ICML 2026 Workshop on Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning

Inverse-Confidence Sampling for Continuous Diffusion Language Models

Andrei Rekesh[1,2,3] **Jarrid Rector-Brooks**[1,4] **Cheng-Hao Liu**[1,5]

[1]California Institute of Technology, Pasadena, CA, USA [2]University of Toronto, Toronto, ON, Canada
[3]The Hospital for Sick Children, Toronto, ON, Canada [4]Mila – Quebec AI Institute, Montreal, QC, Canada [5]FutureHouse

Correspondence to: Andrei Rekesh <arekesh@caltech.edu>.

Abstract

In diffusion language models (DLMs), continuous-space DLMs (CDLMs) are narrowing the quality gap compared to masked diffusion models (MDMs). Despite this, CDLMs denoise tokens uniformly, leaving the design of sampling trajectories largely unexplored. We present a counter-intuitive inference-time procedure, termed INVERSE-CONFIDENCE SAMPLING (INCO). INCO allocates each step's denoising budget *inversely* proportional to the per-token confidence, which allows informative motifs to emerge early and anchor the remaining tokens. We apply INCO to LangFlow, a recent CDLM, and show that, without retraining, INCO substantially improves LangFlow's sample quality on images (MNIST) and natural language (LM1B, OpenWebText), reaching state-of-the-art metrics amongst comparable DLMs.

1. Introduction

Autoregressive (AR) language models commit to a left-to-right generation order, paying a sequential-inference penalty whose cost grows linearly in sequence length (Vaswani et al., 2017; Radford et al., 2019). Diffusion language models (DLMs) reframe text generation as iterative denoising of a full sequence, decoupling generation order from semantic order. Two families have emerged: *continuous embedding-space* DLMs (CDLMs) (Chen et al., 2026; Pynadath et al., 2025; Cheng et al., 2025; Austin et al., 2021; Campbell et al., 2022), which corrupt and denoise token embeddings under Gaussian noise, and *masked* (discrete) diffusion models (MDMs) (Sahoo et al., 2024; Shi et al., 2024; Nie et al., 2026), which mask and reveal tokens through an absorbing-state Markov chain on the vocabulary. Until recently, MDMs dominated the leaderboard and have been the top contender to outperform AR models, in part because of their training stability and in part because of a rich inference-time toolkit. The most recent CDLM, LangFlow, has now closed the quality gap, matching state-of-the-art MDMs on LM1B and OpenWebText through a Bregman-divergence flow-matching formulation, an information-uniform Gumbel noise schedule, and embedding-space self-conditioning (Chen et al., 2026).

The MDM inference-time toolkit now comprises a large body of work that remasks decoded tokens to enable iterative refinement and/or augments the MDM ELBO with a planner that reorders unmaskings by predicted information gain. The CDLM inference-time toolkit, however, remains nearly empty. Existing CDLM samplers integrate a probability-flow ODE (Song et al., 2020) along a *globally shared* noise schedule with a *uniform per-token* step size: at every Euler step, every position advances by the same amount regardless of how confident the denoiser already is at that position. The Gaussian forward process has no discrete mask state to commit to or revert from, so the natural MDM unmasking-order question ("which token to reveal next?") has no direct CDLM analogue. The closest theoretical work, entropic-time schedulers (Stancevic et al., 2026) and the optimal MDM schedules (Chen et al., 2025), prove that information-aware non-uniform schedules are optimal but operate at the *global* level, prescribing a single shared trajectory for all positions. Per-patch heterogeneous-noise schemes exist for images but require retraining with a heterogeneous-noise objective and an auxiliary difficulty head, and none have been brought to language (Schusterbauer et al., 2026).

We describe a CDLM sampling method based on an inference observation: the denoiser's per-position max-softmax confidence varies sharply across the trajectory and across positions (Appendix B). Under a uniform-step sampler, integration resolution would be wasted on positions that have effectively converged. Our contributions are as follows:

- We propose INVERSE-CONFIDENCE SAMPLING (INCO), a training-free, per-token sampling scheme

Accepted to FoGen 2026: Foundations of Deep Generative Models: Understanding Memorization, Generalization, and Reasoning, an ICML 2026 workshop (non-archival).

that allocates per-token step sizes *inversely proportional* to maximum predicted token logits.

- We find that INCO substantially improves Gen. PPL over LangFlow at every step count on language benchmarks while preserving entropy, and reduces FID by nearly half on MNIST relative to the base model.

2. Background

Flow Matching Flow Matching (FM) (Lipman et al., 2023) learns a velocity field $\boldsymbol{u}_t(\boldsymbol{z}_t)$ that transports samples from a simple prior p_0 (e.g. a standard Gaussian) to a typically complex data distribution p_1, such that integrating the ODE $\mathrm{d}\boldsymbol{z}_t = \boldsymbol{u}_t(\boldsymbol{z}_t)\,\mathrm{d}t$ from an initial $\boldsymbol{z}_0 \sim p_0$ yields a sample $\boldsymbol{z}_1 \sim p_1$. In practice, $\boldsymbol{u}_t$ is approximated by a neural network $\boldsymbol{v}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t)$ trained against a tractable conditional flow defined by an affine probability path:

$$\boldsymbol{z}_t = \alpha_t \boldsymbol{z} + \sigma_t \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

where the scalar schedules α_t and σ_t satisfy $\alpha_0 = 0$, $\sigma_0 = 1$ and $\alpha_1 = 1$, $\sigma_1 = 0$, so that $\boldsymbol{z}_0$ is pure noise and $\boldsymbol{z}_1$ is clean data. While natural, this construction couples both the training objective and the sampling dynamics to a particular choice of time schedule.

LangFlow. LangFlow (Chen et al., 2026) is an embedding-space continuous diffusion language model (DLM) that achieves near-MDM performance on a number of language modeling benchmarks. Three of its design choices matter for this work. First, in embedding-space DLMs the cross-entropy loss collapses almost everywhere except in a narrow high-noise band, so most of a uniform-in-t schedule is spent on noise levels that carry no information. LangFlow rectifies this by conditioning the denoiser on the log noise-to-signal ratio $\gamma = \log(\sigma_t^2/\alpha_t^2)$, which stretches the informative band over a wide range of γ ($+\infty$ at pure noise, $-\infty$ at clean data). Second, since each reverse step should be equally informative, LangFlow proposes the *information-uniform principle*: allocate noise density in proportion to the rate of information gain $H'_\gamma = \mathrm{d}H_\gamma/\mathrm{d}\gamma$, where $H_\gamma = \frac{1}{L}\sum_i H(x^{(i)} \mid \boldsymbol{z}_\gamma)$ is the average per-token posterior entropy at noise level γ. Third, H'_γ is empirically fit by a Gumbel density with learnable location and scale, trained jointly with the denoiser; at sampling time with N steps, the schedule $\{\gamma_k\}_{k=0}^N$ is set to the i/N quantiles of this Gumbel. INCO inherits this schedule and operates per-token within it.

Related work. Inference-time methods for diffusion language models have so far focused almost exclusively on the discrete (masked) regime. Path Planning (Peng et al., 2025) and ReMDM (Wang et al., 2026) introduce planner-based and remasking samplers, respectively, that allow MDMs to revise previously decoded tokens, while PAPL (Peng et al., 2026) demonstrates how to align training with planner-based inference. However, none of these approaches extend to continuous embedding-space DLMs. In the image domain, Patch Forcing (Schusterbauer et al., 2026) also adapts per-patch compute based on a learned uncertainty signal, but requires retraining with a Diffusion Forcing–style (Chen et al., 2024) heterogeneous-noise objective and a difficulty head; in contrast, INCO is fully training-free and applies to any pretrained CDLM that allows token logit prediction. Entropic time schedulers (Stancevic et al., 2026) are similar in spirit to LangFlow's Gumbel schedule, allocating sampling steps by total information gained per step, but operate as a single global schedule rather than at the token level. Self-supervised flow matching (Chefer et al., 2026) also varies the noise level across tokens, but does so during training to induce semantic representations; it is therefore orthogonal to the inference-time question we address.

3. Methods

Standard CDLM sampling, including LangFlow, applies a uniform step size $\Delta\gamma_k = \gamma_{k+1} - \gamma_k$ to every token at each Euler step. We observe that at any given step, the denoiser's predicted confidence (here, given by the value of its maximum output logit) varies substantially across positions: some tokens have effectively converged while others remain uncertain (Appendix B). Allocating equal budget to all tokens wastes integration resolution on converged positions, which, in domains like natural language and image, often do not provide valuable context to other tokens.

