WHERE SHOULD DIFFUSION ENTER A LANGUAGE MODEL? GEOMETRY-GUIDED HIDDEN-STATE REPLACEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Continuous diffusion language models require choosing a representation space in which denoising operates, yet it remains unclear which representations are most compatible with diffusion. We study a basic design question: where inside a pretrained language model should continuous diffusion operate? We formulate this as a hidden-state interface-selection problem and hypothesize that diffusion-friendly interfaces can be identified from representation geometry. We operationalize this hypothesis using three training-free geometric proxies: local compactness, global stiffness, and effective rank. Across two 8B-scale backbones, the resulting geometry score predicts fixed-budget diffusion bridgeability beyond the dominant effect of layer depth. We then instantiate **DiHAL**, a Locate-and-Replace framework that replaces the transformer prefix below a selected interface with conditional diffusion while retaining the pretrained suffix and LM head. Under full training, geometry-selected interfaces remain close to validation-loss oracles and outperform worst-layer controls, while the diffusion bridge outperforms a parameter-matched deterministic replacement under matched diagnostic conditions. These results suggest that the representation space in which diffusion operates should itself be treated as a first-class design variable for continuous diffusion in language models.

1 INTRODUCTION

Diffusion models, which generate data through iterative denoising, have achieved remarkable success in continuous domains such as image generation (Song et al., 2021). This success has motivated growing interest in diffusion as an alternative to autoregressive language modeling, with both discrete and continuous formulations (Li et al., 2022; Strudel et al., 2023; Shi et al., 2024; Nie et al., 2025). Despite recent advances, continuous diffusion language models still lag behind autoregressive language models in generation quality (Jo & Hwang, 2025).

In this work, we focus on a challenge for continuous diffusion in language: the interface between continuous denoising and discrete token generation. Prior work has diffused token embeddings, self-conditioned embeddings, or learned text latents, together with mechanisms that map the resulting representations back to language (Li et al., 2022; Strudel et al., 2023; Lovelace et al., 2023; Gong et al., 2023; Zhang et al., 2025). When token recovery depends directly on the denoised representation, errors in representation space can alter the resulting discrete output (Li et al., 2022).

Yet this difficulty does not make continuous representations themselves unsuitable for language generation. Autoregressive language models already compute through a stack of continuous hidden states (Vaswani et al., 2017). An intermediate state is processed by the remaining pretrained transformer layers before the LM head produces token probabilities. Thus, from each intermediate state, the pretrained suffix already provides a learned continuation to the output distribution.

This suggests placing continuous diffusion inside the pretrained model rather than at the continuous-to-token interface. We target an internal hidden state and retain the pretrained computation above it: diffusion reconstructs a representation for the retained suffix, which then continues the mapping to token probabilities.

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.

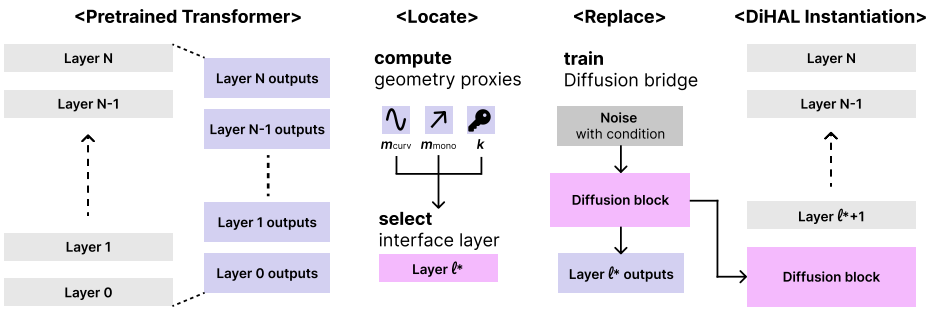

Figure 1: The Locate-and-Replace framework. DiHAL selects a hidden-state interface, replaces the preceding layers with a diffusion bridge, and retains the Transformer suffix and LM head.

