UNDERSTANDING AND ENABLING SCORE DISTILLATION IN CONTINUOUS DIFFUSION LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Score Distillation (e.g., Diff-Instruct and Trajectory Distribution Matching) has substantially accelerated image and video diffusion models, yet transferring it to continuous diffusion language models requires more than reusing the training objective. We identify a central distinction in the relationship between denoising predictions and generated samples: in categorically parameterized language models, continuous distillation updates act on expected token embeddings, whereas generation ultimately requires discrete token decisions. Consequently, increasing prediction confidence can either improve generation or collapse different samples onto a small set of tokens. We propose a unified perspective: effective language distillation should resolve token-level uncertainty while preserving noise-dependent variation across generated samples. A denoiser's conditional mean should therefore not be assumed to provide a suitable generator initialization. This perspective unifies one-step endpoint matching and few-step trajectory matching, guiding generator initialization and score modeling of the student's generated distribution. Building on these principles, we adapt DI and TDM to continuous diffusion language models, enabling effective one-step and few-step generation on LM1B and OpenWebText. Our work clarifies the key considerations that distinguish distribution-matching distillation for language and provides reusable principles for transferring image-diffusion distillation methods to language generation.

1 INTRODUCTION

Diffusion language models offer a promising route to efficient text generation by updating an entire sequence in parallel. Their practical advantage, however, depends on how many rounds of denoising are needed to produce coherent text. Recent continuous diffusion language models (CDLMs), including FLM and LangFlow, demonstrate competitive language modeling by evolving continuous token representations (Lee et al., 2026; Chen et al., 2026). These advances make sampling efficiency a central challenge: realizing the promise of parallel generation requires compressing a lengthy denoising process into one or a few network evaluations while retaining generation quality.

In image generation, score distillation has made such compression possible. Methods including Diff-Instruct (DI), Distribution Matching Distillation (DMD), and Trajectory Distribution Matching (TDM) use a pretrained teacher's score to guide a fast generator toward the target distribution (Luo et al., 2023; Yin et al., 2024b;a; Luo et al., 2025c). This distributional supervision gives the student flexibility to learn shorter generation paths without reproducing every teacher step. CDLMs appear well suited to this paradigm because they share a continuous diffusion formulation. Yet score-based distillation remains underexplored in language, and the conditions that make it effective are poorly understood.

We find that a shared continuous state space does not make image distillation recipes directly transferable: straightforward application can produce severe token repetition in CDLMs. A fundamental difference lies in the prediction interface. Conventional image denoisers predict real-valued pixels or latent features, whereas CDLMs typically express continuous predictions through categorical token distributions. This interface creates two challenges: (1) token prediction can restrict score estimation for generated states, and (2) averaging across token identities can weaken the guidance needed

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

to generate distinct samples. Our theoretical and empirical analyses explain how these challenges affect distribution matching, establishing the requirements for effective CDLM distillation.

Building on this understanding, we introduce **Noise-Interpolated Perturbations for Score Distillation (NIPSD)**, a general framework for one-step and few-step CDLM generation. Its guiding principle is to preserve variation across generated samples as token predictions become more decisive. NIPSD maintains the connection between generation noise and teacher feedback while estimating the scores of the states actually produced by the student. The framework supports both one-step generation and few-step refinement, and can be incorporated into score-based distribution matching across different continuous token representations.

We evaluate NIPSD using DI and TDM as representative distillation objectives, and FLM and LangFlow as representative CDLMs. These combinations test its applicability across both objectives and representations. On OpenWebText with sequences of length 1024, LangFlow+NIPSD achieves the lowest generative perplexity among the compared methods at one, two, and four function evaluations. When initialized from an already-distilled two-step generator, its one-step model reaches a generative perplexity of 43.10, approaching the 512-step teacher's 41.52 with comparable sample entropy. These results demonstrate the potential of score distillation for compressing CDLM sampling into very few steps.

Our contributions are threefold:

- We characterize how categorical representations distinguish CDLM distillation from image distillation, identifying challenges in score estimation and the guidance available for learning distinct samples.
- We introduce NIPSD, a general framework that addresses these challenges within score-based distribution matching for one-step and few-step language generation.
- We validate the framework across representative distillation objectives and CDLMs, achieving strong generation quality with substantially fewer sampling steps.

2 PRELIMINARIES

Continuous diffusion language models CDLMs generate text by denoising continuous representations of a token sequence. For a length-L sequence represented by $X \in \mathbb{R}^{L\times d}$ using a codebook $E \in \mathbb{R}^{V\times d}$, a Gaussian marginal path and a categorical denoising prediction take the form

$$\begin{aligned} Z_t &= \alpha_t X + \sigma_t \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \\ m_{\phi,i}(z,t) &= E^\top p_{\phi,i}(z,t) = E^\top \operatorname{softmax}\big(\ell_{\phi,i}(z,t)\big). \end{aligned} \tag{1}$$

where $\ell_{\phi,i}$ denotes the token logits and $p_{\phi,i}$ the resulting token distribution at position i. Two representative models are *LangFlow* (Chen et al., 2026), which diffuses learned token embeddings along a variance-preserving path ($\alpha_t^2 + \sigma_t^2 = 1$), and the *Flow Language Model* (FLM) (Lee et al., 2026), which uses one-hot token states ($E = I_V$) and a linear path ($\alpha_t = t$, $\sigma_t = 1 - t$). Both refine all positions in parallel and decode the final predictions into tokens.

Score distillation Score distillation transfers a pretrained diffusion teacher into a generator requiring fewer function evaluations. A representative distribution-matching objective compares the teacher marginal P_t with the noisy student distribution $Q_{\theta,t}$:

$$\mathcal{L}_{\mathrm{DM}}(\theta) = \mathbb{E}_t[w_t D_{\mathrm{KL}}(Q_{\theta,t} \| P_t)]\,, \qquad s_R(z) = \nabla_z \log r(z), \tag{2}$$

where $w_t \geq 0$ and r is the density of a law R. Its generator gradient involves $s_{Q_{\theta,t}} - s_{P_t}$, typically estimated using an auxiliary student-score network and the frozen teacher. *Diff-Instruct* (Luo et al., 2023) and distribution matching distillation (DMD) (Yin et al., 2024b) use distribution-level teacher–student signals to train one-step generators. Score identity distillation (SiD) (Zhou et al., 2024; 2025b;c) provides a related data-free route.

Other one-step formulations include score implicit matching (SIM) (Luo et al., 2024), f-divergence distribution matching (Xu et al., 2025), and the unified divergence framework of *Uni-Instruct* (Wang et al., 2025). For few-step generation, DMD2 (Yin et al., 2024a) improves distribution-matching

training and supports multi-step sampling; trajectory distribution matching (TDM) (Luo et al., 2025c) aligns distributions along the generated trajectory. Subsequent approaches use adversarial alignment of score estimators (Lu et al., 2025), phase-wise matching within noise subintervals (Fan et al., 2026). Other works have extended one-step score distillation to flow-matching teachers using related techniques (Liu et al., 2026b;a; Huang et al., 2026; Zhou et al., 2025a; Luo, 2025; Luo et al., 2025a; 2026; 2025b; Wu et al., 2026).

CDLM distillation Flow Map Language Models (FMLMs) (Lee et al., 2026) distill FLM through flow-map objectives. Subsequent work extends flow-map distillation to self-conditioned models (Yoo et al., 2026). These results establish the feasibility of accelerating CDLMs. However, a systematic understanding of score distillation across continuous token representations and generated trajectory states remains limited. Appendix A reviews related work in more detail. The following section examines the differences from image diffusion that make this transfer nontrivial.

3 Understanding the Differences between Image and Language Distillation

Image and continuous language diffusion share a score-based distributional objective, but differ in how denoising predictions represent and recover generated samples. To illustrate the difference, let Y_θ be a differentiable noisy student state with law Q_θ, and let P be the teacher's target law at the same conditioning level. The distribution-matching gradient (Luo et al., 2023; 2025c) is

$$\nabla_\theta D_{\mathrm{KL}}(Q_\theta \| P) = \mathbb{E}\big[J_{Y_\theta}^\top \{s_{Q_\theta}(Y_\theta) - s_P(Y_\theta)\}\big]\,. \tag{3}$$

Here $J_{Y_\theta} = \partial Y_\theta / \partial \theta$, with sequence states flattened into vectors. The identity uses exact scores; the teacher approximates s_P. Score distillation in CDLMs raises two questions: whether the auxiliary model can represent the score of the student's generated states, and whether the teacher provides useful guidance for recovering differences between samples. We examine score estimation first, then the guidance available even with a correct student score.

3.1 Categorical constraints on student-score estimation

A natural choice for the auxiliary score estimator is to reuse the CDLM's categorical denoising head. For noised clean tokens $Z = aX + \sigma\epsilon$, with $a, \sigma > 0$ and standard Gaussian ϵ independent of X, the score of $P = \mathrm{Law}(Z)$ satisfies

$$s_P(z) = \frac{a\,\mathbb{E}[X \mid Z = z] - z}{\sigma^2}. \tag{4}$$

A categorical auxiliary head constrains its denoising prediction at each token position to $m_{\psi,i}(z) = E^\top \mathrm{softmax}(\ell_{\psi,i}(z)) \in \mathcal{C}$, where $\mathcal{C} = \mathrm{conv}\{e_1, \ldots, e_V\}$ is the token convex hull. For independently noised clean or soft token endpoints, the conditional mean remains in $\mathcal{C}^L$, so this output range introduces no irreducible representation error.

However, this compatibility does not necessarily extend to generated intermediate states, which can retain source noise and accumulated model predictions. For a noisy student state with law Q, exactly representing s_Q through the same conversion $s_\psi(y) = (a m_\psi(y) - y)/\sigma^2$ requires

$$m_\psi(y) = m_Q^{\mathrm{req}}(y) := \frac{y + \sigma^2 s_Q(y)}{a}. \tag{5}$$

This effective mean need not lie in $\mathcal{C}^L$. Whenever it falls outside the hull, a categorical head cannot represent the exact student score, regardless of its capacity or optimization. For FLM, the restriction is particularly concrete: its predicted mean must be nonnegative and sum to one at each position, whereas the effective mean required by a general continuous-state score need not satisfy either condition. Conventional image score models use real-valued outputs without a fixed token-codebook constraint (Karras et al., 2022; Peebles & Xie, 2023). Appendix D.2 derives the score-error bound, and Appendix D.3 constructs a generated state whose required mean lies outside the token convex hull.

Finding 1: Categorical heads can incur irreducible student-score error. For generated continuous states, the inherited score conversion may require an effective mean outside the token convex hull. A categorical head then cannot represent the exact student score.

This motivates an auxiliary score parameterization that accommodates the student's actual generated-state distribution. We implement it through a real-valued prediction of freshly added corruption noise, as described in Section 4.1.

3.2 Reduced output variation weakens teacher guidance

Even with a correct student score, distillation relies on useful teacher guidance at the student's noisy states. We focus on generators whose soft outputs become nearly identical across different input noises, as can occur when a high-noise categorical denoiser is reused directly as a generator. Our controlled comparison shows that reducing variation across samples attenuates the teacher's content-restoring response more sharply in the evaluated CDLMs than in image models. Appendix E distinguishes variation across inputs from token uncertainty.

To isolate the effect of reduced variation, we contract real samples toward their mean in each model's native representation. This controlled path models a loss of variation across generated outputs while keeping each sample's original content available as a reference. We then apply independent Gaussian corruption:

$$M_\lambda = (1-\lambda)\mu + \lambda X, \qquad Y_\lambda = aM_\lambda + \sigma\xi, \qquad 0 \le \lambda \le 1, \tag{6}$$

where $\mu = \mathbb{E}[X]$ and $\xi \sim \mathcal{N}(0, I)$ is independent of X. Here $\lambda = 1$ preserves the original samples, while $\lambda = 0$ maps every sample to the shared mean. For language, this interpolation remains a valid soft token mean. For each model, we select a conditioning level giving comparable normalized denoising error at $\lambda = 1$, then keep (a, σ) fixed as λ varies.

Let m_T denote the teacher's denoising prediction. Using the same corruption noise in Y_λ and Y_0, we measure its response along each sample's original content direction:

$$A_T(\lambda) = \mathbb{E}\big\langle X - \mu,\, m_T(Y_\lambda) - m_T(Y_0)\big\rangle, \qquad R_T(\lambda) = \frac{A_T(\lambda)}{A_T(1)}. \tag{7}$$

The expectation averages over samples and corruption noise. Subtracting the shared-mean prediction isolates the response to sample-dependent content, while $R_T(\lambda)$ expresses the fraction retained relative to the same teacher's full-content response. Appendix G provides the measurement protocol.

Figure 1a compares pixel-space EDM, VAE-latent DiT, LangFlow, and FLM. At $\lambda = 0.1$, EDM and DiT retain 9.41% and 6.75% of their full-content response, whereas LangFlow and FLM retain 1.81% and 0.145%. Raising λ from 0.1 to 0.6 increases language responses by factors of 16.3 and 201. The image responses track that amplitude more closely. Under the same relative contraction, the measured CDLMs therefore exhibit a sharper attenuation of teacher guidance along the directions that distinguish samples.

Finding 2: Content averaging attenuates teacher response more sharply in the evaluated CDLMs. Under independent Gaussian corruption, contracted language states retain a smaller fraction of their teacher's full-content response than the evaluated image states.

The next subsection establishes how this response enters the distillation gradient.

3.3 From teacher feedback to distillation updates

Section 3.2 measures how the teacher responds as differences between soft outputs shrink. Since distillation uses both teacher and student scores, we now connect this response to the KL gradient and examine whether the complete update remains aligned with restoring sample content.