3.1. Inverse-Confidence Sampling

Inverse-Confidence Sampling is a per-token adaptive stepping scheme that reallocates the denoising budget toward tokens *with low confidence* while preserving the total integration volume prescribed by the noise schedule. The method modifies only the sampling procedure, has negligible computational costs, and requires **no retraining**.

Per Chen et al. (2026), sampling is solved over a finite range $[a, b]$ of the log noise-to-signal ratio $\gamma = \log(\sigma^2/\alpha^2)$, where a is the clean endpoint and b is the noisy endpoint. Given N Euler steps $k = 0, \ldots, N$, the standard schedule sets $\gamma_0 = b$ and $\gamma_N = a$, with intermediate $\gamma_k \in [a, b]$ distributed according to the Gumbel quantiles, and initializes $z_{\gamma_0} \sim \mathcal{N}(\mathbf{0}, \sigma_b^2\mathbf{I})$. Note that γ *decreases* from b to a over sampling.

Denoising volume. At Euler step k, the global noise schedule prescribes a transition from γ_k to γ_{k+1}, yielding

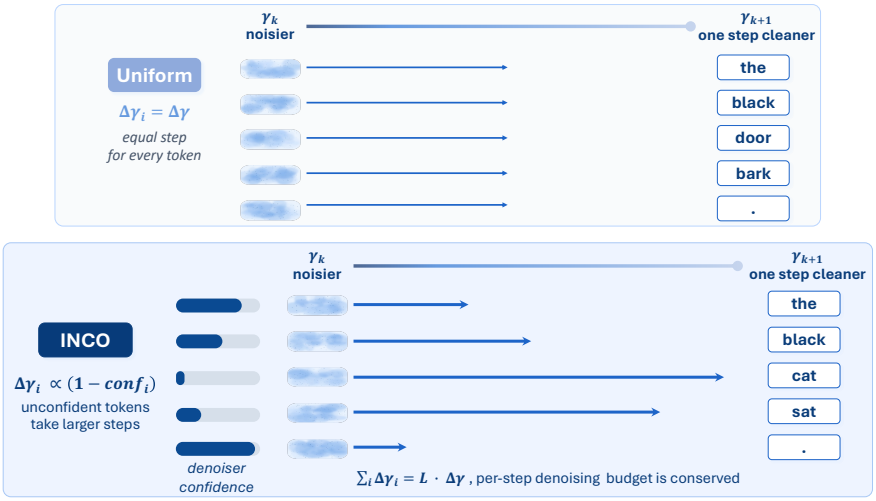


Figure 1. Schematic of uniform sampling in CDLMs and INCO. INCO is a simple inference-time modification with no retraining: at each step, it reallocates the fixed denoising budget toward low-confidence tokens. By resolving these informative motifs earlier, INCO provides scaffolding for the remaining positions and yields higher-quality sequences.

a uniform per-token decrement of $\Delta\gamma_k = \gamma_k - \gamma_{k+1} > 0$. We define the *denoising volume* at step k as the total budget $C_k = L \cdot \Delta\gamma_k$, where L is the sequence length. Under the standard scheme, each token consumes $C_k/L = \Delta\gamma_k$ of this budget.

Per-token weights. Let $\hat{x}^{(i)} = \hat{x}_\theta^{(i)}(z_{\gamma_k}, \gamma_k) \in \mathbb{R}^V$ denote the pre-softmax logits at position i. We first generate token candidate probabilities $p_k^{(i)} = \mathrm{softmax}(\hat{x}^{(i)}) \in [0,1]^V$ and obtain the maximum token probability $\max_v p_k^{(i,v)}$ for each token i. We then calculate a second softmax $u_k = \mathrm{softmax}(\max_v p_k^{(\cdot,v)}/\tau_{conf}) \in [0,1]^L$ over all maximum probabilities to obtain normalized weights proportional to denoiser "confidence"—here, τ_{conf} controls the sharpness of the confidence estimate. The step weight vector is then given by $w_k = 1 - u_k$. Under this weighting scheme, less "confident" tokens (according to $w_k^{(i)}$) are allocated proportionally larger step sizes.

Budget-constrained allocation. Each token i maintains its own noise level $\gamma_k^{(i)}$ and corresponding latent $z_k^{(i)}$, with $\gamma_0^{(i)} = b$ for all i. The sequence-level state at step k is $z_k = (z_k^{(1)}, \ldots, z_k^{(L)})$ and $\boldsymbol{\gamma}_k = (\gamma_k^{(1)}, \ldots, \gamma_k^{(L)})$, where the per-position noise levels are in general distinct under INCO. We define the *headroom* $h_k^{(i)} = \gamma_k^{(i)} - a$ as the largest decrement token i can take at step k without overshooting the clean endpoint a. We seek per-token step sizes $\delta_k^{(i)} \geq 0$ satisfying:

$$\sum_{i=1}^{L} \delta_k^{(i)} = C_k, \qquad \delta_k^{(i)} \leq h_k^{(i)}, \tag{2}$$

$$\delta_k^{(i)} \propto w_k^{(i)} \quad \text{among unclamped tokens.} \tag{3}$$

The first constraint conserves the denoising volume; the second prevents any token from overshooting a; the third ensures the allocation is proportional to the uncertainty weights among tokens that have not yet reached the clean endpoint.

We satisfy the constraints by finding a scalar $\lambda > 0$ such that

$$\sum_{i=1}^{L} \min\left(h_k^{(i)},\ \lambda\, w_k^{(i)}\right) = C_k. \tag{4}$$

The per-token step is then $\delta_k^{(i)} = \min\left(h_k^{(i)}, \lambda\, w_k^{(i)}\right)$. Since the left-hand side of (4) is piecewise-linear and monotonically increasing in λ, we solve for λ by bisection. Tokens for which $\lambda\, w_k^{(i)} \geq h_k^{(i)}$ are clamped to $h_k^{(i)}$, and the remaining budget is absorbed by the unclamped tokens in proportion to their weights.

Proposition 3.1 (Existence, uniqueness, and bisection convergence)**.** *Assume $w_k^{(i)} > 0$ and $h_k^{(i)} \geq 0$ for all i, and $0 \leq C_k < \sum_{i=1}^{L} h_k^{(i)}$. Then (4) admits a unique solution $\lambda^\star \in [0, \lambda_{\max}]$, where $\lambda_{\max} := \max_i h_k^{(i)}/w_k^{(i)}$. Bisection on $[0,\ C_k/\min_i w_k^{(i)}]$ produces iterates $\lambda^{(M)}$ satisfying $|\lambda^{(M)} - \lambda^\star| \leq C_k/(2^{M+1} \min_i w_k^{(i)})$.*

Proof. See Appendix A. □

Per-token Euler step. Given step sizes $\delta_k^{(i)}$, we set $\gamma_{k+1}^{(i)} = \gamma_k^{(i)} - \delta_k^{(i)}$ and update the latent per token: $z_{k+1}^{(i)}$ is computed from $z_k^{(i)}$ via the EDM Euler step of LangFlow , using the shared denoiser prediction $\hat{x}_\theta^{(i)}(z_k, \boldsymbol{\gamma}_k)$ and the per-token noise levels $\gamma_k^{(i)}, \gamma_{k+1}^{(i)}$. Self-conditioning

Algorithm 1 Uncertainty-Weighted Euler Sampling

Require: Model $\hat{x}_\theta$; endpoints a, b; steps N; noise levels $\{\gamma_k\}_{k=0}^{N}$ from Gumbel schedule; logit temperature τ_{logit}; confidence temperature τ_{conf}; bisection iterations M

$z_0^{(i)} \sim \mathcal{N}(\mathbf{0}, \sigma_b^2 \mathbf{I})$ for all i; $\gamma_0^{(i)} \leftarrow b$ for all i; $\hat{z} \leftarrow \mathbf{0}$

for $k = 0, \ldots, N-1$ **do**

  $\hat{x} \leftarrow \hat{x}_\theta(z_k, \boldsymbol{\gamma}_k)$ with self-conditioning input $\hat{z}$