This formulation raises a design question: where should diffusion enter the pretrained transformer? The insertion layer determines both the hidden-state distribution diffusion must model and the suffix that consumes it. Because these distributions vary across depth, their suitability for diffusion may also differ. We therefore view insertion as an interface-selection problem: identifying an internal representation that can be modeled reliably while remaining compatible with the retained suffix.

We hypothesize that the suitability is mirrored in hidden-state geometry. In particular, we want *diffusion-friendly* representations that contract quickly to the target distribution when denoising and are robust to imperfect score estimation, and whose variation is concentrated in a relatively small number of active directions. Idealized results from Langevin dynamics and concentration theory motivate geometric quantities associated with these properties (Villani, 2009; Bakry et al., 2014; Ledoux, 2001). We approximate them by three layer-wise proxies: local compactness, global stiffness and effective rank, which provide a training-free way to screen candidate insertion layers without training a diffusion model at each depth.

Based on this idea, we develop a *Locate-and-Replace* framework for identifying and testing diffusion-friendly hidden-state interfaces, instantiated as **DiHAL** (**Di**ffusion-Transformer **H**ybrid **A**rchitecture for **L**anguage Generation). DiHAL first *locates* a promising insertion layer using these geometric proxies, then *replaces* the transformer prefix below that layer with a diffusion model conditioned on the input token embeddings and trained to reconstruct the hidden states at that layer. The pretrained suffix and LM head remain unchanged, allowing us to test whether the selected interface remains compatible with downstream language generation. The resulting contributions are summarized below.

Interface	We formulate continuous diffusion as a hidden-state interface-selection problem.
Geometry	We introduce geometry-guided interface selection using three theory-motivated proxies—local compactness, global stiffness, and effective rank—validated on synthetic distributions and across backbones.
Evidence	We instantiate this principle in **DiHAL** and show that geometry-selected interfaces transfer to full training and compare favorably with alternative continuous-diffusion interfaces under shared training budgets.

2 Background

Transformer language models take tokens as input, but most computation occurs in continuous hidden spaces. Given $x_{1:T}$, an autoregressive model factorizes $p(x_{1:T}) = \prod_{t=1}^{T} p(x_t \mid x_{<t})$; each token is mapped to an embedding $h_t^{(0)} = E(x_t) + p_t$, hidden states are updated as $H^{(\ell)} = F_\ell(H^{(\ell-1)})$, and the final state is projected to vocabulary logits $z_t = W_{\mathrm{LM}} h_t^{(L)} + b$ (Vaswani et al., 2017). Diffusion models generate samples by reversing a noising process, written in continuous space as $dX_t = f(X_t, t)\,dt + g(t)\,dW_t$, where W_t is Brownian motion; the reverse process depends on the score $\nabla_x \log p_t(x)$, which is approximated by a neural network (Song et al., 2021). Applying diffusion to language thus requires choosing which representation to denoise. Discrete diffusion

models corrupt tokens (Austin et al., 2021; Hoogeboom et al., 2021), while continuous diffusion language models denoise token embeddings or learned latents (Li et al., 2022; Strudel et al., 2023; Lovelace et al., 2023). Token-level diffusion suffers from recovery errors, as small denoising deviations can flip the recovered token (Zhang et al., 2025; Shen et al., 2026), and learned latent diffusion still requires an interface for converting latents to text. We instead target internal transformer hidden states, where recovery becomes hidden-state reconstruction rather than direct token decoding.

3 METHOD

We study where continuous diffusion should operate within a pretrained language model by treating hidden-state selection as an interface-design problem. Figure 1 illustrates our Locate-and-Replace procedure. We first motivate diffusion-friendly representations using geometric principles, then derive layer-wise proxies for selecting a suitable hidden-state interface, and finally test the selected interface by replacing the lower transformer layers with a conditional diffusion bridge while retaining the upper layers and LM head. We call this instantiation `DiHAL`. The bridge reconstructs an internal representation that the retained upper layers can directly process.