Using the same contraction path, let $D = X - \mu$, $Q_\lambda = \mathrm{Law}(Y_\lambda)$, $P = \mathrm{Law}(Y_1)$, and $\mathcal{L}(\lambda) = D_{\mathrm{KL}}(Q_\lambda \| P)$. Let $m_{Q_\lambda}(y) = \mathbb{E}[M_\lambda \mid Y_\lambda = y]$ and $m_P(y) = \mathbb{E}[X \mid Y_1 = y]$ be the exact conditional means. The score-difference gradient gives

$$\mathcal{L}'(\lambda) = \frac{a^2}{\sigma^2}\mathbb{E}\langle m_{Q_\lambda}(Y_\lambda) - m_P(Y_\lambda), D\rangle. \tag{8}$$

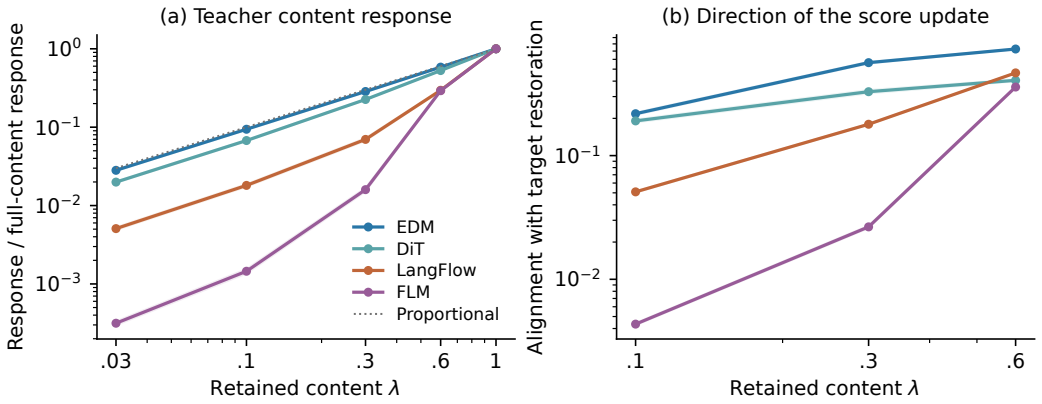


Figure 1: **Soft token averaging weakens the guidance available to score distillation.** (a) Teacher content response $R_T(\lambda)$ after contracting real samples toward their mean and applying fixed Gaussian corruption; the dotted line is proportional response. (b) Alignment of the complete score-difference state update with the direction that restores the target distribution. The student score uses a teacher-based reference obtained from the affine distribution path. Each model uses two independent groups of 64 real samples; shading spans the two group estimates. The measurements concern frozen-model feedback and state-space transport. Definitions, score-reference construction, and conditioning levels appear in Appendix G.

Define A_P as the response A_T in Eq. (7) with the learned denoiser m_T replaced by m_P. Since Y_0 is independent of D and $\mathbb{E}[D] = 0$, the teacher conditional-mean term contributes exactly $-a^2 A_P(\lambda)/\sigma^2$. This connects the unnormalized content response directly to the teacher's contribution toward increasing λ and restoring sample variation. Appendix G provides the derivation and a local curvature analysis.

We next assess the complete update, including the student score. For $0 < \lambda < 1$, the direction $X - M_\lambda = (1-\lambda)D$ restores each contracted sample to its original content. We define the gradient with respect to the contracted state M_λ, averaged over corruption noise, and measure the alignment of its descent direction with D:

$$\bar{g}_\lambda(X) = a\,\mathbb{E}_\xi[s_{Q_\lambda}(Y_\lambda) - s_P(Y_\lambda)]\,,$$
$$\mathcal{A}(\lambda) = -\frac{\mathbb{E}\langle D, \bar{g}_\lambda(X)\rangle}{\sqrt{\mathbb{E}\|D\|_2^2\,\mathbb{E}\|\bar{g}_\lambda(X)\|_2^2}}. \tag{9}$$

With exact scores, $\mathbb{E}\langle D, \bar{g}_\lambda(X)\rangle = \mathcal{L}'(\lambda)$, linking this metric to the same restoration path as Eq. (8). Higher $\mathcal{A}(\lambda)$ indicates better alignment with content restoration relative to the overall update magnitude.

Figure 1(b) evaluates this state-space alignment using the pretrained teacher and a teacher-based student-score reference along the affine path. At $\lambda = 0.1$, the alignment is 0.218 and 0.191 for EDM and DiT, compared with 0.051 and 0.0043 for LangFlow and FLM. Thus, under strong content contraction, the measured CDLM updates also align less with content restoration when the student-score term is included.

These results motivate establishing input-dependent content before distribution matching and preserving the connection between generation inputs and teacher queries during subsequent updates. Appendix C analyzes how paired supervision and source-dependent queries can provide first-order signals for such variation. Together with the score-estimation requirement in Section 3.1, these considerations guide the method in Section 4.

4 Methodology

We propose **Noise-Interpolated Perturbations for Score Distillation (NIPSD)**, a framework for learning the mapping from base noise to text through score-based distribution matching. Section 3

identifies two requirements: the auxiliary score must represent the student's actual continuous states, and the learning signal must distinguish how different inputs should generate different content. NIPSD estimates the score of generated states through continuous denoising score matching, establishes input-dependent predictions through teacher-trajectory initialization, and constructs score queries using *noise-interpolated perturbations (NIP)* between independent noise and the noise used for generation. Together, these designs make the generative mapping the object of both supervision and distribution matching.

Let $S_{\theta,j} = F_{\theta,j}(U)$ be a student state generated from base randomness U, where j selects an endpoint or an intermediate state. We form a score query

$$Y_{j,r} = c_{j,r} S_{\theta,j} + \sqrt{\beta_{j,r}}\,\xi, \qquad \xi \sim \mathcal{N}(0, I), \quad \xi \perp U, \quad \beta_{j,r} > 0, \tag{10}$$

and match its law $Q_{\theta,j,r}$ to the teacher marginal P_r using the objective in Eq. (2). Appendix B derives the distribution-matching gradient. The state selection and coefficients specify which part of the generative mapping receives supervision; Section 4.3 gives our query construction.

4.1 Estimating the score of generated states

Section 3.1 shows that a clean-token posterior need not represent the score required by a generated continuous state. Our ***FiTScore*** module therefore trains the auxiliary network on the actual queries in Eq. (10), using the freshly added noise as its regression target:

$$\mathcal{L}_{\text{fake}}(\psi) = \mathbb{E}\big[\lambda_{j,r} \|\epsilon_\psi(Y_{j,r}, j, r) - \xi\|_2^2\big]\,, \tag{11}$$

$$\widehat{s}_\psi(y, j, r) = -\frac{\epsilon_\psi(y, j, r)}{\sqrt{\beta_{j,r}}}, \qquad s_Q(y) = -\frac{\mathbb{E}[\xi \mid Y = y, j, r]}{\sqrt{\beta_{j,r}}}. \tag{12}$$

Here $\lambda_{j,r} > 0$ weights the regression and the student is frozen during the auxiliary update. The conditional expectation is the population minimizer of this loss, yielding the score of the query distribution. The target ξ is independent of the entire generated state before corruption, including any randomness retained from its generation.

The auxiliary output is real-valued in the continuous state coordinates. It can be implemented by a continuous projection or by adding a signed residual to a categorical score parameterization. This expands the auxiliary score class beyond that induced by a clean-token posterior while the generator retains its categorical predictions. The equivalent conditional-mean formulation and the categorical estimation bound are derived in Appendix D; concrete head parameterizations appear in Appendix F.

4.2 Establishing input-dependent generation

Sections 3.2 and 3.3 identify weak content-restoring guidance when different outputs approach a shared soft state. We address this starting condition with teacher-trajectory initialization. A teacher rollout provides a target tied to each student input, directly supervising the content that the input should generate before distribution matching begins.

Let h be a student input at conditioning level t, and let $T_{t\to s}(h)$ be the teacher's denoising prediction after advancing from h to a cleaner level s. Our consistency-distillation (CD) warmup minimizes

$$\mathcal{L}_{\text{CD}}(\theta) = \mathbb{E}_{h,t,s}[d(m_\theta(h, t), \text{sg}[T_{t\to s}(h)])]\,, \tag{13}$$

where d is a supervised discrepancy in the output representation and sg denotes stop-gradient. We use embedding regression or categorical cross-entropy according to the representation. For one-step generation, the target follows a teacher rollout from the same initial noise as the student. For a multistep generator, a teacher transition supplies the target for the corresponding segment.

This pairing supplies input-dependent learning directions even when the student initially makes similar predictions across inputs. For squared regression, the output-space descent direction at a shared prediction μ is $T_{t\to s}(h) - \mu$: different trajectory targets directly supervise different outputs. CD warmup thus establishes variation tied to teacher-generated content, providing an initial mapping for subsequent distribution matching. Appendix C formalizes this distinction between paired supervision and learning variation from independently corrupted outputs.

4.3 NOISE-INTERPOLATED PERTURBATIONS

During distribution matching, NIP interpolates between independent perturbation noise and the noise used to generate the sample. We realize this interpolation through further corruption of a generated trajectory state. Let $S_{\theta,j}$ correspond to path location t_j, and select a noisier teacher condition r. In Eq. (10), set

$$c_{j,r} = \frac{\alpha_r}{\alpha_{t_j}}, \qquad \beta_{j,r} = \sigma_r^2 - c_{j,r}^2 \sigma_{t_j}^2 > 0, \tag{14}$$

with $\alpha_{t_j} > 0$. This kernel respects the teacher's marginal path: $S_{\theta,j} \sim P_{t_j}$ implies $Y_{j,r} \sim P_r$. It therefore defines the target distribution for the generated state being matched.

The interpolation is explicit for a state constructed as $S_\theta = \alpha_s M_\theta(U) + \sigma_s U$, where $M_\theta(U)$ is a predicted endpoint and $U \sim \mathcal{N}(0, I)$ is its generation noise. Its query can be written as

$$Y_r = \alpha_r M_\theta(U) + \sigma_r \epsilon_\kappa, \qquad \epsilon_\kappa = \kappa U + \sqrt{1-\kappa^2}\,\xi, \qquad \kappa = \frac{\alpha_r \sigma_s}{\alpha_s \sigma_r}. \tag{15}$$

At $\kappa = 0$, the perturbation is independent Gaussian noise; at $\kappa = 1$, it is exactly the generation noise U. For each fixed $\kappa \in [0, 1]$, ϵ_κ remains standard Gaussian, while its dependence on the generated sample changes. Training uses $\kappa < 1$ to retain the positive fresh-noise variance required by Eq. (12). For a multistep generator, we apply Eq. (10) to the actual generated intermediate state, retaining its trajectory history.

Retaining this dependence makes variations across source inputs visible to distribution matching. For a local perturbation $M_\delta(U) = \mu + \delta f(U)$ with $\mathbb{E} f(U) = 0$, independent corruption of the endpoint has zero first-order density change at $\delta = 0$. A query that retains source dependence admits a first-order response through $\mathbb{E}[f(U) \mid Y_0]$ (Proposition 1, Appendix C). This supplies a channel for learning input-dependent changes when endpoint differences are weak. CD warmup establishes these differences through paired targets; NIP retains a route for their supervision during distribution matching.

4.4 TRAINING AND APPLICABILITY

Algorithm 1 summarizes NIPSD training. After CD warmup, we briefly freeze the student and fit the auxiliary score to its generated states. We then alternate auxiliary updates using Eq. (11) with generator updates. For each query, the teacher–student score difference supplies the state direction

$$g_{j,r} = c_{j,r}\{\widehat{s}_\psi(Y_{j,r}, j, r) - s_T(Y_{j,r}, r)\}, \tag{16}$$

which is backpropagated through $S_{\theta,j}$ with the weights of the distribution-matching objective. The score predictions are stopped during the generator update. We also use a teacher-prediction anchor on forward-noised training sequences to retain denoising knowledge. Appendix F gives the anchor, the normalized regression surrogate, and the treatment of multistep trajectories. At inference, only the distilled generator is required.

NIPSD is formulated in terms of continuous generated states, Gaussian marginal paths, and teacher–student score differences. These interfaces allow the same design to be incorporated into score-based distribution matching across CDLM representations and generator step counts. We evaluate this generality using DI and TDM as representative distillation objectives, and LangFlow and FLM as representative CDLMs. Their combinations test the framework across distinct objectives and representations.

5 EXPERIMENTS

Datasets Our experiments cover the One Billion Word benchmark (LM1B) (Chelba et al., 2013) and OpenWebText (OWT) (Gokaslan & Cohen, 2019). For LM1B, we use the `bert-base-uncased` tokenizer and pack the tokenized text into sequences of length $L = 128$; for OWT, we use the GPT-2 tokenizer and $L = 1024$. The corresponding vocabulary sizes are $|V| = 30{,}522$ and $|V| = 50{,}257$.

Training We evaluate NIPSD by distilling two representative CDLM teachers: FLM (Lee et al., 2026), which operates on one-hot token representations, and LangFlow (Chen et al., 2026), which

Table 1: **Few-step generation on OpenWebText ($L = 1024$).** NIPSD uses DI at 1 NFE and TDM at 2 and 4 NFE for both CDLMs, with the same teacher checkpoint within each family. Gen. PPL is measured by GPT-2 Large (lower is better); Ent. is mean per-sample unigram entropy. Shading denotes NIPSD; bold marks the lowest Gen. PPL at each NFE.

Exp set	NFE=1		NFE=2		NFE=4	
	Gen. PPL (↓)	Ent.	Gen. PPL (↓)	Ent.	Gen. PPL (↓)	Ent.
MDLM+SDTT	1260.86	5.26	877.22	5.34	339.73	5.38
MDLM + Di4C	1298.80	5.29	758.23	5.35	239.27	5.40
Duo + DCD	5743.29	6.02	891.16	5.41	250.86	5.37
Duo + Di4C	370.51	3.92	210.22	4.63	154.67	4.85
FMLM	168.30	5.17	133.29	5.25	111.31	5.26
FLM+NIPSD	132.90	5.15	78.77	5.30	66.37	5.24
LangFlow+NIPSD	**70.08**	5.24	**75.50**	5.26	**48.06**	5.25

Table 2: **One-step generation initialized from distilled two-step models on OpenWebText.** For each CDLM, we initialize a one-step NIPSD model from its distilled two-step generator, retaining the DI objective and other one-step training settings. The shaded columns report the resulting one-step model; the remaining columns report its corresponding undistilled teacher at 256, 512, and 1,024 NFE. G. PPL is measured by GPT-2 Large (lower is better), and Ent. denotes mean per-sample unigram entropy.

Exp set	NFE=1024		NFE=512		NFE=256		Ours NFE=1	
	G. PPL	Ent.	G. PPL	Ent.	G. PPL	Ent.	G. PPL	Ent.
FLM	62.23	5.33	70.91	5.41	83.19	5.49	73.55	5.17
LangFlow	36.53	5.25	41.52	5.31	49.24	5.37	43.10	5.26

operates on learned token embeddings. We use 179M-parameter models with 12-block diffusion Transformer (DiT) backbones (Peebles & Xie, 2023). Following prior diffusion language models (Sahoo et al., 2024; 2025), the backbones use rotary positional embeddings (RoPE) (Su et al., 2021) and adaptive layer normalization (AdaLN) for time conditioning.