  $p^{(i)} \leftarrow \mathrm{softmax}(\hat{x}^{(i)}/\tau_{logit})$ for each i

  $u^{(i)} \leftarrow \mathrm{softmax}(\max_v p_k^{(\cdot,v)}/\tau_{conf})^{(i)}$ for each i

  $w^{(i)} \leftarrow 1 - u^{(i)}$ for each i

  $h^{(i)} \leftarrow \gamma_k^{(i)} - a$ for each i

  $C_k \leftarrow L \cdot (\gamma_k - \gamma_{k+1})$

  Solve $\sum_i \min(h^{(i)}, \lambda\, w^{(i)}) = C_k$ for λ by bisection (M iterations)

  $\delta^{(i)} \leftarrow \min(h^{(i)}, \lambda\, w^{(i)})$

  $\gamma_{k+1}^{(i)} \leftarrow \gamma_k^{(i)} - \delta^{(i)}$

  $z_{k+1}^{(i)} \leftarrow \textsc{EulerStep}\big(z_k^{(i)}, \hat{x}^{(i)}, \gamma_k^{(i)}, \gamma_{k+1}^{(i)}\big)$ for each i

  $\hat{z}^{(i)} \leftarrow \mathbf{E}^\top \hat{x}^{(i)}$ ▷ update self-conditioning cache

end for

return $x^{(i)} = \arg\max\, \hat{x}_\theta^{(i)}(z_N, \boldsymbol{\gamma}_N)$

proceeds as in LangFlow, caching the embedded denoiser prediction $\hat{z}_\theta^{(i)} = \mathbf{E}^\top \hat{x}_\theta^{(i)}$ for use at the next step.

4. Experiments

4.1. Text Generation

We evaluate INCO sampling on the standard text benchmarks LM1B (Chelba et al., 2014) and OpenWebText (OWT) (Gokaslan et al., 2019) using checkpoints released by the LangFlow authors. For each dataset, we generate 5000 samples under uniform and INCO sampling and measure quality via generative perplexity (Gen. PPL) and sample entropy. Gen. PPL is computed by scoring generated sequences with GPT-2 Large (Radford et al., 2019) as the reference language model. LM1B experiments use 128 denoising steps; OWT experiments use 1024. All INCO results use confidence temperature $\tau_{conf} = 2.0$ (see Ablations).

Table 1 compares INCO against the LangFlow baseline and a range of discrete-MDM samplers (Lou et al., 2024; Sahoo et al., 2024; Wang et al., 2026) reported in the ReMDM paper. INCO shows highly competitive performance for CDLMs on OWT and LM1B, improving Gen. PPL over the LangFlow baseline at every step budget while preserving the sample entropy characteristic of LangFlow. . Evaluating the quality vs. diversity tradeoff at different temperatures, INCO dominates the Pareto frontier, displaying a $8-13\%$ improvement at matched entropy levels (Figure 2).

Beyond the aggregate metrics, we observe qualitative improvements at the sentence level. The differences range from local repairs to multi-sentence coherence that the base uniform sampler fails to maintain. Three example pairs from OWT are shown below, generated from identical noise seeds. Additional samples can be found in Appendix E.

INCO New York Jets: They've made their first rookie draft.

LangFlow New York Jets: They've had their second rookie breath back.

INCO but generally there are a lot of tools, courses, experiences, and processes

LangFlow but nevertheless there's a very much work, work and research

INCO She said: "The party simply made clear: Brexit is not going to go away. It's about standing on the side that we're staying in the EU."

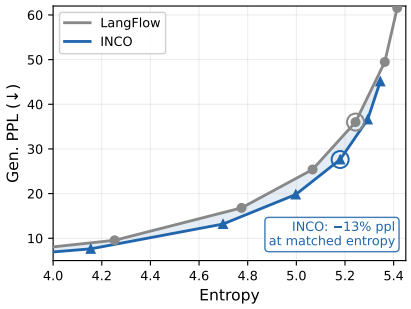
LangFlow She said: "The EU simply made sense because I actually allowed it to go away. It's so protected on the borders that we're running in the EU."

Table 1. Unconditional text generation on OWT (T=1024) and LM1B (T=512). Lower Gen. PPL is better; sample entropy reported for reference. Discrete-diffusion baselines marked † are taken from Wang et al. (2026); continuous-diffusion baselines marked ‡ are taken from Chen et al. (2026) and use 1024 generated samples (versus 5000 for unmarked rows). When a method appears in both, we use the Wang et al. (2026) numbers.

OWT		
Method	Gen. PPL↓	Ent.↑
Reference		
Data	14.8	5.44
AR†	12.1	5.22
MDM		
SEDD†	104.7	5.62
MDLM†	51.3	5.46
MDM + plan		
MDLM+FB†	33.8	5.35
MDLM+DFM†	21.7	5.20
ReMDM†	28.6	5.38
CDLM		
SEDD-U‡	103.6	–
Duo‡	77.6	–
FLM‡	62.2	–
LangFlow	<u>36.7</u>	5.25
INCO	**28.1**	5.18

LM1B	
Method	Gen. PPL↓
MDM	
SEDD†	115.9
MDLM†	103.9
CDLM	
Plaid‡	**77.3**
UDLM‡	99.8
Duo‡	97.6
FLM‡	96.9
LangFlow	93.3
INCO	<u>77.9</u>

(a) OpenWebText



(b) LM1B

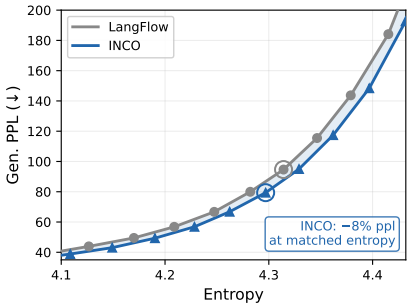


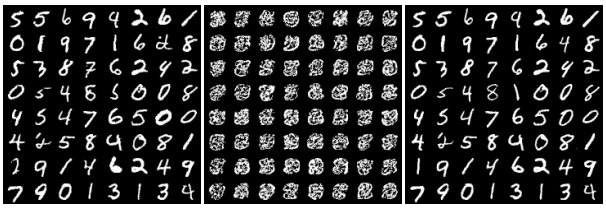
Figure 2. Quality–diversity trade-off of INCO. Generative perplexity (GPT-2 Large) vs. generative entropy as the logit temperature is swept. INCO(blue) dominates the LangFlow baseline (gray) everywhere, cutting Gen. PPL by $8-13\%$. Rings mark $\tau = 1.0$.

4.2. Image Generation

We adapt LangFlow to image generation by treating each 28×28 grayscale MNIST image (lec, 2019) as a sequence of $L = 784$ pixel tokens with vocabulary $|V| = 256$. The backbone is a DiT-style Transformer (8 blocks, 8 heads, hidden size 256, $\sim$5M parameters) operating on $D = 256$-dimensional normalized embeddings. The learnable Gumbel noise schedule and self-conditioning recipe ($p_{\text{sc}} = 0.25$) are kept identical to LangFlow. We train for 200,000 steps with AdamW ($\text{lr} = 10^{-3}$), EMA decay 0.9999, batch size 128, and evaluate with 128 Euler steps. Quality is measured over 10,000 generated samples each by Fréchet Inception Distance (FID) and Feature Likelihood Divergence (FLD) against the MNIST test set using DINO features. All MNIST-INCO results use confidence temperature $\tau_{conf} = 6.0$ (see Ablations). Table 2 compares vanilla LangFlow on MNIST against INCO; Figure 3 shows uncurated samples from each.

Table 2. Unconditional MNIST generation. INCO is applied at inference time only; both rows share the same checkpoint.

Method	FID (↓)	FLD (↓)
LangFlow (MNIST)	4.67	3.20
INCO (MNIST)	**2.68**	**1.08**



(a) Uniform. *(b)* Conf., $\tau = 6.0$. *(c)* INCO, $\tau = 6.0$.