3.1 GEOMETRIC PRINCIPLES FOR DIFFUSION-FRIENDLY LAYER SELECTION

We now make the notion of a *diffusion-friendly* representation space more concrete. Intuitively, a good diffusion space should satisfy three properties: denoising should contract quickly toward the target representation distribution, remain stable under score-estimation error, and have low effective complexity, with variation concentrated in relatively few active directions.

We formalize the first two properties through overdamped Langevin dynamics, an idealized setting with clean convergence and stability guarantees. The third is captured by effective rank, which measures active variance directions. The theorem settings in this section are idealized: they motivate geometric surrogates rather than assume that transformer hidden states exactly satisfy them.

Throughout this section, W_2 denotes the 2-Wasserstein distance and $\mathcal{P}_2(\mathbb{R}^d)$ denotes probability measures with finite second moment. For a density p, define $U(x) = -\log p(x)$. We interpret strong convexity of U as a curvature-like restoring force toward high-density regions.

Theorem 1 (Wasserstein contraction under strong log-concavity (Villani, 2009; Bakry et al., 2014))**.** *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ be a probability measure with density p, and define $U(x) := -\log p(x)$. Assume that $U \in C^2(\mathbb{R}^d)$, U is m-strongly convex, i.e. $\nabla^2 U(x) \succeq mI$ for all $x \in \mathbb{R}^d$, for some $m > 0$, and that ∇U is globally Lipschitz. Let $(X_t)_{t\geq 0}$ satisfy the overdamped Langevin stochastic differential equation (SDE) $dX_t = -\nabla U(X_t)\,dt + \sqrt{2}\,dW_t$, where W_t denotes Brownian motion.*

Then μ is an invariant distribution of (X_t), and for every initial law $\nu_0 \in \mathcal{P}_2(\mathbb{R}^d)$,

$$W_2(\nu_t, \mu) \leq e^{-mt} W_2(\nu_0, \mu), \qquad \nu_t := \mathcal{L}(X_t),$$

where $\mathcal{L}(X_t)$ denotes the distribution of X_t. The invariant distribution μ is unique in $\mathcal{P}_2(\mathbb{R}^d)$.

> **Takeaway (Theorem 1).** Stronger curvature yields faster contraction under idealized Langevin dynamics, motivating curvature as a proxy for diffusion-friendly hidden-state distributions.

Fast contraction alone is not enough. In practice, the score is unknown and must be estimated by a neural network. Theorem 2 characterizes stability under such estimation error.

Theorem 2 (Stability of an invariant measure under score perturbation)**.** *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ have density p and define $U(x) := -\log p(x)$. Assume that $U \in C^2(\mathbb{R}^d)$ is m-strongly convex, and that ∇U is globally Lipschitz. Let $s(x) := \nabla \log p(x) = -\nabla U(x)$, and let $\hat{s} : \mathbb{R}^d \to \mathbb{R}^d$ be globally Lipschitz and satisfy $\sup_{x\in\mathbb{R}^d} \|\hat{s}(x) - s(x)\| \leq \varepsilon$. Consider the two SDEs $dX_t = s(X_t)\,dt + \sqrt{2}\,dW_t$ and $d\hat{X}_t = \hat{s}(\hat{X}_t)\,dt + \sqrt{2}\,dW_t$. Assume that the second SDE admits an invariant distribution $\hat{\mu}$. Then $\hat{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ and*

$$W_2(\hat{\mu}, \mu) \leq \frac{\varepsilon}{m}.$$

Takeaway (Theorem 2). For a fixed score-estimation error, stronger curvature reduces the resulting distributional error, motivating curvature as a proxy for robustness to imperfect score estimation.

Curvature, however, does not capture the full complexity of a representation: variation may still be distributed across many directions. We therefore complement curvature with effective rank as a proxy for dimensionality. For covariance matrix Σ, define $r_{\text{eff}}(\Sigma) := \text{tr}(\Sigma)/\|\Sigma\|$, where $\text{tr}(\Sigma)$ is the total variance and $\|\Sigma\|$ is the largest covariance eigenvalue. Lower effective rank indicates that variation is concentrated along fewer active directions.