Evaluation We assess generation quality using generative perplexity (Gen. PPL) and sample entropy. Following prior work (Sahoo et al., 2024; Schiff et al., 2025), we generate 1,024 samples per setting for the main results and compute Gen. PPL with GPT-2 Large (Radford et al., 2019), exponentiating the average token-level negative log-likelihood across the generated samples. We also report the mean per-sample unigram entropy to measure within-sequence token diversity and help identify repetitive generations that can yield misleadingly low Gen. PPL.

Baselines We compare NIPSD instantiated with FLM and LangFlow against representative distillation methods for both discrete and continuous diffusion language models. The baselines are MDLM+SDTT (Deschenaux & Gulcehre, 2025), MDLM+Di4C (Hayakawa et al., 2025), Duo+DCD (Sahoo et al., 2025), Duo+Di4C (Hayakawa et al., 2025), and FMLM (Lee et al., 2026).

Table 1 shows that NIPSD improves few-step generation across both continuous token representations. With FLM, NIPSD reduces Gen. PPL from FMLM's 168.30, 133.29, and 111.31 to 132.90, 78.77, and 66.37 at 1, 2, and 4 NFE, respectively—relative reductions of 21.0%, 40.9%, and 40.4%. LangFlow+NIPSD achieves the lowest Gen. PPL in every NFE column, reaching 70.08 at one step and 48.06 at four steps. Its one-step result also improves on four-step FMLM by 37.0% while using one quarter as many function evaluations. Performance is not strictly monotonic in NFE: LangFlow's two-step Gen. PPL is 75.50, above its one-step result.

The perplexity gains are accompanied by comparable within-sequence token diversity: NIPSD's entropy ranges from 5.15 to 5.30, close to FMLM's 5.17–5.26. In contrast, the discrete distillation baselines exhibit either substantially higher perplexity or a larger entropy deficit. For example, at one NFE, MDLM+SDTT and MDLM+Di4C exceed 1,200 Gen. PPL despite entropies above 5.2,

Table 3: **Component ablation of DI-based NIPSD on LangFlow (LM1B, $L = 128$, NFE=1).** Red marks severe entropy collapse; the low perplexities reflect repetitive outputs.

Setting	NFE=1	
	Gen. PPL ↓	Ent.
LangFlow	2.02	0.30
+ DI	6.05	0.09
+ CD warmup	150.27	4.32
+ NIP	89.20	4.33
+ FiTScore (NIPSD)	**75.99**	4.28

LangFlow teacher: one-step samples

Official LM1B checkpoint · NFE=1

Sample 1 (excerpt)

```
the the the the the the the the
the the the the the the the the
the the the the the the the the
the the the.  the.  the...
```

Sample 2 (excerpt)

```
the the the the the the the the
the the the the the the the the
the the the the the the the the.
the the the.  the the...
```

while Duo+Di4C reaches 370.51 with entropy 3.92. Together, the two metrics support improved generation quality without a comparable loss of token diversity across the evaluated NIPSD models.

Table 2 examines one-step models initialized from already-distilled two-step generators. This setting yields Gen. PPL of 73.55 for FLM and 43.10 for LangFlow, respectively 44.7% and 38.5% below the corresponding one-step results in Table 1. LangFlow+NIPSD surpasses its 256-NFE teacher (49.24) and approaches its 512-NFE teacher (41.52) with a single function evaluation. Its entropy of 5.26 is also close to the teacher's 5.25–5.37 range. For FLM, one-step NIPSD approaches the 512-NFE teacher's perplexity (73.55 versus 70.91), although its entropy is lower (5.17 versus 5.41). These results show that a distilled few-step generator can provide an effective starting point for further compression to one-step generation.

Component ablation. Table 3 cumulatively adds NIPSD components to DI with LangFlow on LM1B ($L = 128$). The teacher's one-step outputs are dominated by repeated tokens, and direct DI distillation remains collapsed, with sample entropy of 0.09. Thus, directly copying the teacher provides a poor initialization for one-step generation in this setting. CD warmup restores token diversity (entropy 4.32), establishing a viable starting point for distribution matching. Adding NIP lowers Gen. PPL from 150.27 to 89.20 at comparable entropy. FiTScore, which combines continuous score outputs with fresh-noise regression (Section 4.1), further reduces it to 75.99, completing NIPSD. These results support establishing non-degenerate generation before learning from distributional feedback. Appendix I provides additional controlled diagnostics; Appendix K presents generated text.

6 Discussion

Our results highlight that effective CDLM distillation depends on how continuous learning signals interact with categorical representations. A teacher's generation quality alone does not characterize its guidance on student outputs. NIPSD addresses this distinction by estimating the student's actual state distribution and retaining feedback tied to different generation inputs. More broadly, representation and query design should be considered alongside the distillation objective when developing fast language generators.

Distilled few-step models also provide effective starting points for further compression, as demonstrated by the two-step-to-one-step results in Table 2. This motivates studying successive reductions in sampling steps. Our current evaluation covers two CDLM representations and two distillation objectives; larger models and conditional generation remain important extensions. Assessing semantic diversity, long-range coherence, and inference latency would further characterize the benefits of few-step generation beyond generative perplexity and per-sample token entropy.

AI USE STATEMENT

We used generative AI tools to assist in interpreting results and in polishing and revising the manuscript. AI tools were also used to support literature searches, reference formatting, and debugging research code. The authors retain full responsibility for all research decisions and for verifying the correctness, originality, and reproducibility of the work, including the manuscript text, theoretical claims, code, and empirical analyses.

ETHICS STATEMENT

This work studies efficient language generation using existing text benchmarks and pretrained diffusion teachers. Distilled models can inherit biases, harmful content, and memorized information from their teachers and training data. Lower inference costs can broaden access to language generation while also lowering the cost of misuse, including spam and misleading content. Our generation-quality metrics do not establish safety, fairness, or privacy guarantees; deployment requires assessment and safeguards appropriate to the application. Fewer sampling steps also do not by themselves establish a net environmental benefit, since distillation incurs additional training costs.

REFERENCES

Nicholas M. Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *Transactions on Machine Learning Research*, 2025. URL https://arxiv.org/abs/2406.07507.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013. URL https://arxiv.org/abs/1312.3005.

Yuxin Chen, Chumeng Liang, Hangke Sui, Ruihan Guo, Chaoran Cheng, Jiaxuan You, and Ge Liu. LangFlow: Continuous diffusion rivals discrete in language modeling. *arXiv preprint arXiv:2604.11748*, 2026. URL https://arxiv.org/abs/2604.11748.

Justin Deschenaux and Caglar Gulcehre. Beyond autoregression: Fast LLMs via self-distillation through time. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.21035.

Xiangyu Fan, Zesong Qiu, Zhuguanyu Wu, Fanzhou Wang, Zhiqian Lin, Tianxiang Ren, Dahua Lin, Ruihao Gong, and Lei Yang. Phased dmd: Few-step distribution matching distillation via score matching within subintervals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41667–41676, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Fan_Phased_DMD_Few-step_Distribution_Matching_Distillation_via_Score_Matching_within_CVPR_2026_paper.html.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025. URL https://arxiv.org/abs/2505.13447.

Aaron Gokaslan and Vanya Cohen. OpenWebText corpus. https://skylion007.github.io/OpenWebTextCorpus/, 2019.

Ishaan Gulrajani and Tatsunori B. Hashimoto. Likelihood-based diffusion language models. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.18619.

Satoshi Hayakawa, Yuhta Takida, Masaaki Imaizumi, Hiromi Wakaki, and Yuki Mitsufuji. Distillation of discrete diffusion through dimensional correlations. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 22259–22297. PMLR, 2025. URL https://proceedings.mlr.press/v267/hayakawa25a.html.

Keya Hu, Linlu Qiu, Yiyang Lu, Hanhong Zhao, Tianhong Li, Yoon Kim, Jacob Andreas, and Kaiming He. ELF: Embedded language flows. *arXiv preprint arXiv:2605.10938*, 2026. URL https://arxiv.org/abs/2605.10938.

Zemin Huang, Weijian Luo, Zhengyang Geng, and Guojun Qi. One-step diffusion and flow distillation through implicit generator matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(8):9156–9167, 2026. doi: 10.1109/TPAMI.2026.3676894.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2206.00364.

Chanhyuk Lee, Jaehoon Yoo, Manan Agarwal, Sheel Shah, Jerry Huang, Aditi Raghunathan, Seunghoon Hong, Nicholas M. Boffi, and Jinwoo Kim. Flow map language models: One-step language modeling via continuous denoising. *arXiv preprint arXiv:2602.16813*, 2026. URL https://arxiv.org/abs/2602.16813.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. Diffusion-LM improves controllable text generation. *arXiv preprint arXiv:2205.14217*, 2022. URL https://arxiv.org/abs/2205.14217.

Dongyang Liu, Peng Gao, David Liu, Ruoyi Du, Zhen Li, Qilong Wu, Xin Jin, Sihan Cao, Shifeng Zhang, Steven Hoi, and Hongsheng Li. Decoupled dmd: Cfg augmentation as the spear, distribution matching as the shield. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e425b75bac5742a008d643826428787c-Abstract-Conference.html.

Tao Liu, Hao Yan, Mengting Chen, Taihang Hu, Zhengrong Yue, Zihao Pan, Jinsong Lan, Xiaoyong Zhu, Ming-Ming Cheng, Bo Zheng, and Yaxing Wang. Continuous-time distribution matching for few-step diffusion distillation, 2026b. URL https://arxiv.org/abs/2605.06376.

Yanzuo Lu, Yuxi Ren, Xin Xia, Shanchuan Lin, Xing Wang, Xuefeng Xiao, Andy J. Ma, Xiaohua Xie, and Jian-Huang Lai. Adversarial distribution matching for diffusion distillation towards efficient image and video synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16818–16829, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Lu_Adversarial_Distribution_Matching_for_Diffusion_Distillation_Towards_Efficient_Image_and_Video_Synthesis_ICCV_2025_paper.html.

Weijian Luo. Diff-instruct++: Training one-step text-to-image generator model to align with human preferences. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=SeGNvJJjbs.

Weijian Luo, Tianyang Hu, Shifeng Zhang, Jiacheng Sun, Zhenguo Li, and Zhihua Zhang. Diff-Instruct: A universal approach for transferring knowledge from pre-trained diffusion models. *arXiv preprint arXiv:2305.18455*, 2023. URL https://arxiv.org/abs/2305.18455.

Weijian Luo, Zemin Huang, Zhengyang Geng, J. Zico Kolter, and Guo-Jun Qi. One-step diffusion distillation through score implicit matching. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2410.16794.

Weijian Luo, Colin Zhang, Debing Zhang, and Zhengyang Geng. David and goliath: Small one-step model beats large diffusion with score post-training. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 41520–41539, 2025a. URL https://proceedings.mlr.press/v267/luo25s.html.

Yihong Luo, Tianyang Hu, Weijian Luo, Kenji Kawaguchi, and Jing Tang. Reward-instruct: A reward-centric approach to fast photo-realistic image generation. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/f39cbd56877419229aef90dbe463c182-Abstract-Conference.html.

Yihong Luo, Tianyang Hu, Jiacheng Sun, Yujun Cai, and Jing Tang. Learning few-step diffusion models by trajectory distribution matching. *arXiv preprint arXiv:2503.06674*, 2025c. URL https://arxiv.org/abs/2503.06674.

Yihong Luo, Tianyang Hu, Weijian Luo, and Jing Tang. TDM-R1: Reinforcing few-step diffusion models with non-differentiable reward, 2026. URL https://arxiv.org/abs/2603.07700.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2212.09748.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.07524.

Subham Sekhar Sahoo, Justin Deschenaux, Aaron Gokaslan, Guanghan Wang, Justin T. Chiu, and Volodymyr Kuleshov. The diffusion duality. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 52584–52619. PMLR, 2025. URL https://proceedings.mlr.press/v267/sahoo25a.html.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2202.00512.

Yair Schiff, Subham Sahoo, Hao Phung, Guanghan Wang, Sam Boshar, Hugo Dalla-Torre, Bernardo Almeida, Alexander Rush, Thomas Pierrot, and Volodymyr Kuleshov. Simple guidance mechanisms for discrete diffusion models. In *International Conference on Learning Representations*, 2025.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023. URL https://arxiv.org/abs/2303.01469.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021. URL https://arxiv.org/abs/2104.09864.

Yifei Wang, Weimin Bai, Colin Zhang, Debing Zhang, Weijian Luo, and He Sun. Uni-instruct: One-step diffusion model through unified diffusion divergence instruction. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/5297e56ac65ba2bfa70ee9fc4818c042-Abstract-Conference.html.

Junyi Wu, Weijian Luo, Haoyang Zheng, Ruizhe Zhang, and Guang Lin. Diff-instruct with diffused reward: Towards principled one-step generator rl, 2026. URL https://arxiv.org/abs/2605.24001.

Yilun Xu, Weili Nie, and Arash Vahdat. One-step diffusion models with f-divergence distribution matching, 2025. URL https://arxiv.org/abs/2502.15681.

Zhihan Yang, Wei Guo, Shuibai Zhang, Subham Sekhar Sahoo, Yongxin Chen, Arash Vahdat, Morteza Mardani, and John Thickstun. Continuous diffusion scales competitively with discrete diffusion for language. *arXiv preprint arXiv:2605.18530*, 2026. URL https://arxiv.org/abs/2605.18530.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Frédo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. *arXiv preprint arXiv:2405.14867*, 2024a. URL https://arxiv.org/abs/2405.14867.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b. URL https://arxiv.org/abs/2311.18828.

Jaehoon Yoo, Wonjung Kim, Floor Eijkelboom, Chanhyuk Lee, Nicholas M. Boffi, Seunghoon Hong, and Jinwoo Kim. Self-conditioned flow map language models via fixed-point flows. *arXiv preprint arXiv:2607.00714*, 2026. URL https://arxiv.org/abs/2607.00714.

Mingyuan Zhou, Huangjie Zheng, Zhendong Wang, Mingzhang Yin, and Hai Huang. Score identity distillation: Exponentially fast distillation of pretrained diffusion models for one-step generation. *arXiv preprint arXiv:2404.04057*, 2024. URL https://arxiv.org/abs/2404.04057.

Mingyuan Zhou, Yi Gu, Huangjie Zheng, Liangchen Song, Guande He, Yizhe Zhang, Wenze Hu, and Yinfei Yang. Score distillation of flow matching models, 2025a. URL https://arxiv.org/abs/2509.25127.