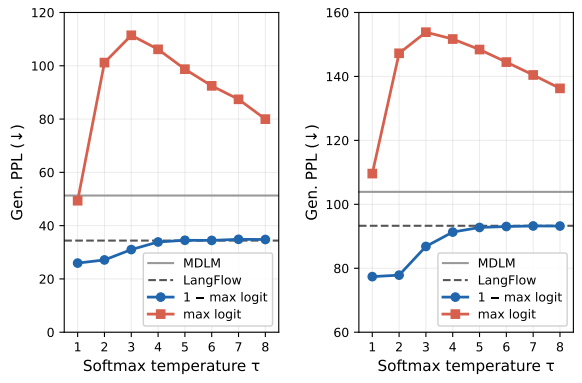
Figure 3. Uncurated MNIST samples from the same checkpoint and seed, generated with 128 Euler denoising steps. Left: the uniform Euler sampler. Middle: a confidence-proportional variant of INCO, which surprisingly collapses when denoising high-confidence tokens first. Right: the confidence-complement variant of INCO , which produces noticeably cleaner digits with fewer ambiguous strokes and broken structures.

4.3. Ablations

Confidence. We sweep $\tau_{conf} \in \{1, ..., 8\}$ for weighting proportional to $1 - \max_v p_k^{(i,v)}$ (low confidence tokens prioritized), and also proportional to $\max_v p_k^{(i,v)}$ (confident tokens prioritized). Across tasks, upweighting step sizes of low confidence tokens consistently improves over uniform sampling, while upweighting high confidence tokens by $\max_v p_k^{(i,v)}$ consistently degrades it (Figure 4 , Appendix D). This is unexpected, especially since recent MDM inference planning methods have shown improvements when prioritizing the unmasking of confident tokens (Peng et al., 2025). Empirical evidence suggests model confidence correlates with content complexity (Figure 11), and we hypothesize larger step size may help providing context for the remainder tokens.

For text, performance peaks at $\tau_{conf} = 2.0$; for MNIST, at $\tau_{conf} = 6.0$. We note that low temperatures improve Gen. PPL for both weighting schemes, but at the cost of significantly lower entropy when $\tau_{conf} < 1.0$. As a sanity check, we modulate the initial token candidate probabilities using logit temperature τ_{logit} (step 4 of Algorithm 1) such that $p_k'^{(i)} = \text{softmax}(\hat{x}^{(i)}/\tau_{logit})$ with uniform Euler step sizes for MNIST (modulating per-token step direction, rather than size) and find that performance declines at both lower and higher temperatures (Appendix D).

Number of denoising steps. Over different Euler step sizes ({8, 16, 32, 64, 128, 256, 512, 1024}, Appendix C), we compare uniform sampling against INCOat the benchmark-optimal temperature on LM1B ($\tau_{conf} = 2.0$), OWT ($\tau_{conf} = 2.0$), and MNIST ($\tau_{conf} = 6.0$). Compared to uniform sampling, INCO consistently demonstrates superior sample quality across all step counts.



(a) OpenWebText *(b)* LM1B

Figure 4. Confidence temperature ablation. Generative perplexity (GPT-2 Large, lower is better) as a function of τ for the two weighting schemes, with the uniform-Euler baseline shown as a dashed line. Prioritizing low confidence tokens (blue) improves over uniform sampling on both datasets; prioritizing confident tokens (red) degrades it.

Conclusion

In this work, we present INVERSE-CONFIDENCE SAMPLING (INCO), a training-free inference procedure that addresses the inefficiencies of uniform step sizes in CDLMs. We hypothesize that by reallocating the denoising budget inversely proportional to the denoiser's per-token confidence, INCO allows informative motifs to emerge early and anchor the generation process, reserving crucial integration capacity for highly uncertain tokens.

Empirically, this counterintuitive intervention substantially improves generation quality across modalities. Without requiring architectural modifications, auxiliary difficulty heads, or expensive retraining, INCO achieves state-of-the-art performance among comparable CDLMs on standard natural language benchmarks (OpenWebText, LM1B) and drastically reduces Fréchet Inception Distance on MNIST.

Ultimately, our results demonstrate that the sampling trajectories of CDLMs harbor significant untapped potential that can be unlocked entirely at inference time. As an inference method based on logit confidence, the utility of INCO inevitably depends on the dynamics of logit confidence over time in different domains. We hope INCO serves as a foundation for future exploration into dynamic, token-aware path planning in continuous diffusion models, closing the remaining flexibility gap with discrete masked diffusion paradigms.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

The mnist database of handwritten digits, 2019.

Austin, J., Johnson, D. D., Ho, J., Tarlow, D., and Van Den Berg, R. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

Campbell, A., Benton, J., De Bortoli, V., Rainforth, T., Deligiannidis, G., and Doucet, A. A continuous time framework for discrete denoising models. *Advances in Neural Information Processing Systems*, 35:28266–28279, 2022.

Chefer, H., Esser, P., Lorenz, D., Podell, D., Raja, V., Tong, V., Torralba, A., and Rombach, R. Self-supervised flow matching for scalable multi-modal synthesis. *arXiv preprint arXiv:2603.06507*, 2026.

Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., and Robinson, T. One billion word benchmark for measuring progress in statistical language modeling. In *Proc. Interspeech 2014*, pp. 2635–2639, 2014.

Chen, B., Martí Monsó, D., Du, Y., Simchowitz, M., Tedrake, R., and Sitzmann, V. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37: 24081–24125, 2024.

Chen, S., Cong, K., and Li, J. Optimal inference schedules for masked diffusion models. *arXiv preprint arXiv:2511.04647*, 2025.

Chen, Y., Liang, C., Sui, H., Guo, R., Cheng, C., You, J., and Liu, G. Langflow: Continuous diffusion rivals discrete in language modeling. *arXiv preprint arXiv:2604.11748*, 2026.

Cheng, C., Li, J., Fan, J., and Liu, G. α-flow: A unified framework for continuous-state discrete flow matching models. *arXiv preprint arXiv:2504.10283*, 2025.

Gokaslan, A., Cohen, V., Pavlick, E., and Tellex, S. Openwebtext corpus, 2019.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

Lou, A., Meng, C., and Ermon, S. Discrete diffusion modeling by estimating the ratios of the data distribution. *International Conference on Machine Learning*, 2024.

Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., Zhou, J., Lin, Y., Wen, J.-R., and Li, C. Large language diffusion models. *Advances in Neural Information Processing Systems*, 38:50608–50646, 2026.

Peng, F. Z., Bezemek, Z., Patel, S., Rector-Brooks, J., Yao, S., Bose, A. J., Tong, A., and Chatterjee, P. Path planning for masked diffusion model sampling. *arXiv preprint arXiv:2502.03540*, 2025.

Peng, F. Z., Bezemek, Z., Rector-Brooks, J., Zhang, S., Zhang, A. R., Bronstein, M., Tong, A., and Bose, A. J. Planner aware path learning in diffusion language models training. *International Conference on Learning Representations*, 2026.

Pynadath, P., Shi, J., and Zhang, R. Candi: Hybrid discrete-continuous diffusion models. *arXiv preprint arXiv:2510.22510*, 2025.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Sahoo, S. S., Arriola, M., Schiff, Y., Gokaslan, A., Marroquin, E., Chiu, J. T., Rush, A., and Kuleshov, V. Simple and effective masked diffusion language models. *Advances in Neural Information Processing Systems*, 37: 130136–130184, 2024.

Schusterbauer, J., Gui, M., Li, Y., Ma, P., Krause, F., and Ommer, B. Denoising, fast and slow: Difficulty-aware adaptive sampling for image generation. *arXiv preprint arXiv:2604.19141*, 2026.

Shi, J., Han, K., Wang, Z., Doucet, A., and Titsias, M. Simplified and generalized masked diffusion for discrete data. In *Advances in Neural Information Processing Systems 37*, 2024.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

Stancevic, D., Handke, F., and Ambrogioni, L. Entropic time schedulers for generative diffusion models. *Advances in Neural Information Processing Systems*, 38: 44222–44252, 2026.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, G., Schiff, Y., Sahoo, S., and Kuleshov, V. Remasking discrete diffusion models with inference-time scaling. *Advances in Neural Information Processing Systems*, 38:147282–147339, 2026.

A. Theory

Proposition 3.1 (Existence, uniqueness, and bisection convergence)**.** *Assume* $w_k^{(i)} > 0$ *and* $h_k^{(i)} \geq 0$ *for all* i*, and* $0 \leq C_k < \sum_{i=1}^{L} h_k^{(i)}$*. Then* (4) *admits a unique solution* $\lambda^\star \in [0, \lambda_{\max}]$*, where* $\lambda_{\max} := \max_i h_k^{(i)}/w_k^{(i)}$*. Bisection on* $[0,\, C_k/\min_i w_k^{(i)}]$ *produces iterates* $\lambda^{(M)}$ *satisfying* $|\lambda^{(M)} - \lambda^\star| \leq C_k/(2^{M+1}\min_i w_k^{(i)})$.