Lemma 1 (Approximate manifold support implies low effective rank)**.** *Let X be an $\mathbb{R}^d$-valued random variable with covariance $\Sigma := \text{Cov}(X)$. Assume there exist a k-dimensional C^2 manifold $\mathcal{M} \subset \mathbb{R}^d$ and a measurable map $\Pi : \mathbb{R}^d \to \mathcal{M}$ such that $\text{tr}(\text{Cov}(\Pi(X))) \le C_1 k$, $\mathbb{E}\|X - \Pi(X)\|^2 \le C_2(\delta^2 + \eta)$, and $\|\Sigma\| \ge c > 0$.*

Then

$$r_{\text{eff}}(\Sigma) := \frac{\text{tr}(\Sigma)}{\|\Sigma\|} \le \frac{2C_1}{c}\, k + \frac{2C_2}{c}\,(\delta^2 + \eta).$$

In particular, if δ, η are controlled constants and $\|\Sigma\|$ is bounded above and below by constants, then

$$r_{\text{eff}}(\Sigma) = O(k).$$

Takeaway (Lemma 1). Near a k-dimensional manifold, effective rank scales with intrinsic rather than ambient dimension, motivating it as a proxy for representation complexity.

We next combine this dimension control with curvature-based concentration under strong log-concavity (Bakry et al., 2014; Ledoux, 2001).

Theorem 3 (Intrinsic dimension and effective representation complexity)**.** *Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ have density $p(x) = Z^{-1}e^{-U(x)}$ for some $U \in C^2(\mathbb{R}^d)$, and let $\Sigma := \text{Cov}(\mu)$. Assume that $\nabla^2 U(x) \succeq mI$ for all $x \in \mathbb{R}^d$, for some $m > 0$, and that ∇U is globally Lipschitz.*

Assume moreover that there exist a k-dimensional C^2 manifold $\mathcal{M} \subset \mathbb{R}^d$ and a measurable map $\Pi : \mathbb{R}^d \to \mathcal{M}$. For $X \sim \mu$, suppose that $\text{tr}(\text{Cov}(\Pi(X))) \le C_1 k$, $\mathbb{E}\|X - \Pi(X)\|^2 \le C_2(\delta^2 + \eta)$, and $\|\Sigma\| \ge c_0 > 0$.

Then

$$r_{\text{eff}}(\Sigma) \le \frac{2C_1}{c_0}\, k + \frac{2C_2}{c_0}(\delta^2 + \eta),$$

and hence

$$\left(\mathbb{E}\|X - \mathbb{E}X\|^2\right)^{1/2} = \sqrt{\text{tr}(\Sigma)} \le \|\Sigma\|^{1/2} \left(\frac{2C_1}{c_0}\, k + \frac{2C_2}{c_0}(\delta^2 + \eta)\right)^{1/2}.$$

Furthermore, there exists an absolute constant $c > 0$ such that for all $t \ge 0$,

$$\mathbb{P}(|\|X - \mathbb{E}X\| - \mathbb{E}\|X - \mathbb{E}X\|| \ge t) \le 2\exp(-cmt^2).$$

Takeaway (Theorem 3). Low intrinsic dimension controls effective representation complexity through effective rank, while stronger curvature yields tighter concentration around the typical representation scale. Together, they motivate effective rank and curvature as complementary geometric criteria for diffusion-friendly hidden states.

Taken together, these results motivate three properties of diffusion-friendly representations: contraction, robustness to score-estimation error, and low effective complexity. While reverse diffusion uses time-dependent scores of noisy marginals rather than the fixed target score of Langevin dynamics, these principles provide geometric guidance for representation selection. We operationalize them using local covariance concentration, global precision-based stiffness, and effective rank, yielding testable geometric hypotheses for Transformer hidden states. We first validate the predicted relationships on controlled synthetic distributions, where each factor can be varied independently, and then examine whether the resulting proxies transfer to bridgeability across pretrained Transformer layers. Proofs of all results are provided in Appendix A.