Mingyuan Zhou, Zhendong Wang, Huangjie Zheng, and Hai Huang. Guided score identity distillation for data-free one-step text-to-image generation. In *International Conference on Learning Representations*, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/6246f93ebf4f2d76ad2bcb3158220d34-Abstract-Conference.html.

Mingyuan Zhou, Huangjie Zheng, Yi Gu, Zhendong Wang, and Hai Huang. Adversarial score identity distillation: Rapidly surpassing the teacher in one step. In *International Conference on Learning Representations*, 2025c. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ce953d71deeb33d9ffa2c879b518d273-Abstract-Conference.html.

Algorithm 1 NIPSD training

Require: Frozen teacher T; teacher-initialized student θ; auxiliary noise predictor ϵ_ψ;
training data $\mathcal{D}$; Gaussian path (α_t, σ_t); phase lengths $N_{\mathrm{CD}}, N_{\mathrm{aux}}, N_{\mathrm{DM}}$;
auxiliary updates per generator update n_ψ.
Ensure: Distilled generator F_θ.

Teacher-trajectory initialization
for $k = 1, \dots, N_{\mathrm{CD}}$ **do**
  Sample a student input h at t and a cleaner target level s.
  $\bar{m} \leftarrow \mathrm{sg}[T_{t\to s}(h)]$ ▷ Teacher target from the same input
  Update θ using $d(m_\theta(h,t), \bar{m})$ as in Eq. (13).
end for
Auxiliary-score warmup
for $k = 1, \dots, N_{\mathrm{aux}}$ **do**
  $\psi \leftarrow \textsc{FitScore}(\theta, \psi)$ ▷ Student fixed
end for
Alternating distribution matching
for $k = 1, \dots, N_{\mathrm{DM}}$ **do**
  for $a = 1, \dots, n_\psi$ **do**
    $\psi \leftarrow \textsc{FitScore}(\theta, \psi)$
  end for
  $(S, Y, \xi, j, r, c, \beta) \leftarrow \textsc{NIPQuery}(\theta)$
  $\widehat{s}_Q \leftarrow -\epsilon_\psi(Y, j, r)/\sqrt{\beta}$; $\widehat{s}_P \leftarrow s_T(Y, r)$
  $g \leftarrow \mathrm{sg}[c(\widehat{s}_Q - \widehat{s}_P)]$
  Form $S^\star = \mathrm{sg}(S - \eta\bar{g})$ with the normalization in Eq. (54).
  Sample forward-noised data from $\mathcal{D}$; compute $\mathcal{L}_{\mathrm{anchor}}$ via Eq. (53).
  Update θ using Eq. (55) with $S, S^\star, \mathcal{L}_{\mathrm{anchor}}$; keep ψ fixed.
end for
return F_θ

function $\textsc{FitScore}(\theta, \psi)$
  $(S, Y, \xi, j, r, c, \beta) \leftarrow \textsc{NIPQuery}(\theta)$
  $\ell_\psi \leftarrow \lambda_{j,r}\|\epsilon_\psi(\mathrm{sg}[Y], j, r) - \xi\|_2^2$
  Update ψ using ℓ_ψ; **return** ψ. ▷ Regress only the fresh noise ξ
end function

function $\textsc{NIPQuery}(\theta)$
  Sample generation noise $U \sim \mathcal{N}(0, I)$ and query indices (j, r).
  $c \leftarrow \alpha_r/\alpha_{t_j}$; $\beta \leftarrow \sigma_r^2 - c^2\sigma_{t_j}^2$, with $\beta > 0$.
  Generate $S \leftarrow F_{\theta,j}(U)$; detach any multistep prefix.
  ▷ For an endpoint predictor, construct $S = \alpha_{t_j} M_\theta(U) + \sigma_{t_j} U$.
  Sample fresh noise $\xi \sim \mathcal{N}(0, I)$ independently of U.
  $Y \leftarrow cS + \sqrt{\beta}\,\xi$ ▷ Retain the generation noise already present in S
  return $(S, Y, \xi, j, r, c, \beta)$
end function

A Related Work

A.1 Continuous diffusion language models

Continuous diffusion language models generate text by denoising real-valued representations and decoding them into discrete tokens. Early work established this approach through embedding-space generation in Diffusion-LM (Li et al., 2022) and likelihood-based training in Plaid (Gulrajani & Hashimoto, 2023). Recent models explore different representations and training objectives. Flow Language Models (FLMs) apply continuous flow matching to one-hot token representations with categorical denoising objectives (Lee et al., 2026). LangFlow operates in learned token embeddings and connects flow matching with cross-entropy training through a Bregman-divergence formulation (Chen et al., 2026). RePlaid modernizes Plaid's architecture and demonstrates competitive scaling with discrete diffusion language models (Yang et al., 2026). ELF instead models frozen contextual embeddings, keeping denoising continuous and performing categorical decoding only at the end (Hu et al., 2026).

These developments establish continuous representations as a competitive basis for language generation, while exposing different interfaces between continuous dynamics and discrete tokens. We study the implications of this interface for distillation, focusing on models such as LangFlow and FLM whose categorical predictions are converted into continuous denoising outputs. In this setting, both the representable auxiliary score and the information retained by soft token states affect how distribution matching trains a fast generator.

A.2 Diffusion distillation

One route to faster diffusion generation learns mappings along denoising trajectories. Progressive distillation repeatedly compresses teacher sampling steps (Salimans & Ho, 2022), while consistency models learn to map points on a trajectory to a common endpoint (Song et al., 2023). Flow map matching generalizes this perspective to mappings between arbitrary time pairs (Boffi et al., 2025). Flow Map Language Models (FMLMs) adapt flow-map distillation to FLMs using objectives that respect the categorical output geometry (Lee et al., 2026); subsequent work extends this approach to self-conditioned models through fixed-point flows (Yoo et al., 2026). MeanFlow offers a related route to one-step generation by learning average velocities over time intervals, with its original formulation supporting training from scratch without a pretrained teacher (Geng et al., 2025). ELF also explores progressive distillation for few-step language generation (Hu et al., 2026).

Score-based distillation instead supplies distributional supervision through teacher and student scores. Diff-Instruct (DI) minimizes an integral KL divergence between noisy teacher and generator distributions (Luo et al., 2023). Distribution Matching Distillation (DMD) combines a score-based distribution-matching update with paired regression (Yin et al., 2024b); DMD2 removes this regression requirement and stabilizes training through two-time-scale updates and adversarial supervision (Yin et al., 2024a). Trajectory Distribution Matching (TDM) extends distribution matching to intermediate sampling states for few-step generation (Luo et al., 2025c). Complementary formulations include Score identity Distillation (SiD), which uses score identities to optimize a Fisher-divergence objective (Zhou et al., 2024), and Score Implicit Matching (SIM), which derives tractable gradient estimators for a broader family of score-based divergences (Luo et al., 2024).

Our work investigates how score-based distribution matching transfers from image diffusion to CDLMs. We characterize how categorical score parameterizations and soft token states alter score estimation and sample-specific learning signals relative to image diffusion. NIPSD addresses these challenges within a common framework for score-based distribution matching across CDLM representations and generator step counts.

B Distributional identities and derivations

We work in flattened sequence coordinates $\mathbb{R}^{Ld}$. For the differentiations below, assume that the state map is differentiable in its parameters, all relevant expectations and KL divergences are finite, differentiation can be exchanged with integration, and boundary terms vanish under integration by parts. A sufficient setting for the elementary examples is a smooth bounded state map with integrable parameter derivatives, positive Gaussian corruption variance, and a finite Gaussian-mixture target. The distributional identities use the exact reference score s_{P_t}, which the pretrained teacher approximates in training.

B.1 Score-difference gradient

For the generated states and queries in Eq. (10), the distribution-matching objective and its exact gradient are

$$\mathcal{L}_{\mathrm{DM}}(\theta) = \mathbb{E}_{j,r}[w_{j,r} D_{\mathrm{KL}}(Q_{\theta,j,r} \| P_r)]\,, \tag{17}$$

$$\nabla_\theta \mathcal{L}_{\mathrm{DM}} = \mathbb{E}\big[w_{j,r} c_{j,r} J_{F_{\theta,j}}(U)^\top \{s_{Q_{\theta,j,r}}(Y_{j,r}) - s_{P_r}(Y_{j,r})\}\big]\,, \tag{18}$$

where the weights $w_{j,r} \geq 0$ and the sampling distribution of (j, r) are independent of θ.

For a scalar parameter coordinate θ_k, let $v_k(y) = \mathbb{E}[\partial_{\theta_k} Y \mid Y = y]$. The density of a differentiable random transport satisfies the continuity equation $\partial_{\theta_k} q = -\nabla \cdot (q v_k)$. For a fixed target density p,

$$\partial_{\theta_k} D_{\mathrm{KL}}(q\|p) = \int (1 + \log(q/p)) \partial_{\theta_k} q \, dy \tag{19}$$

$$= \int q(y) v_k(y)^\top \{\nabla \log q(y) - \nabla \log p(y)\} \, dy \tag{20}$$

$$= \mathbb{E}\big[(\partial_{\theta_k} Y)^\top \{s_Q(Y) - s_P(Y)\}\big] . \tag{21}$$

For Eq. (10), $\partial_{\theta_k} Y = c_{j,r} \partial_{\theta_k} F_{\theta,j}(U)$. Averaging over the parameter-independent kernel indices and weights gives Eq. (18). If the prefix of a shared-parameter rollout is detached, this calculation applies to the current map with its input distribution held fixed. It does not include the derivative of the prefix distribution.

B.2 Gaussian smoothing and joint-distribution identifiability

For $Y = \alpha X + \sigma \xi$, where $\alpha, \sigma > 0$ and ξ is independent standard Gaussian, the characteristic function is

$$\varphi_Y(\omega) = \varphi_X(\alpha\omega) \exp(-\sigma^2 \|\omega\|_2^2 / 2). \tag{22}$$

The Gaussian factor is nonzero at every finite ω. Equality of two smoothed laws therefore implies equality of the characteristic functions, and hence laws, of their clean variables. With distinct token codewords, the embedding of an entire token sequence is injective, so the statement identifies the full joint token law. The result concerns equality of distributions under exact smoothing.

B.3 Consistency of the trajectory kernel

Suppose S has law P_s, represented as $S \overset{d}{=} \alpha_s X + \sigma_s \epsilon$. Draw a fresh independent ξ and set $c = \alpha_r / \alpha_s$, $\beta = \sigma_r^2 - c^2 \sigma_s^2 > 0$. Then

$$cS + \sqrt{\beta}\xi \overset{d}{=} \alpha_r X + c\sigma_s \epsilon + \sqrt{\beta}\xi \overset{d}{=} \alpha_r X + \sigma_r \epsilon', \tag{23}$$

where ϵ' is standard Gaussian independent of X. This proves the stated consistency for Gaussian marginal paths. It constructs a compatible Markov kernel between marginals, without identifying it with the samplewise coupling of a deterministic interpolant.

C Teacher-Trajectory Initialization and Noise Interpolation

C.1 Paired teacher supervision

Under the reference forward process, the law of total covariance gives

$$\mathrm{Cov}(\mathbb{E}[X \mid Z_t]) = \mathrm{Cov}(X) - \mathbb{E}[\mathrm{Cov}(X \mid Z_t)]. \tag{24}$$

At zero signal, the optimal denoiser is the input-independent prior mean. Using that mean as a generator endpoint does not reproduce the variance contained within the posterior.

To examine how paired supervision learns variation, write $M_\delta(U) = \mu + \delta f(U)$ with $\mathbb{E} f(U) = 0$, and let $T(U)$ be a fixed teacher-trajectory target paired with U. For squared regression,

$$\frac{d}{d\delta} \frac{1}{2} \mathbb{E}\|M_\delta(U) - T(U)\|^2 \Big|_{\delta=0} = -\mathbb{E}\langle f(U), T(U) - \mu \rangle. \tag{25}$$

This derivative is negative for variations positively correlated with the paired target. In particular, if the perturbation $f(U) = T(U) - \mathbb{E}T(U)$ is available, it equals $-\operatorname{tr} \mathrm{Cov}(T(U))$. Thus nonconstant paired targets provide a first-order direction for learning variation, even at a constant prediction. The next result contrasts this with independent endpoint corruption and characterizes the effect of retaining the source in a score query.

C.2 First-order visibility of latent variation

Proposition 1 (Local visibility of latent variation). *Let $G_\delta(U) = \mu + \delta f(U)$, with $\mathbb{E}f(U) = 0$, and assume differentiation under the density and KL integrals is valid. For $a, \sigma > 0$, the density of $Y_\delta = aG_\delta(U) + \sigma\xi$, with independent Gaussian ξ, satisfies $\partial_\delta q_\delta|_{\delta=0} = 0$. Its KL derivative against a fixed positive target density is therefore zero. If instead $U = z \sim \mathcal{N}(0, I)$ and*

$$Y_\delta = aG_\delta(z) + bz + c\xi, \qquad b^2 + c^2 = \sigma^2, \quad c > 0, \tag{26}$$

then

$$\partial_\delta q_\delta(y)|_0 = -a\nabla_y \cdot [q_0(y)\mathbb{E}[f(z) \mid Y_0 = y]], \tag{27}$$

which need not vanish.

Proof. Let ϕ_σ denote the centered Gaussian density with covariance $\sigma^2 I$. With independent corruption,

$$q_\delta(y) = \mathbb{E}_U \phi_\sigma(y - a\mu - a\delta f(U)), \qquad \partial_\delta q_\delta(y)|_0 = -a\nabla\phi_\sigma(y - a\mu)^\top \mathbb{E}f(U) = 0. \tag{28}$$

The KL derivative is the integral of this derivative against $1 + \log(q_0/p)$ and is therefore zero. For the noise-interpolated query,

$$q_\delta(y) = \mathbb{E}_z \phi_c(y - a\mu - bz - a\delta f(z)). \tag{29}$$

Differentiation gives $-a\nabla_y \cdot \mathbb{E}_z[f(z)\phi_c(y - a\mu - bz)]$. The quantity inside the divergence equals $q_0(y)\mathbb{E}[f(z) \mid Y_0 = y]$, proving Eq. (27). The covariance makes the distinction explicit:

$$\mathrm{Cov}(Y_\delta) = \sigma^2 I + ab\delta\{\mathrm{Cov}(f(z), z) + \mathrm{Cov}(z, f(z))\} + a^2\delta^2 \,\mathrm{Cov}(f(z)). \tag{30}$$

The linear term is absent for independent corruption. A nonzero density derivative need not imply descent against every target; the next example supplies a target and an improving direction.