Proof. For Proposition 3.1. Fix a sampling step k and write $w_i := w_k^{(i)}, h_i := h_k^{(i)}, C := C_k$, $\lambda_i^\star := h_i/w_i$, $\lambda_{\max} := \max_i \lambda_i^\star$, and

$$f(\lambda) := \sum_{i=1}^{L} \min\big(h_i,\, \lambda w_i\big).$$

Step 1: Continuity. For each i, $\lambda \mapsto \min(h_i, \lambda w_i)$ is the pointwise minimum of two affine functions of λ and is therefore continuous. Hence f is continuous on $[0, \infty)$ as a finite sum of continuous functions.

Step 2: Monotonicity. For each i, $\lambda \mapsto \min(h_i, \lambda w_i)$ is non-decreasing on $[0, \infty)$ (it equals λw_i for $\lambda \leq \lambda_i^\star$, then is constant at h_i). Hence f is non-decreasing.

We show f is *strictly* increasing on $[0, \lambda_{\max}]$. Let $j \in \arg\max_i \lambda_i^\star$, so $\lambda_j^\star = \lambda_{\max}$. For any $0 \leq \lambda_1 < \lambda_2 \leq \lambda_{\max}$, both $\lambda_1, \lambda_2 \leq \lambda_j^\star$, so $\lambda_1 w_j, \lambda_2 w_j \leq h_j$ and

$$\min(h_j, \lambda_2 w_j) - \min(h_j, \lambda_1 w_j) = (\lambda_2 - \lambda_1) w_j > 0,$$

since $w_j > 0$. The remaining terms are non-decreasing, so $f(\lambda_2) - f(\lambda_1) > 0$.

Step 3: Existence. We have $f(0) = 0 \leq C$. At $\lambda = \lambda_{\max}$, every i satisfies $\lambda_{\max} w_i \geq \lambda_i^\star w_i = h_i$, so $\min(h_i, \lambda_{\max} w_i) = h_i$ and $f(\lambda_{\max}) = \sum_i h_i > C$ by assumption. By continuity (Step 1) and the intermediate value theorem, there exists $\lambda^\star \in (0, \lambda_{\max})$ with $f(\lambda^\star) = C$.

Step 4: Uniqueness. Suppose $\tilde{\lambda} \neq \lambda^\star$ also satisfies $f(\tilde{\lambda}) = C$. We show this leads to a contradiction.

- If $\tilde{\lambda} \in [0, \lambda_{\max}]$: strict monotonicity on this interval (Step 2) gives $f(\tilde{\lambda}) \neq f(\lambda^\star) = C$, contradicting $f(\tilde{\lambda}) = C$.
- If $\tilde{\lambda} > \lambda_{\max}$: then $\tilde{\lambda} w_i > h_i$ for all i, so $f(\tilde{\lambda}) = \sum_i h_i > C$, again a contradiction.

Hence $\lambda^\star$ is unique.

Step 5: Bisection convergence. Let $g(\lambda) := f(\lambda) - C$ and $w_{\min} := \min_i w_i > 0$. By Steps 1–3, g is continuous on $[0, C/w_{\min}]$ with $g(0) = -C \leq 0$ and $g(C/w_{\min}) \geq 0$. The standard bisection theorem then yields iterates $\lambda^{(M)}$ with

$$|\lambda^{(M)} - \lambda^\star| \;\leq\; \frac{C}{2^{M+1} w_{\min}}. \qquad \square$$

B. Confidence Variation Analysis

This section motivates INCO by showing that per-token confidence is highly non-uniform across positions and across the sampling trajectory. Trajectories are obtained from a pretrained LangFlow model sampled with T=128 ODE steps.

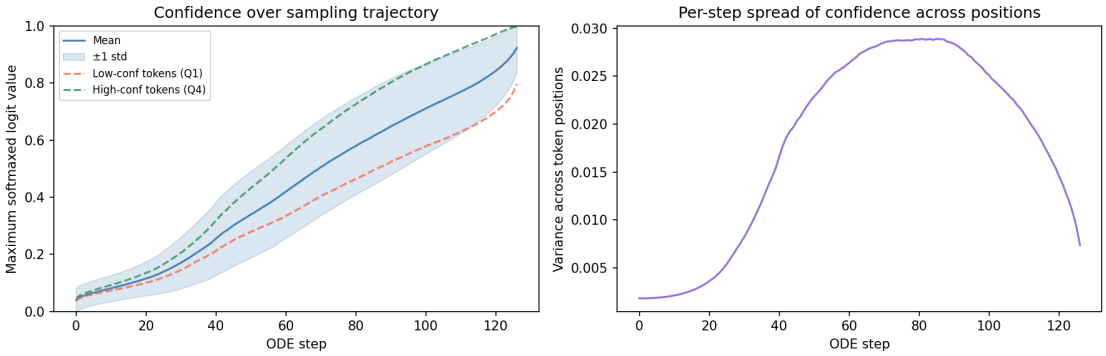


Figure 5. **Confidence statistics on OpenWebText.** Left: mean (solid) and ± 1 std (shaded) of per-token max-softmax confidence over the ODE trajectory. Confidence rises monotonically and Q1 lags Q4 by a substantial margin throughout sampling. Right: variance of confidence across token positions at each step, peaking in the middle of the trajectory.

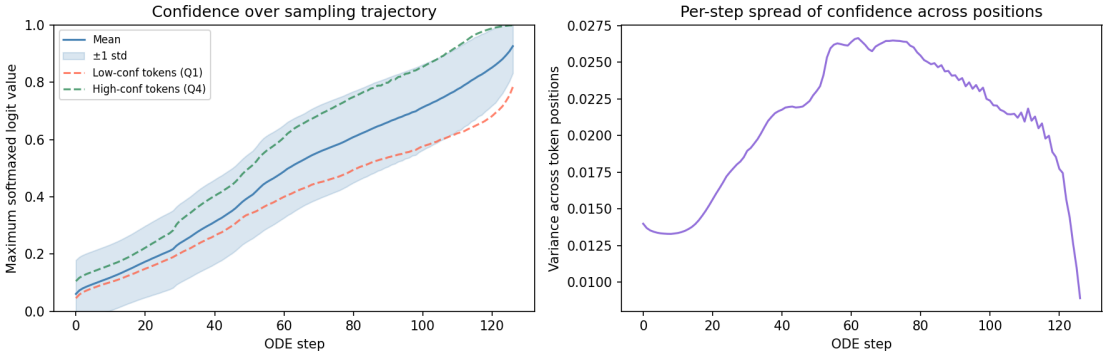


Figure 6. **Confidence statistics on LM1B.** Same axes as Figure 5. The same pattern holds: a persistent Q1–Q4 gap and a mid-trajectory peak in cross-position variance. The effect is qualitatively unchanged on the shorter LM1B sequences, suggesting it is a property of the diffusion process rather than of sequence length.

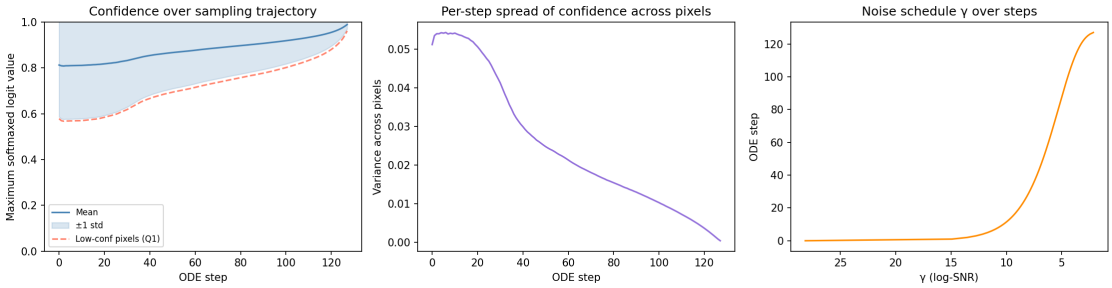


Figure 7. **Confidence statistics on MNIST.** Left: per-pixel confidence trajectory (mean ± 1 std and Q1). Middle: variance across pixels, which decays as the image resolves. Right: the noise schedule γ (log-SNR) used at inference. Unlike text, image pixels start with high mean confidence (background pixels are easy) and the bottleneck is concentrated on a small set of foreground pixels.