3.2 LOCATE: FINDING DIFFUSION-FRIENDLY LAYERS

We translate the three properties above into empirical spectral proxies computed from hidden-state activations (see Appendix B for details). For each layer, let $x \in \mathbb{R}^{M \times D}$ denote the activation matrix over M tokens and D hidden dimensions.

We compute three statistics on x. First, the local curvature proxy $\hat{m}_{\mathrm{curv}}$ is obtained from the covariance of k-nearest-neighbor neighborhoods: $m^{(i)}_{\mathrm{curv}} = 1/\lambda_{\max}(\Sigma^{(i)}_{\mathrm{local}})$, with the layer-level value taken as the median; larger values indicate compact neighborhoods. Second, the monotonicity proxy $\hat{m}_{\mathrm{mono}}$ captures global directional stiffness. With $P = (\Sigma + \lambda I)^{-1}$ denoting the regularized precision of the empirical covariance Σ, we compute $m_{ij} = (x_i - x_j)^\top P (x_i - x_j)/\|x_i - x_j\|^2$ for sampled pairs and take the median. Third, effective intrinsic dimension is estimated as $\hat{k} = r_{\mathrm{eff}}(\Sigma) = \mathrm{tr}(\Sigma)/\|\Sigma\|$. Diffusion-friendly layers should have large curvature-related proxies and small effective rank.

We combine these into a selection score: $\mathrm{selection_score}(\ell) = z(\log \hat{m}_{\mathrm{curv}}(\ell)) + z(\log \hat{m}_{\mathrm{mono}}(\ell)) - z(\log \hat{k}(\ell))$, where $z(\cdot)$ denotes layer-wise z-score normalization. The score rewards curvature-related structure while penalizing effective rank. We define *bridgeability* as reconstructability of a layer's hidden state by the diffusion bridge under a matched training protocol, measured by validation loss. The layer sweep evaluates how well the geometry score predicts bridgeability under this fixed protocol, and we select $\ell^* = \arg\max_\ell \mathrm{selection_score}(\ell)$. We use this score as a low-cost operational criterion for interface selection. Details are in Appendix C.

> **Takeaway (Layer Selection).** DiHAL favors layers with large local curvature $\hat{m}_{\mathrm{curv}}$, large global stiffness $\hat{m}_{\mathrm{mono}}$, and low effective rank $\hat{k}$. Together, these proxies capture complementary aspects of diffusion-friendly representation geometry.

3.3 REPLACE: HIDDEN-STATE DIFFUSION MODULE

Given the selected insertion layer ℓ^*, we test whether the selected interface can be reconstructed by conditional diffusion while remaining compatible with the pretrained suffix. Let $F_{1:\ell^*}$ and $F_{\ell^*+1:L}$ denote the original prefix and retained suffix, respectively. For input x, the original model produces $h_{\ell^*} = F_{1:\ell^*}(x)$, while the diffusion bridge reconstructs $\hat{h}_{\ell^*} = D_\theta(c(x))$ from an embedding-derived condition $c(x)$. The reconstructed state is passed through the retained suffix, $h_L = F_{\ell^*+1:L}(\hat{h}_{\ell^*})$, and the LM head produces token probabilities.

At inference, D_θ bypasses the lower layers and directly reconstructs $\hat{h}_{\ell^*}$. We instantiate the bridge as a Stable-Diffusion-style UNet: hidden states are projected to a latent tensor, conditionally denoised, and projected back to the model hidden space. Mirroring the training setup, the condition at step t contains only the observed prefix $x_{\le t}$, and the retained causal suffix produces the next-token distribution with consistent masking. A small-scale architecture ablation is provided in Appendix D.4.

