a zero auxiliary input.

Sampling: always enabled

1. Start with a zero SC input.
2. Predict and integrate.
3. Reuse the preceding predicted embedding at the next step.

No second pass is added to every sampling step.

The auxiliary input enters through a zero-initialized projection before the DiT blocks.

Paper, Sec. 4.2, p. 8; Algorithm 1, p. 15; Appendix C.1, p. 18.

Self-conditioning changes the continuous model's ranking

LM1B setting	Gen. PPL ↓	PPL ↓
MDLM	103.9	31.0
MDLM + SC	94.9	32.7
LangFlow	154.2	49.0
LangFlow + SC	81.5	30.0

SC improves both metrics for LangFlow. MDLM's sample score improves while its validation PPL worsens.

Keep the reported results separate

This ablation reports 81.5 Gen. PPL; the main comparison reports 92.2. The paper does not explain the difference in protocol.

Paper, Table 1, PDF p. 8; contrast Table 2, p. 9, and Table 5, p. 20.

Two datasets test different sequence lengths

	LM1B	OpenWebText
Context length	128	1024
Tokenizer	BERT uncased	GPT-2 Large
Training steps / batch	1M / 512	1M / 512
Main sampling steps	128	1024
Generated samples	1024	1024

- Approx. 130M parameters; hidden size 768, 12 heads, RoPE.
- Validation PPL: likelihood on real data; DLM values are bounds.
- Gen. PPL: generated samples scored by GPT-2 Large.

Paper, Sec. 5, PDF p. 9. The text specifies 12 layers; Table 4 says 16.

Plaid uses its own 108M architecture and 16-dimensional embeddings. FLM uses 1024 LM1B sampling steps.

Validation PPL closes the gap with discrete diffusion

Model	LM1B PPL ↓	OWT PPL ↓
AR Transformer	22.8	17.5
MDLM	31.0	23.2
SEDD Absorb	32.0	24.1
Plaid	32.4	—
UDLM	33.8	27.4
Duo	33.6	25.2
LangFlow	**30.0**	**24.6**

Best reported DLM bound on LM1B; third-best on OWT. Different DLM bounds need not have the same tightness.

Paper, Table 2, PDF p. 9. Selected rows; all DLM PPLs are upper bounds.

OWT generation scores approach the AR reference

Model	LM1B Gen. PPL ↓	OWT Gen. PPL ↓
AR Transformer	66.7	35.9
MDLM	103.9	104.9
Plaid	77.3	—
Duo	97.6	77.6
FLM	96.9*	62.2
LangFlow	**92.2**	**36.5**

Best DLM sample score on OWT; second on LM1B. This measures a reference model's judgment of generated text, not the generator's likelihood.

Paper, Table 2, PDF p. 9. GPT-2 Large scores 1024 generated sequences.

* FLM: 1024 LM1B sampling steps; other listed diffusion models: 128. OWT: 1024 steps.

Zero-shot gains over AR occur on four of seven corpora

Target corpus	AR	MDLM	LangFlow
PTB	82.05	95.26	81.20
Wikitext	25.75	32.83	32.28
LM1B	51.25	67.01	68.21
Lambada	51.28	47.52	46.93
AG News	52.09	61.15	69.41
Pubmed	49.01	41.89	46.74
Arxiv	41.73	37.37	38.47

All are **corpus perplexities** after OWT training; Lambada is not a last-word accuracy result. LangFlow also beats MDLM on PTB, Wikitext, and Lambada.

Paper, Table 3, PDF p. 10. DLM perplexities are upper bounds.

Fewer steps preserve a lead, but quality still degrades

OWT budget	Gen. PPL ↓			LF entropy
Reported NFE	LangFlow	Duo	MDLM	
1024	36.53	77.69	104.85	5.25
512	41.52	78.14	104.43	5.31
256	49.24	78.62	112.70	5.37
128	60.09	80.02	120.77	5.43
64	80.34	85.62	143.88	5.51

One LangFlow checkpoint, no distillation or special few-step training. The gap to Duo narrows at 64 steps; no end-to-end speedup is measured here.

Paper, Appendix D.2, Table 6, PDF p. 21. Baselines quoted from Duo. Data entropy: 5.44.

Paper budget labels retained. Algorithm 2 also includes a final decoding forward pass.

Lower sample entropy leaves a quality question open

What entropy measures here

Token-frequency entropy within each sequence, averaged across samples. It mixes vocabulary diversity, topic concentration, and repetition.

- OWT at 1024 steps: LangFlow 5.25; data reference 5.44.
- Authors attribute part of the gap to content words recurring across sentences.
- A selected LangFlow example repeats *health* 11 times; this is not an aggregate repetition statistic.
- The shown texts still contain malformed words and semantic errors; the sample score alone does not establish coherence.

Paper, Appendix C.4, p. 19; Tables 6–7, p. 21; qualitative OWT samples, pp. 26–30.

The contribution is a coherent recipe for continuous DLMs

1. **A categorical objective can learn a continuous flow.**
   Posterior matching connects CE to an embedding denoiser.
2. **Noise geometry is modality-dependent.**
   An entropy-change proxy guides both training and sampling.
3. **Self-conditioning materially changes the comparison.**
   Its treatment in training and evaluation matters.

What remains to establish

Scaling beyond 130M, coherent long generation, and actual few-step latency gains. ODE distillation is a proposed next step, not a result of this paper.

Paper, Secs. 3–5 and 7, PDF pp. 4–11; Appendix D.2, pp. 20–21.