C.3 A binary categorical example with an exact gradient

Take codewords $\{-1, +1\}$, logits $(-\theta z, \theta z)$, and the corresponding expected codeword $G_\theta(z) = \tanh(\theta z)$, with $z \sim \mathcal{N}(0, 1)$. Let

$$P = \tfrac{1}{2}\mathcal{N}(-a, \sigma^2) + \tfrac{1}{2}\mathcal{N}(a, \sigma^2), \qquad Y_\theta = a\tanh(\theta z) + bz + c\xi, \qquad b^2 + c^2 = \sigma^2. \tag{31}$$

At $\theta = 0$, $Q_0 = \mathcal{N}(0, \sigma^2)$ for every admissible b, while $\partial_\theta G_\theta|_0 = z$. The two scores satisfy

$$s_{Q_0}(y) - s_P(y) = -\frac{a}{\sigma^2}\tanh\Big(\frac{ay}{\sigma^2}\Big). \tag{32}$$

Using $\mathbb{E}[z \mid Y_0 = y] = by/\sigma^2$ in the score-difference gradient yields

$$\partial_\theta D_{\mathrm{KL}}(Q_\theta \| P)|_0 = -\frac{a^2 b}{\sigma^4}\mathbb{E}_{Y_0}\left[Y_0 \tanh\left(\frac{aY_0}{\sigma^2}\right)\right]. \tag{33}$$

This is zero for $b = 0$ and strictly negative for $b > 0$. The different gradients occur at the same initial student distribution and target, through the different dependence between the probe and the source. Thus source dependence alone changes whether this latent-varying direction receives a first-order descent signal.

C.4 Noise reuse can change the target's fixed point

Consider a generator $G(z) = \mathrm{sign}(z)$ whose endpoint law already matches the equally weighted binary target. A query that retains source noise nevertheless has variance

$$\mathrm{Var}(aG(z) + bz + c\xi) = a^2 + \sigma^2 + 2ab\mathbb{E}|z|, \tag{34}$$

whereas the standard independently corrupted target has variance $a^2 + \sigma^2$. For $a, b > 0$ these laws differ. Thus arbitrary source-noise coupling cannot be justified solely by the Gaussian marginal of the mixed noise. Matching a defined boundary through Eq. (14) avoids this ambiguity about which distribution constitutes the target.

D CONTINUOUS STUDENT-SCORE ESTIMATION AND CATEGORICAL CONSTRAINTS

D.1 DENOISING THE ACTUAL STATE

For $Y = cS + \sqrt{\beta}\xi$ with independent ξ, differentiate its Gaussian-mixture density to obtain

$$\nabla_y \log q(y) = \mathbb{E}\left[-\frac{y - cS}{\beta} \,\middle|\, Y = y\right] = -\frac{\mathbb{E}[\xi \mid Y = y]}{\sqrt{\beta}}. \tag{35}$$

Conditional expectation minimizes squared error, proving Eq. (12). The equivalent conditional-mean parameterization is

$$s_Q(y) = \frac{c\,\mathbb{E}[S \mid Y = y] - y}{\beta}. \tag{36}$$

For $Y = aM(z) + bz + c\xi$, with $\sigma^2 = b^2 + c^2$, the pre-corruption variable is $aM(z) + bz$ and the fresh variance is c^2. Regressing the endpoint and applying the original score conversion instead would give

$$s_{\text{endpoint}}(y) = \frac{a\mathbb{E}[M \mid Y = y] - y}{\sigma^2}, \tag{37}$$

$$s_Q(y) = \frac{a\mathbb{E}[M \mid Y = y] + b\mathbb{E}[z \mid Y = y] - y}{c^2}. \tag{38}$$

The expressions coincide under independent endpoint corruption ($b = 0$). In the conventional conversion with total variance $\sigma^2 = b^2 + c^2$, the required effective mean is

$$m_{\text{eff}}(y) = \frac{y + \sigma^2 s_Q(y)}{a} = \mathbb{E}\left[\frac{Y - \sigma^2\xi/c}{a} \,\middle|\, Y = y\right]. \tag{39}$$

The score is determined by this conditional mean, whose support relative to the token convex hull determines the representation bound below.

For comparison, consider independent endpoint corruption with $M_\theta(U) = E^\top p_\theta(U)$. Cross-entropy regression to $p_\theta(U)$ has population optimum $p_\psi(y) = \mathbb{E}[p_\theta(U) \mid Y = y]$. Multiplication by $E^\top$ gives $\mathbb{E}[M_\theta(U) \mid Y = y]$, which lies in the convex hull. The ordinary categorical score conversion is therefore exact in this case. More generally, unrestricted state-mean and fresh-noise predictors are related by Eqs. (12) and (36); matching their losses under this affine change of variables gives equivalent estimation problems.

D.2 THE REQUIRED MEAN FOR A GENERATED CONTINUOUS STATE

The representability condition in Eq. (5) can be made explicit for a generated state containing a soft token prediction and a continuous residual. Write

$$Y = S_\theta(U) + c\xi, \qquad S_\theta(U) = aM_\theta(U) + R_\theta(U), \qquad c > 0, \tag{40}$$

where $M_\theta(U) \in \mathcal{C}^L$ and $\xi \perp U$ is standard Gaussian. Differentiating the smoothed density gives

$$s_Q(y) = \frac{\mathbb{E}[S_\theta(U) \mid Y = y] - y}{c^2} = -\frac{\mathbb{E}[\xi \mid Y = y]}{c}. \tag{41}$$

This is the same denoising–score identity as for native diffusion, applied to the actual state and fresh corruption variance. For an inherited conversion with coefficients $a, \sigma > 0$, substituting this score into Eq. (5) yields

$$m_Q^{\text{req}}(y) = \mathbb{E}[M_\theta(U) \mid Y = y] + \frac{1}{a}\mathbb{E}\left[R_\theta(U) - \frac{\sigma^2 - c^2}{c}\xi \,\middle|\, Y = y\right]. \tag{42}$$

The residual term explains why a conditional mean of the token endpoint alone need not provide the correct score. Regressing the wrong conditional target would also produce an error for an image

model. The further restriction of a categorical head is its output range, which gives the pointwise optimal error

$$\inf_{m\in\mathcal{C}^L}\left\|\frac{am-y}{\sigma^2}-s_Q(y)\right\|^2=\frac{a^2}{\sigma^4}\operatorname{dist}(m_Q^{\mathrm{req}}(y),\mathcal{C}^L)^2. \tag{43}$$

For independent corruption of an endpoint, $R_\theta = 0$ and $c = \sigma$, so the correction vanishes and the bound is zero. The proof below establishes that generated states can also require means outside the hull. Appendix H separately gives an existence example in which this restriction affects the generator gradient.

D.3 Categorical score constraint

Proposition 2 (Categorical score constraint)**.** *Let $\mathcal{C}_L = \mathcal{C}^L$ be the product token-codebook convex hull. For $a, \sigma > 0$, define $m_{\mathrm{eff}}(y) = (y + \sigma^2 s_Q(y))/a$. Any score of the form $s_\psi(y) = (am_\psi(y) - y)/\sigma^2$, with $m_\psi(y) \in \mathcal{C}_L$, satisfies*

$$\|s_\psi(y)-s_Q(y)\|_2^2\geq\frac{a^2}{\sigma^4}\operatorname{dist}(m_{\mathrm{eff}}(y),\mathcal{C}_L)^2. \tag{44}$$

For a noise-interpolated query, m_{eff} can lie outside $\mathcal{C}_L$ even when $M_\theta(z) \in \mathcal{C}_L$ for every z.

Proof. By definition,

$$s_\psi(y)-s_Q(y)=\frac{a}{\sigma^2}(m_\psi(y)-m_{\mathrm{eff}}(y)). \tag{45}$$

Minimizing its squared norm over $m_\psi(y) \in \mathcal{C}_L$ gives Eq. (44). Integrating over Q gives a population mean-squared score-error lower bound, which is strictly positive when the conditional effective mean lies outside the hull on a set of positive Q probability.

For a concrete example of this possibility, take $\mathcal{C} = [-1, 1]$, $M(z) = \operatorname{sign}(z)$, and $a, b, c > 0$. Set $\sigma^2 = b^2 + c^2$, let ϕ_σ be the one-dimensional Gaussian density, and let Φ be the standard Gaussian cumulative distribution function. Direct integration over $z > 0$ and $z < 0$ gives

$$q(y)=\phi_\sigma(y-a)\Phi\left(\frac{b(y-a)}{c\sigma}\right)+\phi_\sigma(y+a)\Phi\left(-\frac{b(y+a)}{c\sigma}\right). \tag{46}$$

Write $q(y) = \phi_\sigma(y - a)R(y)$. As $y \to +\infty$, $R(y) \to 1$. At $y = a$, letting $k = b/(c\sigma) > 0$,

$$R(a)=\tfrac{1}{2}+\exp(-2a^2/\sigma^2)\Phi(-2ak)<1. \tag{47}$$

Since R is smooth and approaches a larger limit, $R'(y) > 0$ at some $y > a$, and by continuity on an open interval. On that interval,

$$m_{\mathrm{eff}}(y)=\frac{y+\sigma^2\partial_y\log q(y)}{a}=1+\frac{\sigma^2}{a}\frac{R'(y)}{R(y)}>1. \tag{48}$$

The bound is therefore positive on a set of positive probe probability. The example can also use a smooth categorical mean: $M_n(z) = \tanh(nz)$ converges almost surely to $\operatorname{sign}(z)$, and Gaussian smoothing makes both $q_n(y)$ and $q_n'(y)$ converge at each fixed y. A strict violation thus persists for sufficiently large finite n.

E Diagnostics for the categorical generator

For $m = E^\top p$ and $p = \operatorname{softmax}(\ell)$, a continuous instruction $r = \nabla_m \mathcal{L}$ reaches the logits through

$$\nabla_\ell\mathcal{L}=(\operatorname{diag}(p)-pp^\top)Er. \tag{49}$$

At uniform p, the categorical Jacobian has gain $1/V$ on the centered-logit subspace; as p approaches a one-hot vector, it approaches zero. Sharpening a prediction shared across inputs therefore leaves latent variation absent while suppressing its response to a continuous instruction. This identity concerns effective logits; a preceding cap or temperature transformation contributes its own Jacobian. For comparison, teacher-target cross-entropy has effective-logit gradient $p - p_T$. These statements

distinguish the fake's output restriction from the generator's response to a continuous score instruction.

To distinguish posterior uncertainty from variation across latent inputs, we consider the position-averaged quantities

$$H_{\text{pred}} = \frac{1}{L}\sum_i \mathbb{E}_U H(p_{\theta,i}(U)), \qquad H_{\text{mix}} = \frac{1}{L}\sum_i H(\mathbb{E}_U p_{\theta,i}(U)), \qquad I_{\text{latent}} = H_{\text{mix}} - H_{\text{pred}}. \tag{50}$$

For an auxiliary token $A_i \mid U \sim \text{Cat}(p_{\theta,i}(U))$, the difference equals $L^{-1}\sum_i I(A_i; U)$. It is neither the mutual information of argmax-decoded sequences nor a measure of joint language quality. A uniform posterior at every latent has high H_{mix} but zero I_{latent}; a shared one-hot posterior has low H_{pred} and also zero I_{latent}. Distinct confident predictions can have positive I_{latent} while still producing poor language. These quantities therefore accompany discrete sample duplication, position-conditioned token statistics, and generation-quality evaluation. The average unigram entropy *within* a generated sequence is a separate statistic and does not measure diversity across generated sequences.

F NIPSD Parameterizations and Optimization

This appendix gives concrete parameterizations and optimization details for the representative experimental settings.

F.1 Continuous representations and student trajectories

LangFlow uses a learned embedding codebook, which we hold fixed during distillation. Its variance-preserving path has $\alpha_t^2 = \text{sigmoid}(-\gamma_t)$ and $\sigma_t^2 = \text{sigmoid}(\gamma_t)$, where γ_t is the log noise-to-signal ratio. FLM uses one-hot token states, with $E = I_V$, $\alpha_t = t$, and $\sigma_t = 1 - t$; its lower-dimensional network input embedding is a separate map.

For a K-step student, coarse denoising transitions start from $S_{\theta,0} = \sigma_{t_0} U$ and take the form

$$S_{\theta,j+1} = \frac{\sigma_{t_{j+1}}}{\sigma_{t_j}} S_{\theta,j} + \left(\alpha_{t_{j+1}} - \frac{\sigma_{t_{j+1}}}{\sigma_{t_j}}\alpha_{t_j}\right) m_\theta(S_{\theta,j}, t_j), \tag{51}$$

where t_{j+1} is cleaner than t_j and $\sigma_{t_j} > 0$. For a linear path, this is the Euler update with velocity $(m_\theta - S)/(1-t)$; for the variance-preserving path, it is the update in noise-scaled coordinates. Generated intermediate states provide the inputs to Eq. (10). A one-step endpoint predictor can instead construct a state at a selected path location as $S_\theta = \alpha_s M_\theta(U) + \sigma_s U$, giving Eq. (15).

The initialization in Eq. (13) uses a teacher rollout from the same starting state for one-step generation. For segmentwise initialization, a forward-noised training sequence provides the starting state, and a local teacher transition supplies the lower-noise target. Expected-embedding regression is used for embedding-valued outputs; categorical supervision is used in one-hot coordinates.

F.2 Auxiliary output parameterizations

In embedding coordinates, a continuous head predicts the fresh noise directly. In vocabulary coordinates, we retain the analytic state term and augment a categorical prediction with a signed residual:

$$\widehat{s}_\psi(y,j,r) = \frac{\alpha_r p_\psi(y,j,r) - y}{\sigma_r^2} + R_\psi(y,j,r), \qquad \epsilon_\psi(y,j,r) = -\sqrt{\beta_{j,r}}\,\widehat{s}_\psi(y,j,r). \tag{52}$$

The residual is projected from hidden features into vocabulary coordinates, expanding the categorical score class without constructing a separate dense vocabulary-state backbone. Both parameterizations are trained using the fresh-noise target in Eq. (11).