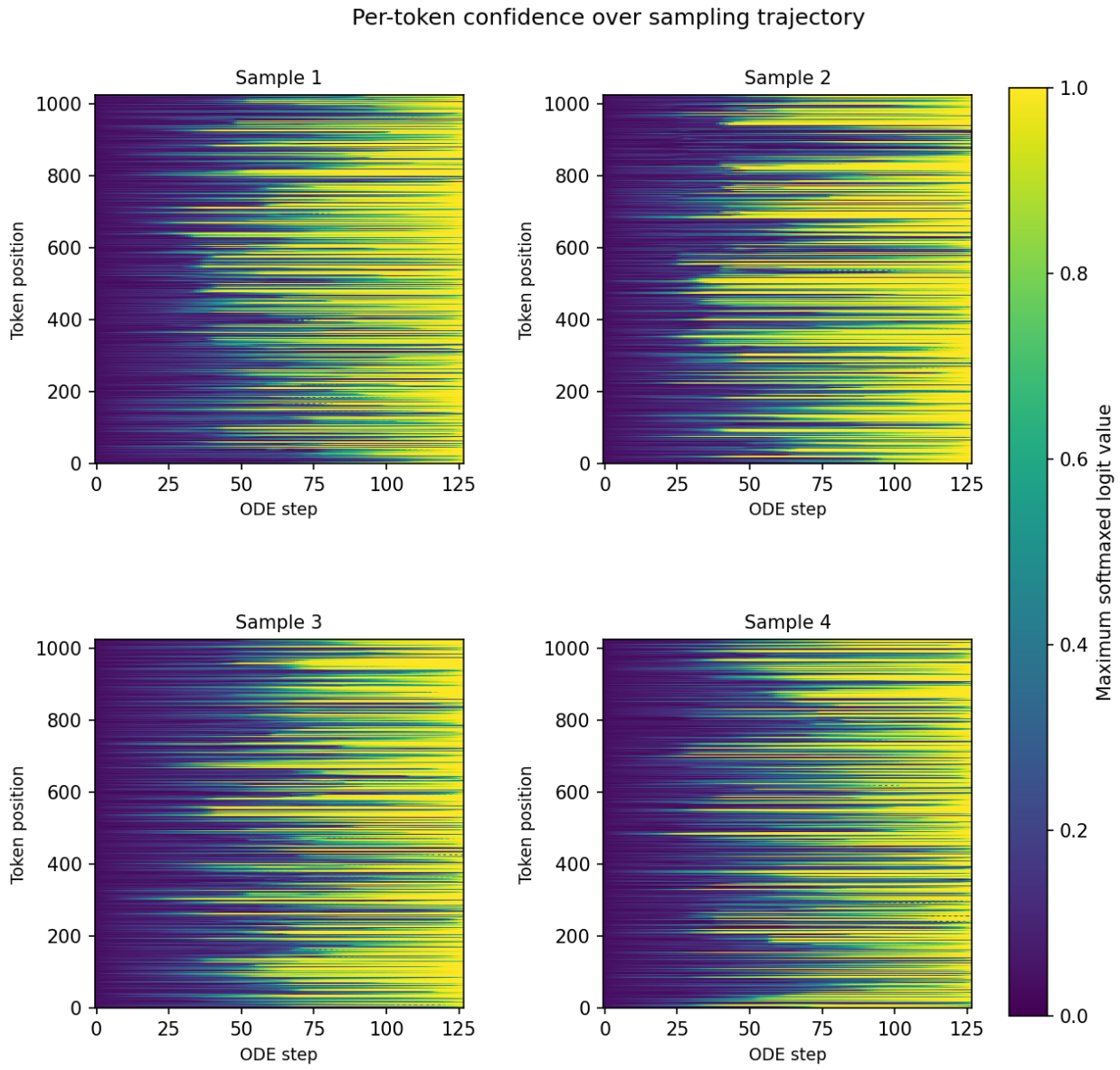


Figure 8. **Per-token confidence heatmaps on OpenWebText (4 samples).** Rows are token positions, columns are ODE steps; color encodes max-softmax confidence. Tokens resolve at very different times along the trajectory, with clear horizontal bands of late-resolving positions interleaved with positions that lock in early. A uniform denoising schedule treats all of these positions identically.

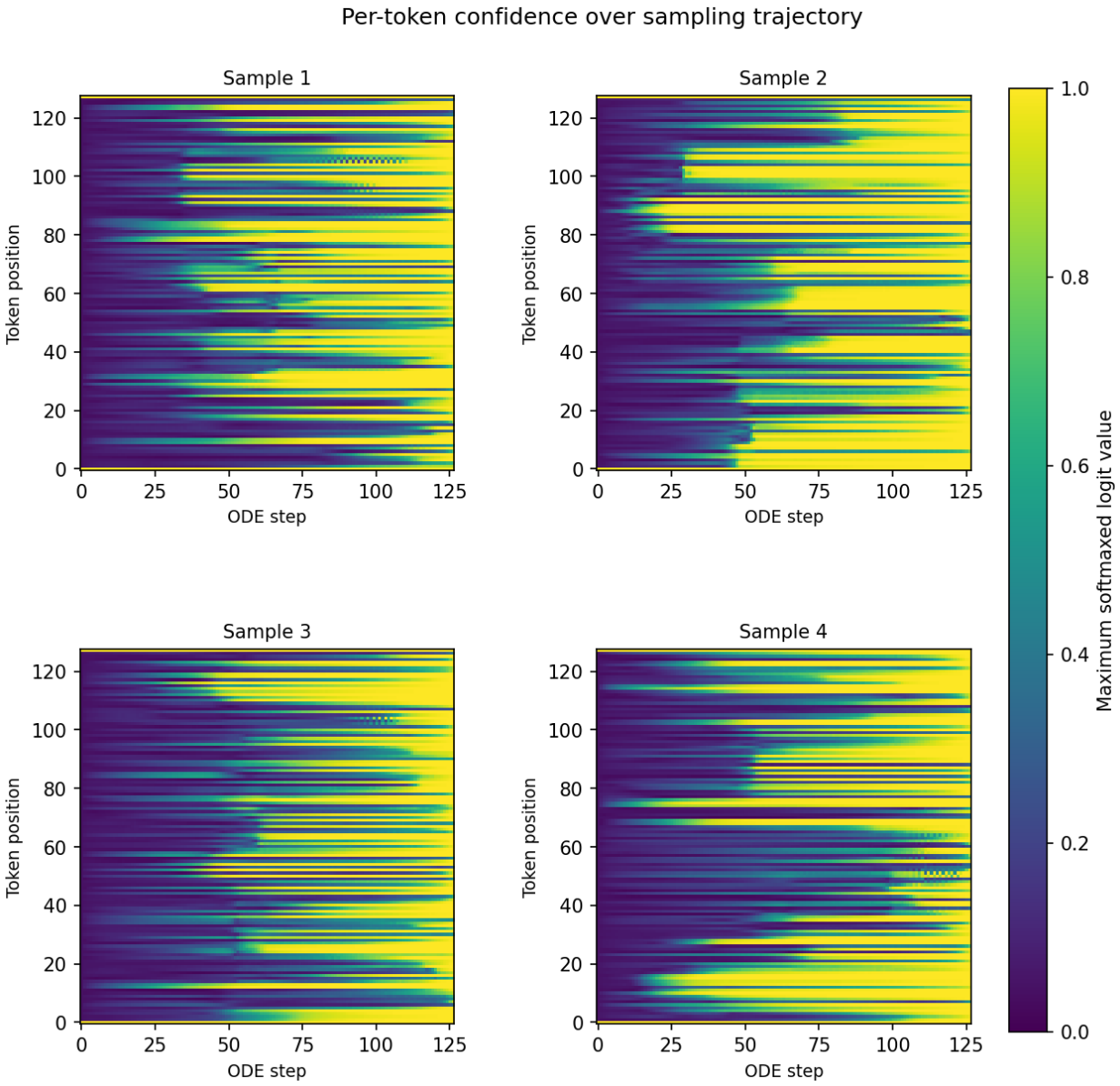


Figure 9. **Per-token confidence heatmaps on LM1B (4 samples).** Same visualization as Figure 8. The position-dependent resolution times persist on shorter sequences and across samples.