Reference and discussion prompts

Yuxin Chen, Chumeng Liang, Hangke Sui, Ruihan Guo, Chaoran Cheng, Jiaxuan You, and Ge Liu.

LangFlow: Continuous Diffusion Rivals Discrete in Language Modeling.
arXiv:2604.11748v3, April 15, 2026.

Paper: arxiv.org/abs/2604.11748v3
Author project: github.com/nealchen2003/LangFlow

- Does an entropy-change schedule retain its shape at larger scale?
- Can distillation preserve the quality–diversity balance?
- How should the SC history be represented in likelihood evaluation?

All experimental numbers in this presentation come from this paper. Backup slides follow.

Backup: the likelihood bound follows three steps

1. **Introduce a continuous latent endpoint.**

$$\log p(\mathbf{x}) \geq \mathbb{E}_{q(\mathbf{z}_a|\mathbf{x})}[\log p_a(\mathbf{z}_a) + \log \hat{p}(\mathbf{x} \mid \mathbf{z}_a) - \log q(\mathbf{z}_a \mid \mathbf{x})].$$

2. **Track the density along the ODE.**

$$\log p_a(\mathbf{z}_a) = \log p_b(\mathbf{z}_b) + \int_a^b \nabla \cdot \mathbf{v}_\theta \, d\gamma.$$

3. **Cancel the Gaussian scale terms.**

$$\nabla \cdot \mathbf{v}_\theta = LD \frac{\dot{\sigma}_\gamma}{\sigma_\gamma} - \frac{\alpha_\gamma}{2} \nabla \cdot \hat{\mathbf{z}}_\theta.$$

The $\log \sigma$ terms cancel against encoder entropy, leaving $LD/2$. A regular differentiable flow is required for the change-of-variables step.

Paper, Appendix B.1, Eqs. (30)–(43), PDF pp. 16–17.

Backup: the architecture adds three small mechanisms

1. **Input adaptation for SC:**

$$\mathbf{z}_\gamma \leftarrow \mathbf{z}_\gamma + W_{\text{in}}\mathbf{z}_\gamma + W_{\text{SC}}\mathbf{z}_{\text{SC}}.$$

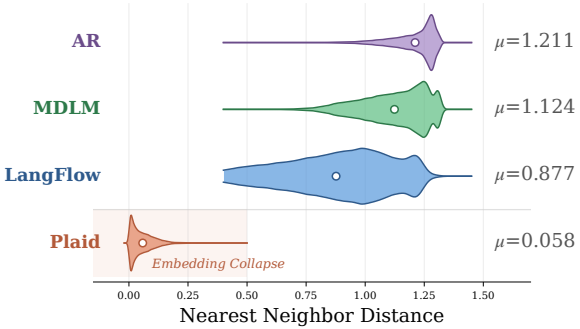
   Both projections are zero-initialized.
2. **Embedding normalization:** each embedding has norm $\sqrt{D}$.
3. **Analytic logit bias for candidate token k:**

$$r \log p(\mathbf{z}_\gamma^{(i)} \mid x^{(i)} = k) = r\frac{\alpha_\gamma}{\sigma_\gamma^2}\mathbf{e}_k^\top \mathbf{z}_\gamma^{(i)} + C.$$

   Ramp r from 0 to 1 during the first 5000 training steps.

Paper, Appendix C.1, Eq. (44), PDF pp. 17–18.

Backup: embedding geometry motivates avoiding MSE

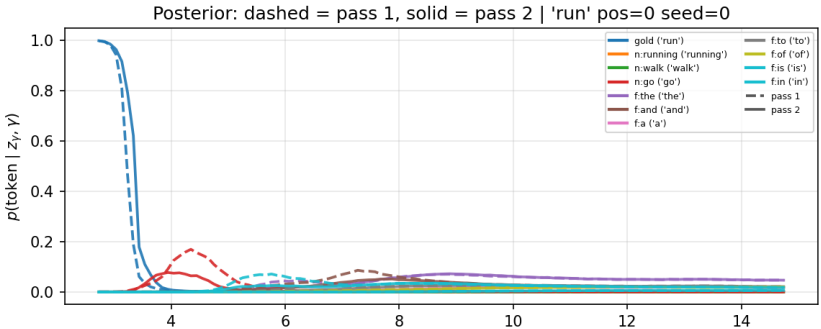


- Nearest-neighbor distance on normalized embeddings.
- Plaid's vectors are much closer together.
- MSE can reduce error by contracting the representation.

This is a cross-model observation: Plaid uses 16-dimensional embeddings; LangFlow uses 768. The comparison does not isolate loss choice.

Paper, Appendix C.3, Fig. 3 and Eqs. (45)–(46), PDF pp. 18–19. Original figure.

Backup: SC changes how probability shifts under noise



Higher γ moves from the correct token toward related words, then frequent function words. Dashed: pass 1. Solid: pass 2 with SC. This is one illustrative token position.

Paper, Appendix D.1, Fig. 4, PDF p. 20. Original figure.

Backup: LM1B budget and reproduction details

NFE	Gen. PPL	Entropy
128	92.24	4.31
64	104.83	4.32
32	127.32	4.33
16	179.60	4.35

Data entropy: 4.32.

No distillation.

- Table 1 gives 81.5 Gen. PPL; Tables 2/5 give 92.2/92.24.
- Sec. 5 and C.1 specify 12 layers; Table 4 specifies 16.
- SC uses history in Algorithm 2; the likelihood proof does not spell out its treatment in the state and divergence.

Paper, Table 5, p. 20; Table 1, p. 8; Table 2, p. 9; Table 4, p. 18; Appendix B.1, pp. 16–17.