F.3 Denoising anchor

We retain denoising knowledge with a forward-KL anchor on forward-noised training sequences:

$$\mathcal{L}_{\text{anchor}} = \mathbb{E}_{X\sim P_{\text{data}},t,\epsilon}\frac{1}{L}\sum_i D_{\text{KL}}(p_{T,i}(Z_t,t)\|p_{\theta,i}(Z_t,t))\,. \tag{53}$$

Its effective-logit gradient $p_\theta - p_T$ supplies token-level supervision alongside the continuous distribution-matching instruction. The anchor uses training data, while distribution matching uses generated states.

F.4 Practical generator update

For the state direction $g = g_{j,r}$ in Eq. (16), we normalize per sequence and construct a detached target:

$$\bar{g} = \frac{g}{\max\{\mathrm{RMS}(g), \rho\} + \varepsilon}, \qquad S^\star = \mathrm{sg}(S_{\theta,j} - \eta\bar{g}), \tag{54}$$

$$\mathcal{L}_{\mathrm{student}} = \mathbb{E}\left[w_{j,r}\left\{\sqrt{\mathrm{RMS}(S_{\theta,j} - S^\star)^2 + h^2} - h\right\}\right] + \lambda_{\mathrm{anchor}}\mathcal{L}_{\mathrm{anchor}}. \tag{55}$$

Here $\mathrm{RMS}(v) = \|v\|_2/\sqrt{Ld}$, $\rho \geq 0$ is a normalization floor, and $\varepsilon, \eta, h > 0$. The weights $w_{j,r}$ are those of Eq. (17). This loss implements a normalized robust surrogate for the score-difference update. It is distinct from taking an exact gradient step on the population KL.

For multistep students, we generate the prefix without gradients and differentiate through the current segment, yielding a local update with its prefix distribution held fixed. During auxiliary updates the generated states are detached; during generator updates the auxiliary and teacher predictions are stopped. The auxiliary network tracks the changing student distribution throughout training. Only the student generator is used for inference.

G Teacher Content Feedback and Distributional Transport

G.1 From posterior response to KL curvature

Let $P_{\mathrm{data}} = \mathrm{Law}(X)$, $\Sigma_X = \mathrm{Cov}(X)$, and $C_P(y) = \mathrm{Cov}(X \mid Z = y)$, where $Z = aX + \sigma\epsilon$. We use flattened sequence coordinates and assume $a, \sigma > 0$, finite moments, and sufficient integrability to exchange differentiation and expectation and to discard boundary terms in integration by parts. All identities in this subsection use the exact target conditional mean m_P. The experiments substitute the pretrained denoiser m_T and evaluate its response directly.

Conditional-mean Jacobian. For $Z = aX + \sigma\epsilon$, the posterior measure is proportional to $\exp(-\|y - ax\|^2/(2\sigma^2))\, dP_0(x)$. Its logarithmic derivative with respect to y is $a(x - m_P(y))/\sigma^2$. Differentiating the conditional mean therefore gives

$$\partial_{y_j} m_{P,k}(y) = \frac{a}{\sigma^2}\left(\mathbb{E}[X_k X_j \mid Z = y] - \mathbb{E}[X_k \mid Z = y]\mathbb{E}[X_j \mid Z = y]\right). \tag{56}$$

Equivalently, $\nabla_y m_P(y) = (a/\sigma^2) C_P(y)$. This identity includes cross-position dependencies. Differentiating Eq. (7) at $\lambda = 0$ gives

$$\begin{aligned} A_P'(0) &= a\,\mathbb{E}\big[D^\top \nabla m_P(Y_0) D\big] \\ &= a\ \mathrm{tr}(\Sigma_X\, \mathbb{E}_{Y_0} \nabla m_P(Y_0)) = \frac{a^2}{\sigma^2}\,\mathrm{tr}(\Sigma_X\, \mathbb{E}_{Y_0} C_P(Y_0))\,. \end{aligned} \tag{57}$$

The second equality uses $D \perp Y_0$ and $\mathbb{E}DD^\top = \Sigma_X$. The conditional covariance is that of the *target* posterior, evaluated at a query drawn from the *averaged student* distribution. Replacing that query distribution by the target's native noisy marginal would characterize a different response.

Distribution-matching curvature. Let ϕ_σ be the centered Gaussian density with covariance $\sigma^2 I$. The path density is

$$q_\lambda(y) = \mathbb{E}_D \phi_\sigma(y - a\mu - a\lambda D). \tag{58}$$

Because $\mathbb{E}D = 0$, its first derivative vanishes at zero. Its second derivative is

$$\partial_\lambda^2 q_\lambda(y)\big|_0 = a^2\,\mathrm{tr}\big(\Sigma_X \nabla_y^2 \phi_\sigma(y - a\mu)\big)\,. \tag{59}$$

The zero first derivative and conservation of density mass imply

$$\begin{aligned}\mathcal{L}'(0) &= 0,\\ \mathcal{L}''(0) &= \int \partial_\lambda^2 q_\lambda(y)\big|_0 \log \frac{q_0(y)}{p(y)}\, dy\\ &= a^2 \mathbb{E}_{Y_0} \operatorname{tr}\big[\Sigma_X\{\nabla^2 \log q_0(Y_0) - \nabla^2 \log p(Y_0)\}\big]. \end{aligned} \tag{60}$$

Now $\nabla^2 \log q_0 = -I/\sigma^2$ and $\nabla^2 \log p = -I/\sigma^2 + a^2 C_P/\sigma^4$. The isotropic terms cancel, yielding $\mathcal{L}''(0) = -(a^2/\sigma^2)A_P'(0)$. In particular,

$$\mathcal{L}(\lambda) = \mathcal{L}(0) - \frac{a^2}{2\sigma^2} A_P'(0)\lambda^2 + o(\lambda^2). \tag{61}$$

The content-response slope and the restoring curvature describe the same local obstruction. This is a statement about a distribution path near its shared mean, not about the full parameter Hessian of a neural generator.

G.2 A TARGET-RESTORING DIRECTION WITH A DISTRIBUTIONAL MEANING

Write $\widetilde{Q}_\lambda = \mathrm{Law}(M_\lambda)$ for the clean contracted law. For $\lambda > 0$, the affine map $T_\lambda(m) = \mu + (m - \mu)/\lambda$ takes $\widetilde{Q}_\lambda$ to P_0 and attains

$$W_2^2(\widetilde{Q}_\lambda, P_0) = (1-\lambda)^2 \operatorname{tr}\Sigma_X. \tag{62}$$

To verify optimality, any coupling of M_λ and X' with $X' \sim P_0$ can be written using $X = \mu + (M_\lambda - \mu)/\lambda \sim P_0$. Its squared cost is

$$(1+\lambda^2)\operatorname{tr}\Sigma_X - 2\lambda\, \mathbb{E}\langle X - \mu, X' - \mu\rangle. \tag{63}$$

Cauchy–Schwarz bounds the last expectation by $\operatorname{tr}\Sigma_X$, with equality at $X' = X$. Thus $D = X - \mu$ points along a minimum-squared-distance restoration of the target law, rather than an arbitrary pairing of samples.

For independent Gaussian corruption define the state gradient and its conditional noise average by

$$g_\lambda(X,\xi) = \frac{a^2}{\sigma^2}\{m_{Q_\lambda}(Y_\lambda) - m_P(Y_\lambda)\}, \qquad \bar{g}_\lambda(X) = \mathbb{E}_\xi g_\lambda(X,\xi). \tag{64}$$

The reported alignment is the cosine in the space of square-integrable sample-dependent updates:

$$\mathcal{A}(\lambda) = \frac{-\mathbb{E}\langle D, \bar{g}_\lambda(X)\rangle}{\sqrt{\mathbb{E}\|D\|^2\, \mathbb{E}\|\bar{g}_\lambda(X)\|^2}}. \tag{65}$$

It measures progress in the restoration direction per unit root-mean-square state update. In the measurements, $\bar{g}_\lambda$ is estimated with eight independent noise draws per sample. The coordinates are the model's native pixels, latents, or token representations; this squared-distance criterion is distinct from a decoded-sample quality metric.

G.3 STUDENT-SCORE REFERENCE ALONG THE AFFINE PATH

The affine path permits a student-score reference without fitting an auxiliary model. Suppose the target family is $P_t = \mathrm{Law}(a_t X + \sigma_t \epsilon)$. At a fixed probe condition (a, σ), choose a supported condition t' and a positive scalar k satisfying

$$k a_{t'} = a\lambda, \qquad k\sigma_{t'} = \sigma. \tag{66}$$

Then Y_λ has the same distribution as $a(1-\lambda)\mu + kZ_{t'}$. With $y' = [y - a(1-\lambda)\mu]/k$, a change of variables yields

$$\begin{aligned} s_{Q_\lambda}(y) &= k^{-1} s_{P_{t'}}(y'),\\ m_{Q_\lambda}(y) &= (1-\lambda)\mu + \lambda m_{P_{t'}}(y'). \end{aligned} \tag{67}$$

For a variance-preserving schedule, $k = \sqrt{a^2\lambda^2 + \sigma^2}$, $a_{t'} = a\lambda/k$, and $\sigma_{t'} = \sigma/k$. For EDM's variance-exploding coordinates, $k = \lambda$ and $\sigma_{t'} = \sigma/\lambda$. For FLM's linear path, $k = a\lambda + \sigma$ and $t' = a\lambda/k$.

Equation (67) is exact for the population target score. The experiment substitutes the same pretrained teacher at the original and transformed conditions, yielding a *teacher-based score reference*. The source data are held-out real samples rather than exact draws from a law recovered by the neural teacher. Consequently, the measured alignment includes teacher approximation error; the teacher-response measurement in Figure 1a does not require this student-score construction. DiT uses the nearest supported discrete conditioning level and the corresponding actual contraction; nominal $\lambda = 0.1$ is approximately 0.100414 in its score-reference measurement.

G.4 Cross-modal measurement protocol

We evaluate the four frozen teachers described in Appendix J: EDM on CIFAR-10 pixels, DiT on ImageNet VAE latents, and LangFlow and FLM on length-1024 OWT token sequences. Each model uses two independent data/noise groups of 64 examples. The mean μ and, for language, each $\bar{p}_i$ are computed within the group. EDM uses a fixed class within each group (classes 0 and 5); DiT uses its trained null class. No parameter updates enter the intervention.

Each teacher is evaluated at a fixed conditioning level selected by the earlier denoising calibration targeting normalized reconstruction error 0.5. Table 4 gives the selected conditions and the achieved full-content errors on the present groups. These conditioning levels remain fixed as λ changes in the content-response measurement.

Teacher	Signal a	Noise σ	Full-content error, two groups
EDM	1	6.4048	0.530, 0.569
DiT	0.32627	0.94528	0.574, 0.546
LangFlow	0.08002	0.99679	0.503, 0.511
FLM	0.76409	0.23591	0.483, 0.500

Table 4: Conditions for the content intervention. DiT uses timestep 468, LangFlow uses log noise-to-signal ratio 5.04457, and FLM uses physical time 0.76409. Errors are normalized squared denoising errors, rather than generation-quality scores.

Teacher content response uses two independent noise draws per sample, shared across λ and the baseline query. Each group's response is normalized by its own $A_T(1)$; the main figure reports the mean and range across the two groups. These ranges are not confidence intervals or variability across independently trained models. The statistical unit is the whole example. Inference and readouts use FP32, with scalar reductions in FP64.

The content-response grid is $\lambda \in \{0.03, 0.1, 0.3, 0.6, 1\}$. The main score-alignment panel uses $\lambda \in \{0.1, 0.3, 0.6\}$, for which EDM's transformed score queries remain within its supported noise range ($\sigma_{t'} \leq 64.05 < 80$). The content-response panel always uses the fixed probe condition and entails no transformed-noise extrapolation.

As a check on the interpretation of content response, the teachers' token top-1 denoising accuracies at $\lambda = 0.1$ are 1.84% for LangFlow and 0.87% for FLM; at $\lambda = 1$ they are 48.36% and 60.61%. These refer to noisy real-source reconstruction under the intervention. The experiment isolates how the fixed teachers respond to controlled content loss and how their score differences act in state space; it does not estimate the full neural generator's parameter update.

Finally, independent exact-density calculations validate the score transformation, the KL directional derivative, and the local curvature identity on a two-point Gaussian mixture. The maximum absolute errors are 7.11×10^{-15} for the transformed score, 5.07×10^{-10} for the first derivative, and 5.42×10^{-8} for the second derivative.

H A Population-Level Reversal of the Distillation Direction

Let U, ξ be independent standard Gaussian variables, $M_\theta(U) = \tanh(\theta U)$, and $Y = M_\theta(U) + bU + c\xi$, with $b^2 + c^2 = 1$. This generator is the expected embedding under two-token logits $(-\theta U, +\theta U)$ and embeddings $(-1, +1)$. With ϕ the standard Gaussian density, its probe density and the teacher

target are

$$q_\theta(y) = \int \phi(u)\frac{1}{c}\phi\left(\frac{y - \tanh(\theta u) - bu}{c}\right) du, \tag{68}$$

$$p(y) = \tfrac{1}{2}\phi(y-1) + \tfrac{1}{2}\phi(y+1), \qquad s_P(y) = \tanh(y) - y. \tag{69}$$

We compute $s_Q = q'_\theta/q_\theta$ directly from this integral. The optimal token-CE prediction has mean $m_{\mathrm{CE}}(y) = \mathbb{E}[M_\theta(U) \mid Y = y]$ and induces $s_{\mathrm{CE}}(y) = m_{\mathrm{CE}}(y) - y$. The mean required for the correct score is $m^*_Q(y) = y + s_Q(y)$. Minimizing population score MSE subject to the categorical range $[-1, 1]$ gives

$$s_{\mathrm{catDSM}}(y) = \mathrm{clip}(m^*_Q(y), -1, 1) - y. \tag{70}$$

This is the optimum in the closed categorical function class. Finite softmax logits approach its boundary values; the strict gradient-sign differences below persist for sufficiently accurate approximations. An unrestricted fresh-noise DSM estimator gives s_Q at its population optimum.

The choice $a = \sigma = 1$ is a coordinate convention. Scaling the observed student and target states by $1/\sqrt{2}$ gives $\alpha^2 + \sigma^2 = 1$; scaling by $1/2$ gives $\alpha + \sigma = 1$. Such common invertible rescaling leaves the KL and its parameter derivative unchanged, so the example also admits variance-preserving and linear-path coordinates.