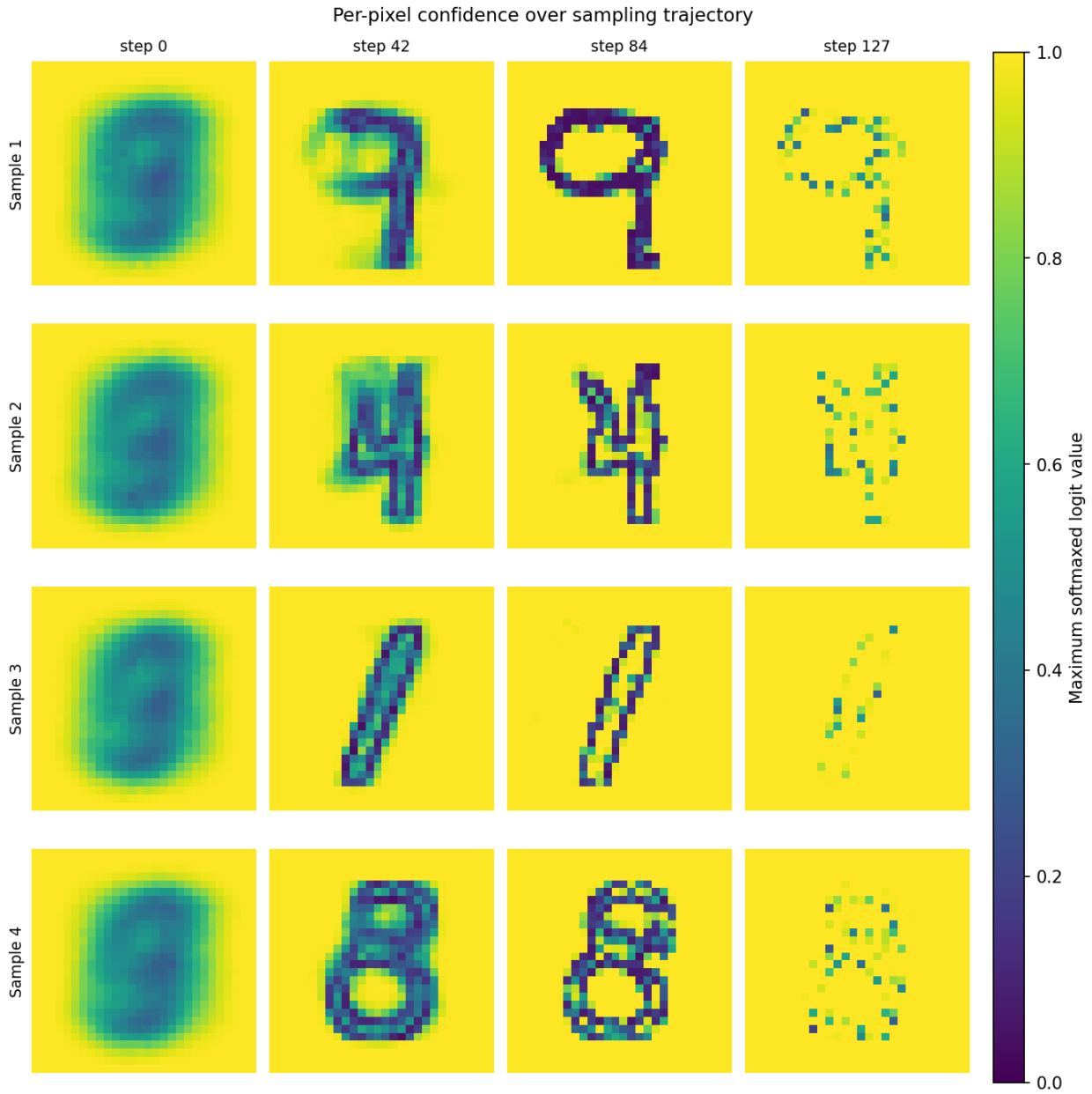


Figure 10. **Per-pixel confidence over the sampling trajectory on MNIST (4 samples).** Columns show steps 0, 42, 84, and 127. Background pixels are confident from the start, while the digit silhouette resolves progressively, with thin strokes and ambiguous junctions remaining low confidence until the very end. This spatial concentration of confidence is what INCO exploits.

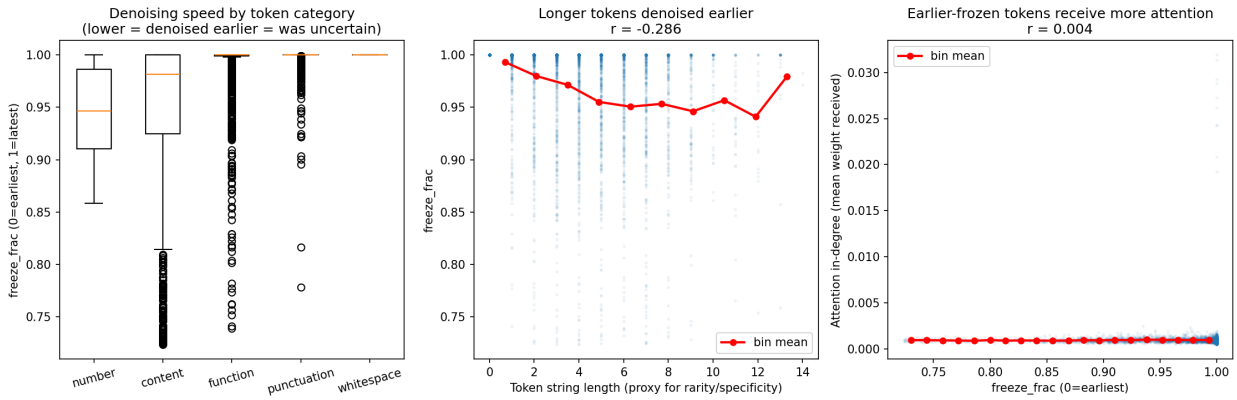
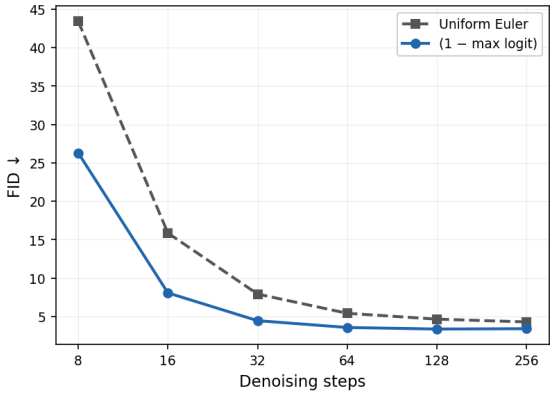
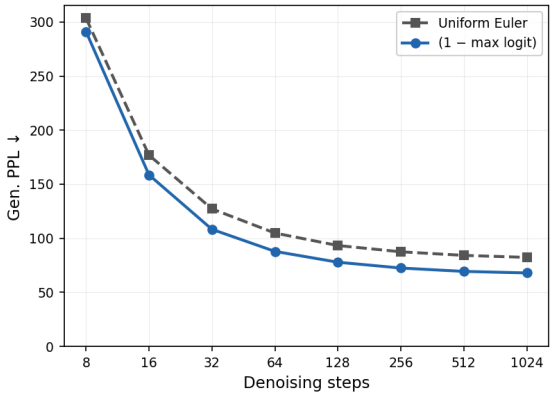


Figure 11. **What the model has low confidence about, on OpenWebText.** `freeze_frac` is the normalized step at which a token's confidence first crosses a fixed threshold (0=earliest, 1=latest). Left: by token category; numbers and content words are denoised earliest, while function words, punctuation, and whitespace lock in last. Middle: longer tokens (a proxy for rarer, more specific sub-words) are denoised earlier. Right: attention in-degree is essentially uncorrelated with freeze time, indicating the ordering reflects intrinsic predictive difficulty rather than how much each token is attended to. Together, these support the view that confidence tracks information content and that resolving it first is what allows INCO to anchor the rest of the sequence.