We optimize hidden-state reconstruction using the diffusion loss $\mathcal{L}_{\mathrm{diff}} = \mathbb{E}_{t,\epsilon}[\|\hat{\epsilon}_\theta(z_t, t, c) - \epsilon\|_2^2]$ and reconstruction loss $\mathcal{L}_{\mathrm{rec}} = \|\hat{h}_{\ell^*} - h_{\ell^*}\|_2^2$, together with next-token and logit-distillation losses $\mathcal{L}_{\mathrm{LM}}$ and $\mathcal{L}_{\mathrm{KD}}$ to preserve suffix compatibility. The overall objective is $\mathcal{L} = \mathcal{L}_{\mathrm{diff}} + \lambda_{\mathrm{rec}}\mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{LM}}\mathcal{L}_{\mathrm{LM}} + \lambda_{\mathrm{KD}}\mathcal{L}_{\mathrm{KD}}$. Implementation details are in Appendix D.1.

4 EXPERIMENTS

Our central experimental question is whether theory-motivated hidden-state geometry predicts fixed-budget diffusion bridgeability beyond layer depth, and whether this prediction transfers to fully trained models. We first validate the geometric factors on controlled synthetic distributions, then test the geometry score across layers of two pretrained 8B LMs with depth-controlled held-out analyses (Section 4.2). We further compare diffusion with deterministic and alternative generative bridges (Section 4.3), and fully train DiHAL at the geometry-selected layer against oracle and worst-layer controls while measuring fidelity and inference cost (Section 4.4). Finally, we compare our hidden-state interface with existing continuous diffusion language models (Section 4.5).

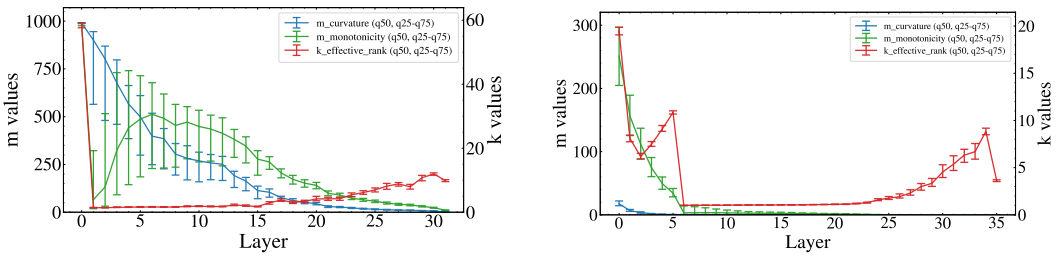

Figure 2: Layer-wise geometry for Llama-3.1-8B-Instruct (left) and Qwen3-8B (right). The three geometric factors vary across depth. The score selects layer 3 for Llama and layer 2 for Qwen.

4.1 Experimental Setup

Backbones and activation corpus. We evaluate DiHAL on Llama-3.1-8B-Instruct (Grattafiori et al., 2024) and Qwen3-8B (Yang et al., 2025), with 32 and 36 Transformer layers, respectively, and hidden size 4096. For each backbone, we run the source model on 300K *Dolma v1.7* sequences (Soldaini et al., 2024) and store the embedding output and layer-wise hidden states.

Geometry estimation. For each candidate layer, we compute local compactness $\hat{m}_{\mathrm{curv}}$, global stiffness $\hat{m}_{\mathrm{mono}}$, and effective rank $\hat{k} = r_{\mathrm{eff}}(\Sigma)$. We estimate them from 100 repeated 3K-example subsamples and assess score stability on 30 additional 500-example subsamples (Appendix C). The selection-score coefficients are fixed rather than fitted to bridge losses.

Bridge training. We train one embedding-conditioned bridge per candidate layer for one epoch on 150K examples with a 9:1 train/validation split, using the same bridge family, data, optimization, and budget across layers. Selected-layer bridges are trained on 300K examples for four epochs within 40 H100 GPU-hours. Unless noted, we use Stable-Diffusion-v1.5-style components (Rombach et al., 2022), with a frozen VAE and trainable U-Net and projections for hidden-state reconstruction; no image supervision or text-to-image conditioning is used (Appendix D.1).