For each score $\widehat{s}$, the stopped-score distillation update uses

$$\widehat{g}(\theta) = \mathbb{E}\big[U \operatorname{sech}^2(\theta U)\{\widehat{s}(Y) - s_P(Y)\}\big]. \tag{71}$$

Only $\widehat{s} = s_Q$ necessarily recovers $\partial_\theta D_{\mathrm{KL}}(Q_\theta \| P)$. At $\theta = 1$, $b = 0.6$, and $c = 0.8$, we obtain:

Auxiliary score	$\widehat{g}$	Actual KL change at step 0.01
Correct score / unrestricted DSM	$+0.008625858$	-7.43815×10^{-7}
Optimal token CE	-0.056176572	$+4.85585 \times 10^{-6}$
Optimal categorical DSM	-0.000686591	$+5.92259 \times 10^{-8}$

The final column recomputes the actual KL at $\theta - 0.01\widehat{g}$, rather than estimating improvement from a surrogate. With equal parameter displacement 10^{-3} in each proposed descent direction, the correct update changes KL by -8.59367×10^{-6} and both categorical updates change it by $+8.65796\times10^{-6}$. Thus the reversal is not an update-scale artifact. With $b = 0, c = 1$, all three gradients at $\theta = 1$ equal -0.052299001, providing an independent-corruption control.

Gaussian expectations are evaluated with 128, 256, and 512 Hermite nodes; the corresponding y grids have 6001, 12001, and 24001 points on $[-13, 13]$. Independent adaptive integration over $u \in [-12, 12]$ agrees with the highest-resolution gradients to within 6.36×10^{-9}. A central difference of the KL with parameter displacement 10^{-4} agrees with the exact-score gradient to within 4.27×10^{-10}. The numerical density mass error is below 2.3×10^{-16}. This example establishes that a categorical estimator's mismatch can survive the generator Jacobian and reverse the learning direction. It does not assert that such a reversal occurs at every state or dominates the full-model experiments.

I Controlled CDLM Transfer Diagnostics

Initialization and source correspondence. The LangFlow-DI comparison uses OWT sequences of length 1024, three independent training seeds, a common 1000-step teacher-trajectory initialization for initialized branches, 512 fake-only steps, and 1024 joint steps. Final states are evaluated on 256 generated sequences per run with the same GPT2-large evaluator. Table 5 reports arithmetic means over the three runs. Source shuffling preserves the added-noise marginal while breaking its correspondence with the current generated sample. Under this construction the shuffled and independent probes have the same full-sequence marginal law; their finite-training results need not be identical.

Teacher-copy branches with independent and noise-interpolated probes have GenPPL 1.87 and 1.97, respectively, alongside repeated-4-gram fractions 0.9836 and 0.9834. These low perplexities reflect severe repetitive degeneration. Among initialized branches, the paired-seed GenPPL ratio of noise-interpolated to independent is 0.869 with 95% interval $[0.841, 0.897]$; relative to source-shuffled it is

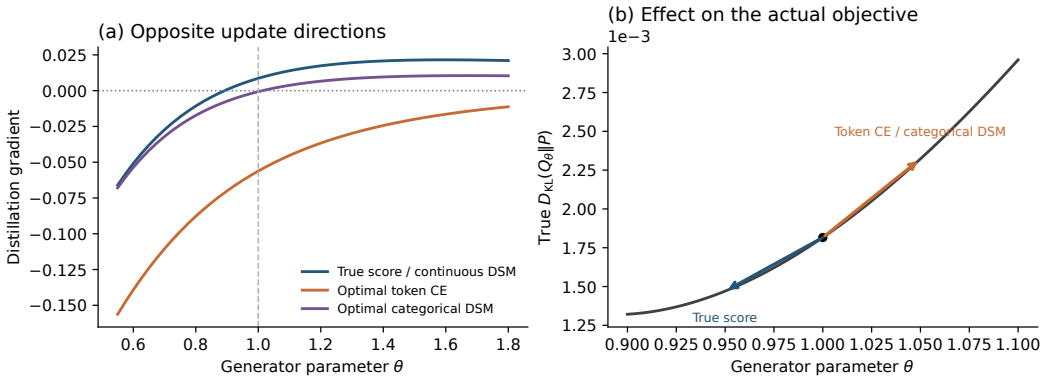


Figure 2: **A categorical score mismatch changes the direction of distillation.** A one-token example with population-optimal auxiliary estimators. Left: the actual KL gradient and the updates induced by token CE and by optimal categorical score MSE. At $\theta = 1$, both categorical estimates have the opposite sign from the true gradient. Right: their descent directions increase the actual objective. Arrows use equal parameter displacement to separate direction from magnitude.

Generator start	Probe	Repeated 4-grams	GenPPL
Teacher copy	Independent	98.36%	—
Teacher copy	Noise-interpolated (NIP)	98.34%	—
Trajectory-initialized	Independent	0.29%	266.29
Trajectory-initialized	Noise-interpolated (NIP)	0.43%	231.36
Trajectory-initialized	Source-shuffled	0.29%	261.59

Table 5: Controlled LangFlow-DI transfer diagnostics on OWT-1024, averaged over three training seeds. Severe repetition makes the low perplexities of the teacher-copy branches unsuitable as quality scores; their values are reported in the text.

0.885 with interval $[0.851, 0.920]$. Intervals use paired run-level mean-token NLL differences and a three-seed Student-t interval before exponentiation. The noise-interpolated branch also has slightly lower within-sequence unigram entropy and higher repetition. This grid establishes the effect of the initialization intervention in this setting; it does not isolate posterior variance from all other changes learned during initialization, nor does it establish initialization or NIP as universal necessities across the four instantiations.

Auxiliary target and range. For frozen generators, we train token-CE, categorical-DSM, and signed-residual auxiliary variants for 512 steps on matched probes. In the initialized source-dependent setting, the paired held-out DSM-risk difference, categorical DSM minus token CE, is -3.091×10^{-4} with 95% interval $[-3.575, -2.608] \times 10^{-4}$ for LangFlow, and -1.739×10^{-5} with interval $[-2.321, -1.157] \times 10^{-5}$ for FLM. These are differences in noisy-label DSM risks; their common irreducible conditional variance cancels. Independent-corruption controls do not show the same CE-to-DSM improvement.

A separate conditional-support test evaluates a necessary inequality for $m^*_Q(Y) \in \mathcal{C}^L$ using directions that depend only on the observed full sequence Y. Across 144 tests, 18 remain positive after Bonferroni correction; both language families have detections, and no independent-corruption control does. These tests establish a representational mismatch, separately from finite-network fit. A full-space score-error lower bound alone does not imply a nonzero generator-gradient error; Appendix H supplies a separate constructed example where the latter does occur.

Categorical DSM nevertheless yields effective generators in both language families. In the controlled LangFlow runs its GenPPL is 201.01 for DI and 150.43 for TDM, compared with 231.36 and 157.12 for the tested signed-epsilon variant. FLM's categorical-DSM branches also succeed. Thus neither the support test nor the one-token counterexample establishes universal superiority of an epsilon parameterization. The transferable requirement is to estimate the appropriate student-state

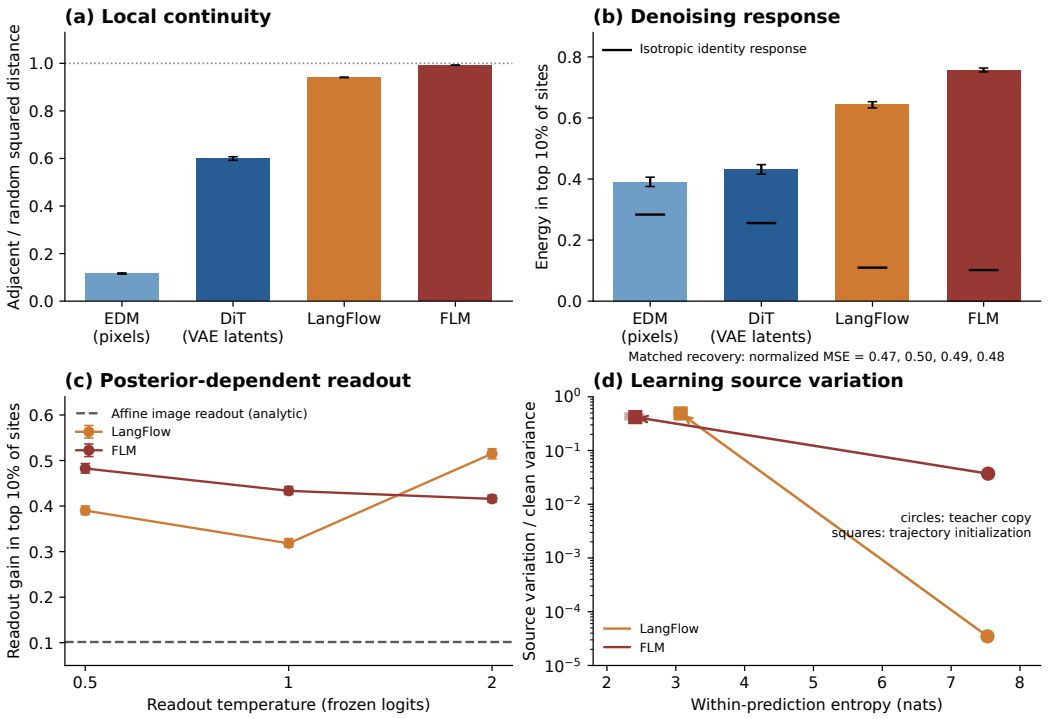


Figure 3: **Auxiliary representation diagnostics.** (a) Clean-state local continuity. (b) Frozen-teacher finite-response concentration at matched recovery, with dimension-matched isotropic identity controls. (c) Categorical readout gain under a frozen-logit temperature intervention; the affine image reference is analytic. (d) Soft-output variation across source noise versus within-prediction entropy, before and after teacher-trajectory initialization. Error bars in (a)–(c) are whole-example bootstrap intervals. These measurements are distinct from the actual distillation-gradient analysis in Figure 2.

score; the practical sufficiency of a restricted approximation depends on the state and its effect on the generator update.

J Representation and Response Analysis

Role of these measurements. The statistics below characterize clean-state geometry, frozen-teacher input responses, and categorical readout gains. They do not measure the student parameter update in Eq. (3), its alignment with the true KL gradient, or the effect of a distillation step on generation quality. They serve as representation diagnostics; the population-level score counterexample in Appendix H and the cross-modal content interventions in Figure 1 establish transfer difficulties.

Across 1536 examples per model, adjacent-to-random squared-distance ratios are 0.116, 0.600, 0.941, and 0.992 for EDM, DiT, LangFlow, and FLM. At normalized recovery error near 0.5, the most responsive 10% of sites contain 39.1%, 43.1%, 64.3%, and 75.7% of finite-response energy. These observations establish stronger local continuity in the measured image states and greater response concentration in the measured CDLMs. Concentration alone does not determine whether a score instruction points in a useful direction.

Models and data. We use the official class-conditional EDM CIFAR-10 checkpoint, DiT-XL/2 at image resolution 256, and the frozen OWT LangFlow and FLM teachers. EDM operates on 32×32 RGB pixels. DiT operates on $4 \times 32 \times 32$ VAE latents; its cached real-image inputs are posterior means scaled by 0.18215. The language inputs are length-1024 validation sequences. Image conditioning uses the ground-truth class without classifier-free guidance. The diffusion site count is 1024 in each case; it is distinct from the number of transformer patches in DiT. FLM

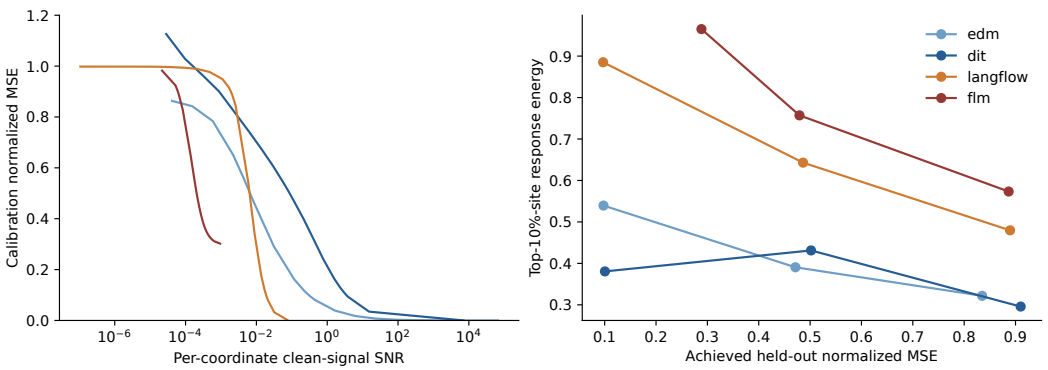


Figure 4: **Calibration and retained response measurements.** Left: normalized reconstruction error over each model's native conditioning range, plotted against per-coordinate signal-to-noise ratio. Right: finite-response concentration at every evaluated condition, located by its achieved held-out error. The FLM endpoint does not reach normalized error 0.1. The main comparison uses the shared recovery target 0.5.

is analyzed in its native vocabulary coordinates, rather than its lower-dimensional network input embedding.

Static measurements. For each model, three disjoint subsets contain 512 examples each. Image adjacency uses the four-neighbor lattice; text adjacency uses consecutive tokens. The random-pair denominator uses different positions in the same example. Row energies and feature covariance use one global feature mean per subset. Column-variance statistics are compared before and after an orthogonal feature rotation; for FLM an orthonormal DCT permits exact sparse computation. Covariance-spectrum participation ranks and analytic finite-dimensional Gaussian references are retained as additional diagnostics. Gaussian reference approximations are not measurements of model outputs.

Recovery calibration. On 128 independent examples, we evaluate a native conditioning grid and refine conditions targeting normalized reconstruction errors 0.9, 0.5, and 0.1. DiT uses supported integer scheduler timesteps. EDM uses the official sampler range $[0.002, 80]$. LangFlow uses its native log-noise range and two-pass self-conditioning; FLM uses its native conditioning-to-physical-time conversion. The variance normalizer is fixed from the first static subset. Three new subsets of 64 examples provide held-out measurements. Figure 4 shows the calibration curves and retained measurements. A target passes the recovery gate when its held-out mean error is within 0.05 of the target. All four models pass at 0.5, which is the comparison in Figure 3b. EDM does not pass at 0.9 (achieved 0.835), and FLM does not pass at 0.1 (achieved 0.288). These points are retained in the full results and are excluded from matched comparisons at those targets.