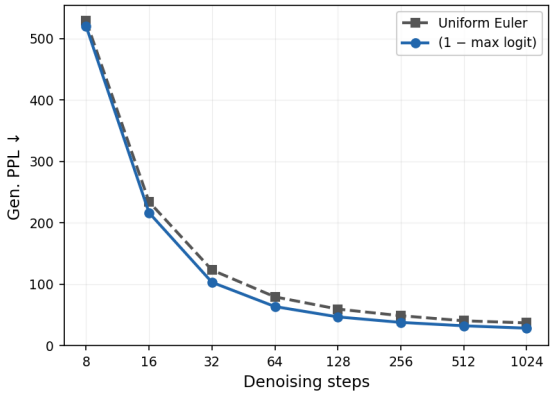
C. Step Count Ablations



(a) MNIST.



(b) LM1B.



(c) OWT.

Figure 12. Euler denoising step-count ablation across benchmarks (FID for MNIST, Gen. PPL for LM1B and OWT) under uniform sampling and INCO at the benchmark-optimal temperature. INCO improves over uniform sampling at every step count tested.

D. MNIST Ablation Discussion

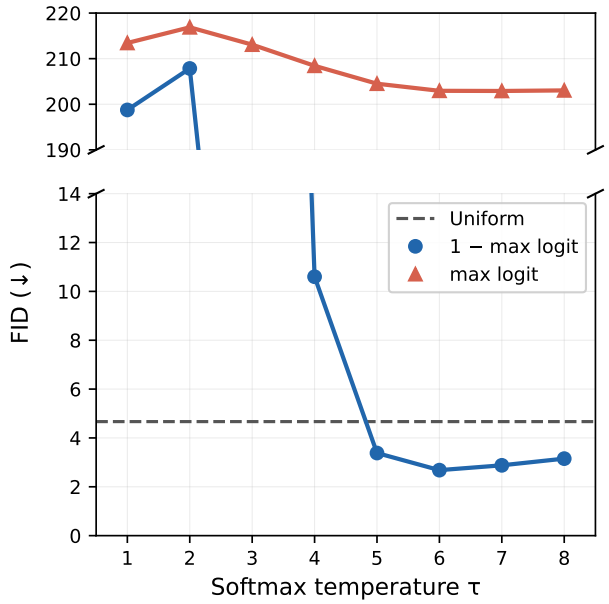


Figure 13. Confidence temperature ablation, MNIST. We hypothesize that at low τ_{conf}, the abundance of high-confidence background tokens forces λ to grow large and concentrate the entire budget C_k on a few uncertain tokens.

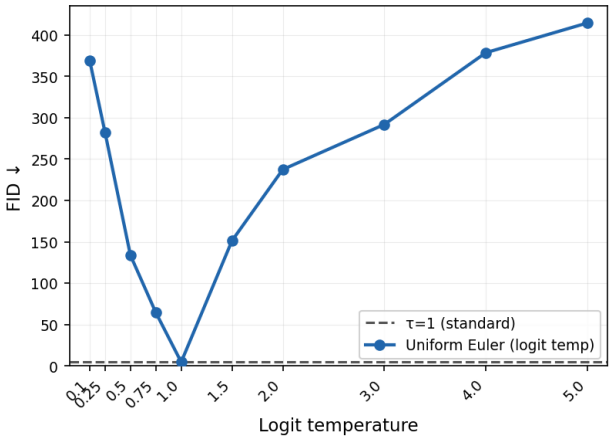


Figure 14. Logit softmax ablation with uniform Euler sampling step size (baseline method), MNIST. Sampling-time changes to τ'_{conf} (default = 1.0) in the baseline method severely degrade performance at both high and low values, verifying that trivial changes to Euler step direction is insufficient to produce improvement gains.

E. Additional Text Samples

We group qualitative comparisons by the type of failure exhibited by the uniform sampler. All pairs are unconditional generations from the same OWT checkpoint at $T{=}1024$ Euler steps.

E.1. Lexical and local repairs

We observe that INCO can improve locally implausible token (a duplicate, a polarity flip, a near-synonym in the wrong register).

INCO The EU is not alone facing restricted supplies of solar plants.
LangFlow The EU are not alone facing unlimited supplies of solar plants.

INCO There are certain key factors that seem to support robust growth.
LangFlow There are certain key applications that seem to support healthy jobs.

INCO It said close-up footage of the explosion had been viewed by activists and social media.
LangFlow It said three-scene footage of the explosion had been viewed by local state media channels.

E.2. Within-sentence coherence

We observe cases where INCOimproves sentences with tokens that are individually plausible but jointly inconsistent with the sentence's argument structure.

INCO Subjects that have started to become standardized will include classes offering specifics on how AI can be applied to and the skills needed to inform such concepts.
LangFlow Subjects that have started to become standardized will be courses offering specifics on how AI can be applied to and the level needed to inform such concepts.

INCO "It's not a bad question," McCannon added recently, "but of course that always leads to trouble. I'm just going to try to make people act and make better decisions to protect the people."
LangFlow "It's not a bad question," McCannon added recently, "because the reason that he leans to trouble is we're all going to try to make people exist and make better decisions to protect the people."

E.3. Cross-sentence and sustained coherence

We observe that INCOcan sometimes help maintain a more consistent stance in longer passages.

INCO Couchboy is one of my favorite video applications. It is a wonderful application for uploading all the features in the movie and it is a very simple media application. It is a fun application and allows you to download and edit each audio file in disk.
LangFlow Buddytie is one of my favorite UI applications. It is a wonderful application for displaying all the features in the project like everyone is shafting in a news conference. It is a fun application and allows you to download and edit each feature docs in dd.

INCO "When you're not fully up and working one full day a week, you expend a little time in doing what you need to do. I've training every day, and as much as my activities come from me, I really spend about three months cycling, swimming, playing video games, and doing things independently. As long as I do it and feel busy—waiting doesn't always limit the motivation on my stuff."
LangFlow "When you're not fully up and working one full day a week, you need to pick up a plan. I have 5 emails that I've constructed every day, and as much as my emails come from me, I can take out three months [a regular schedule] a year for the completed project plan. As long as I have all my emails ready [at 3] on a 'small' plan, I don't have to come up with an email telling me to drop whatever bug I want or take care of updates to me."

INCO Of course, that doesn't mean HTML5 is not fully compatible! That's probably the biggest obstacle to changing our style, but we'll never lose it.

LangFlow Of course, that doesn't mean scm.ts file is not automatically open! That's probably the easiest excuse to overwrite our own, but we'll never regret it.

INCO Researchers have long known that the sensory regions of our brain have a strong and abundant ability to both bring out the illusion of information, and that these regions cope with each sense of information. Now, a new study from scientists from the University of Chicago has found that, due to the connections within the brain, our frontal region play more fundamental roles in communication.

LangFlow Researchers have previously suggested that the sensory regions of our brain have a strong and abundant ability to both bring out the effects of information processing and that these regions cope with each realm of information. Now, a new study from economists from the University of Hollywood has found that, due to the connections within the brain, our frontal region play more fundamental roles in communication.

E.4. Shared failure modes

INCO redistributes integration budget but does not add model capacity. We observe many cases, mostly related to rare technical vocabulary, specific named entities, or quantitative claims outside the training distribution, where reallocating compute toward uncertain positions does not improve the generation. We highlight three representative cases.

INCO In addition, going according to vehicle behaviour, dogs were more than six times more likely to have noicative stimuli related to the mean numberal ant ($P < 0.001$).

LangFlow For example, dogs responded to the tail cage group with one or fewer responses that the stranger answered vice versa with responses to the predicted numberal ant ($p < 0.01$).

INCO "Senator, I did not want to support the selection of Jeff Sessions MORE (D-N.J.), but I knew he would already represent me on my committee, so I have begun debating the traditions of the Senate."

LangFlow "Jeff, I didn't want to support the selection of Andy Gorsuch MORE (D-N.D.), but I knew he would already represent me on my Committee, so I have begun debating the traditions of the Senate."

INCO Regibels isolate in room with body and waste xcerella Sringine and gelatinous worms from flavorometer (clone 2. T culture) for incubates at 37 °C after 15 min and bacterial culture reaction E7438.

LangFlow from garbage in bacteria in room slime bodies and waste xpcb spenine and gelatinary worms from pulled paper (made large. body culture) for incubating at 40 °C for 5 min and clagDOE #34s for reaction k-in 8.