An initial pilot used 16 calibration examples and native mixed precision. Recovery mismatch in that pilot motivated the larger calibration set and new evaluation examples. The main response measurements use FP32 for all models. FLM uses its existing explicit attention implementation, preserving rotary embeddings and attention softcapping. Codebook readout gains are computed in FP64. Identical-input repeats produce zero difference in the first batch of each condition and seed. Paired perturbation directions are shared between doses $\eta = 0.05$ and 0.10; their measured responses are finite differences, with no assumption of an exact linear regime.

Concentration and statistical units. The finite response is $d_\eta = [m(z + \eta\sigma v) - m(z - \eta\sigma v)]/(2\eta\sigma)$ for $v \sim \mathcal{N}(0, I)$. For the categorical readout, $B_i = E^\top(\mathrm{diag}(p_i) - p_i p_i^\top)$ and $g_i = \mathrm{tr}(B_i B_i^\top) = \sum_v p_{iv}^2 \|e_v - m_i\|^2$. The latter is a direction-averaged local gain, rather than the response to the actual teacher–student score difference.

For nonnegative per-site energies, concentration is the fraction in the largest $\lceil 0.1L \rceil$ entries. Denoising response uses all 1024 sites. Readout gains use the same uniformly selected panel of 128

positions in both language models, so the uniform-gain reference is $13/128$. An additional isotropic identity-map control simulates the exact χ^2_d distribution of independent Gaussian row energies at the native L, d, giving top-site fractions of 28.35%, 25.58%, 10.98%, and 10.17% for EDM, DiT, LangFlow, and FLM. This controls the sampling concentration introduced by feature dimension alone. Confidence intervals resample complete examples $4{,}000$ times; positions are never treated as independent examples. Frozen-teacher seeds index data and noise repetitions, rather than independently trained teachers. The temperature intervention holds logits fixed; the reported gain is with respect to effective logits ℓ/T. Differentiating with respect to the original logits adds a common $1/T^2$ factor, leaving concentration unchanged.

Source variation. We evaluate three existing 1000-step one-step trajectory-initialization checkpoints per language family, using 64 independent noise pairs for each checkpoint. Teacher-copy and initialized generators share the same pairs and native DI endpoint condition, with zero self-conditioning. These measurements use native language inference. Source variation describes continuous soft outputs, separately from decoded token disagreement or generation quality. In particular, FLM's teacher copy already has 99.26% token disagreement between paired argmax outputs despite its much smaller soft-output variance. Image teachers also exhibit posterior averaging: at their native initial conditions, paired source variations are 0.0151 for EDM and 0.1246 for DiT, with class fixed within each pair. Their initial signal-to-noise ratios differ, so these image values are descriptive and are not used as a matched cross-modal comparison.

The readout Jacobian characterizes the categorical conversion, rather than the full parameter Jacobian. The four-model response comparison includes differences in architecture and training data. It establishes the reported response geometry for these models; the fixed-logit intervention separately isolates the dependence of the output conversion on the posterior.

K GENERATED TEXT SAMPLES

Examples are grouped by model and distillation method. Each box records the dataset and number of function evaluations (NFE).

DI–LangFlow Example 1

Dataset: OpenWebText NFE: 1

<|endoftext|> She tells me about her job with Google because she was able to work a new company last coming from Harvard's Met Beijing last year. She says that she had been working with her because she did not work and she was want to make her company.

Carger says she's working for this job $1 million. She has no question that She will not give a new conference in January 2017. She wants't living in these same China.

There are two other people (1) She is very young. That fact, she says she work with her family because she is trying to work her life is not very good. For example, I still won't have an job for the company but about $1,000. Because these company is very expensive, do we allow other people to work? There are 2 million people working at Google (they no money? How are going to work?

But she doesn is going to tell them?

She does not know that she's not working because she's got an pay out of 1 million dollars.

Her Google

Also, she says that she doesn't work. Because the company has only about $1,000 last salary and She's working at Google. China says she knows that $1,000 in doing their job (the tech company).

Conger also called it "pure" people who want to work and have very money. However, she says that he or her new income will not work when the company comes back for the company. He doesn't know this is only because his wife want to work but he is working because he doesn't even see his money but doesn't.

They're working here in China because they are have any money for the same time, because They have a lot of companies who tell me there are no "very many companies" in a "new company" don't come back to them when they work here because they have a lot more than us. They don't have more than 10% and they want to at any more than one (1) or "big companies"). Some people say they want to pay their $1,million) but for the first time which is that we do not work. In China, we have a good $100 (or) little10 million (day) $10 million. You will not lose $10, but you don't want to come anything like this for more money because there are many people in China who say you can do them. You may be a few problems because more than 2 million people will not work in the country but do you work again.

And you won't work here and when most other people will not work at it. There are many other companies working in China. There's difficult to work because I will not have any other company in China because I am not working to replace one of my own company. In China, I don't think able to get money. But I don't be concerned about this because you don't want to work in the United States this way. This is possible because many people have wanted to work at another company in China because of what happening and there are some people who make their company's work. If you work in the other part of this company, there are more than 100 people who work. It is unclear how many of these other people are not goodt going to work, but the problem will be at now.

And for the first time, company they're trying in work with China in China. This is working from other company outside of China. That means they have over 100 jobs.

Over 10 million companies

(My company: There are 10 million companies companies who are two-thirds ($1,000. dollars), but they will get some much money jobs when they China, but it's not too much. people don't know to get them back in the year. There new high

The company doesn't have a new company on it, but there is no way to get it'1.1 million ($10 million) jobs in the company, Google or "new company," which is about 100% new companies working in China, even because they are trying to work with other companies around China. They say company says they will not be doing work because of the government? Who will any money about a new company because he can say that jobs will be used by no one an investment worth Of than10 million dollars? There are some people who don't even know more than 100 large-based companies, which is an new business

DI–LangFlow Example 1 (continued)

company, but there has no more than other companies in China Korea, China, and the United States. With other companies such as China,<|endoftext|>

DI–LangFlow Example 2

Dataset: OpenWebText NFE: 1

<|endoftext|> when they work:

I'll be like to work at my company in Mexico. The job is being work in 10 country. I think this would be going to go later than me. I didn't want to work in New York, but I think he came up in Boston, I thought he would be launching his own company (for his new company) about $100 for the last 10 years for all other companies companies. They are also paid with at least $1 million. The money been broken out by Corporations that for more than 10 years. I know that that these companies have a financial money of $1 million, revenues of $1 per million. So How do we want to work for here? A "So we'll want to pay a lot more than $1 million when we worked here: This is why many people people (sic) but right least she wanted to come back for their money: Of fact that really don't think what's family for 1 or 10 dollars at our company here. Many of us are new people who have more than 10 years in the US, but they are now. It's good thing to work here than $1 million. These companies will be paid for least $1. They can make an offer, but it isn't true. There are some of the people who make money from the new company. These companies who just pay $1 million and take over 10 million dollars from our company. They're just 10 years after they came to Chicago (more than 10 years ago) have been own money for me. There have been several New Zealand American companies in Boston and New York (and the company) since I started, but I will pay out my job backeven if I do). So I have 10 years ago we got $10 million here during the last 10 years. They have been by having so many people who paid 10 million in a German company; How can they have paid over $1 million million dollars?

I think whose jobs was one of all of them. They were leave an States in the America. But they have a good situation. We will be so many other companies these companies' businesses because they won't leave their job? So also won't give them "real money" come work. But I think there are many people in the own company and trying to come to this: The company who still trying to start her own company here. She is not saying that she will have an income (an salary of $1,000) (which is here). She said she will come back for $1 million for her own company. She will work because she still doesn't have any money to come back. But now she does not going to pay more than a new salary than 10 years, if there is no money where she will have her job: In 10 week ago she said 'fry company" or 'new company" when she would here:

And she said she was 100%-Freeting:

"Our company will have so much money, so we will want to come back.

If she goes for our family job (even) she gets not 10). There's been this company for 2008, these they are not in the same way and go get an company for $ least-$ $1 dollars. It will happen if they come up and get their company in the same way:

Don't do it again, but I don't want to come up more than two companies. That's no best more than them (and possibly others) in the company: Msacey (sic):

But I don't want to go up more than one company. "civilatives" as I am wrong. I're more than 1 can feel like there's no US here. It will be far too much than because they don't want to work and what is only going to keep the company's company: But she doesn't know if she won't come back in the US States. She says she will go to her Public Department office in the United States. She says she will come back to help's company. But she's hard because she works back in Chicago. She will not go to everyone's lending. She's worried to have a working as a company. When she sees her job in New Mexico (in other country). She works again with her family so she will not come back. She says she is back in Mexico since 2008. She has a family, and she wants to be able to get to Mexico (above) She comes to Canada (but she hasn't had more than go to work:

(dot) She doesn't working because she wants to help with her<|endoftext|>

TDM–LangFlow Example 1

Dataset: OpenWebText NFE: 4

<|endoftext|> surgery because he was not going to get away from a season with the Cubs back in two years.

The surgery didn't lead to his recovery, as Bisby said. But it's still not like he was able to stay in St. City, and it still wasn't as that Bisby was leaving.

"We all know that he's coming out for a year now, and so our momentum has changed," he said. "We all know that we are going to stay in that situation and go as soon as possible, and what we'll do."

The USB's 2017 process will coincide with the growth of the team. It is in order for him to find out in a team that can be successful, but his process has gotten slower so far this season.

Heneata said he has told Bisby that he will continue playing 2017 at a team that started in May of 2017. If the team does not at Bisby's contract extension, Bisby has a chance to go back anytime soon.

"It's going to move him," Heeaton said. "He has a good career. "He has a good career. He wants to want to play a lot. Now he has a good career. I know he likes him, but I will push him back soon."

Heneata started Bisby's career, but he went into a stint in Chicago.

Back in January of 2017, Dolan Heeat, Bisby's best partner in the spring and began last season against the Toronto Red Sox, said he will continue to stay on the team until more than later. "He's spent a lot of time with lots of the different ideas of how he comes out of our team," he said. "If he's an important player for our team, then I will find out and hold him in my direction and see what they do."

Maybe Heeat's going to stay behind him.

But this season, Bisby's problem has been I'm spending much more time than what he does and is going to need to stay full-time more, but at the end least, I'll be more comfortable at getting a long-term job.

Ballyby's team has been one of the best teams assembled I've ever seen: an team up at the bottom of the season at Mr.'s age.

Heeas said he is not considering returning to Chicago.

"I miss that a few years from now," he said. "If you do slap me in my direction when I think about this move, I think it's something I have enjoyed over the last year and years. I'm still looking forward to it."

However, Heeat has already earned a $4.5 million deal in his deal with the Cubs.

He's going to have to exit with Mr. Louis's new team. He's 30.

But Heeat also said that going from age six to six hours a year ago, isn't really in his favor. "I saw my father last week and he didn't know about it anymore, but I think he's 30," he said. "He will be back to Detroit. That's what's really going to happen. Do you do on coming to Detroit? That's going to change our business.

Diki Alonso (@dalawkettel)

Two weeks ago I learned that my son would have been returning to the USB. With Bisby's first contract, the news is that the USB is doing the right thing. He wants to stay in Detroit — scheduled to take place on June 4, 2016 — after he's been on Chicago for four weeks.

Ballyan's wife, Susan told me about Saturday's first team game with Bisby: "We're in a good place. I knew we had to talk about it. Again, too." Worani said that when she heard why she's being unsettled, she's a fan, and it really excited one. "I told myself, the fan, it'll be like, 'What's against me?" Hadani said. "I'll be there my team and Aaron (my team) and my team. By the end of the season, we want it. We want it."

Ballyan has a little time to meet up with her<|endoftext|>

TDM–LangFlow Example 2

Dataset: OpenWebText NFE: 4

TDM–LangFlow Example 2 (continued)

<|endoftext|> going to a conference. There are people in Canada talking there about how we're going to move.

Don't know here. It is Rusrik Berchov from Toronto Capitals. There is no way he's ready for another year.

This is playing on this year's NHL UHL Junior League Team. Every team he is playing. How are they going to expect him to be this season?

And it's worse than they were about.

"We look at them. We'retying with them but we want to see how the future is what it is going to be because we have the options and the best as we can" Cessel says. That's the best deal they have of their own players right now. But there's only a lot of players who aren't free agents.

Kkechuk speaks to a lot of people that the Capitals will have to move by.

It's Bergenov, because because we know what moves him. He's not being that a lot of people are right now. NHL

The decision to trade Philchuk didn't know whether he would go again, but they told us that he has a preterm contract that allows him to go for one year.

So, Gretchuk is thinking he will move a long enough to have to be traded.

Rusrik Berchov was signed last year. He is a 26-year-old former Leafs Blue Jackets goalie, who plays with hockey and the Montreal Canadiens, where he has played for years.

Berchov signed three years ago and he was a 22-year-old that he was playing for the team.

When it started to talking about, Bergenov just wanted to move.

We didn't have to make Stanley Cup (two wins from 2014 in three seasons) or what we wanted to do.

"We had know that we're going to do a lot because he's a really skilled player and he is a great playerist," he tells me. "And he is a great player who knows that you can help the team."

Bergen wanted to come back last year and realized that he would be able to stay this season.

After his first NHL season with the Capitals of the Islanders, Berchov was selected to become the NHL's top 25 player and he's earned the MVP of the award this year. He's a player who earns up to $28,000 in summer bonuses. He was also named a CM to the club's team and was named Stanley U Hockey Team each this year. Bergenov has played 17 games for the Leafs. Before making any part with NHL, he played a season-high 11 plus-minus games with the Leafs.

At the end of year, the Bergenars played a total of 40 plus-minus. When Berchov plays for the Islanders and was traded to the Leafs, it took about a year before he's going to make the team. If we started out we'd tell him.

Now, if you are saying us we'd be a better better player for the Canadiens because we know what we're going to do! If we can't have contracts" then we have to go on what we're doing now.

We are taking players now.

However, a lot of people will say about it if you are wondering if you should or not, do you know how people go around the world?

Kkechuk says that he thought players were too close. He describes how many times of the career in which players have made their moves.

Bergenov says that he did not know, "You can't make a move. That's not what people are really coming up now. We already know that we know that if players are going to get off, we have to do it.

"We're going to do it."

Kkechuk.

In the long run, the NHL knows that players are many better players. There's not a trade deal and Cessel will return and they have a deal with another team. They extend their contracts for a couple of years. But we want to be successful. I don't believe we're going to do for what we did for a year. We "We do it. We can make these decisions." "We want to make the best, get back the best. If he fails, we're going to get into the future. . . and we want to be working on to make these players as well.

TDM–LangFlow Example 2 (continued)

If he’<|endoftext|>