Evaluation. We report Spearman and Kendall correlations between geometry score and negative bridge loss, along with predicted and observed best layers. We test beyond depth using LOOCV prediction of log bridge loss. For trained models, we report NLL, perplexity, and output-distribution KL on *WikiText-103* and held-out *Dolma v1.7*. For diffusion LM comparisons, we report GPT-2 Gen.PPL and diversity (Appendix G.2). Details are in Appendix E.

4.2 Geometry Predicts Bridgeability

Controlled synthetic validation. Section 3.1 motivates geometric factors for diffusion-friendly representations without assuming that Transformer hidden states satisfy the idealized theory. We validate these factors on synthetic target distributions using the same fixed-budget Stable-Diffusion-style U-Net denoiser while varying one factor at a time. Denoising difficulty follows the predicted direction, with Spearman correlations of 0.947, 0.946, and 0.968 for global stiffness, effective rank, and local compactness, respectively, although the relationship weakens under strong multimodality. We next test whether these proxies transfer to pretrained Transformer hidden spaces.

Layer-wise geometry. Figure 2 shows substantial geometric variation across depth for both backbones. Input-adjacent layers generally exhibit stronger local compactness, while global stiffness and effective rank follow different trends. The score selects layer 3 for Llama-3.1-8B and layer 2 for Qwen3-8B, while the lowest fixed-budget bridge loss occurs at layer 1 for both. Thus, bridgeability is shallow, and we test whether geometry explains variation beyond this depth trend.

Fixed-budget layer sweep. Figure 3 compares the geometry score with validation bridge loss across all candidate layers. Higher-scoring layers generally exhibit lower bridge loss. As shown in Table 1, the score achieves Spearman $\rho = 0.9143 \pm 0.0069$ for Llama and $\rho = 0.9267 \pm 0.0157$

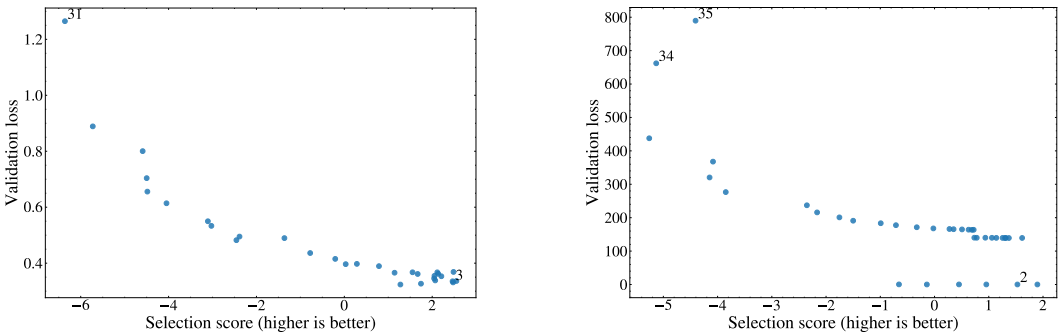

Figure 3: Geometry score versus bridge loss for Llama-3.1-8B-Instruct (left) and Qwen3-8B (right). Each point denotes a candidate layer. Higher scores generally correspond to lower fixed-budget loss.

Table 1: Agreement between geometry score and fixed-budget bridgeability. Correlations use negative validation bridge loss; $\pm$ denotes standard deviation over 30 repeated score-estimation runs.

Model	Spearman $\rho \uparrow$	Kendall $\tau \uparrow$	Best pred.	Best obs.
Llama-3.1-8B	0.9143 ± 0.0069	0.8011 ± 0.0086	3	1
Qwen3-8B	0.9267 ± 0.0157	0.8208 ± 0.0256	2	1

for Qwen across repeated score estimates, with rank gaps of 2 and 1, respectively, between the geometry-selected and observed best layers.

Does geometry add information beyond depth? The preference for shallow layers raises an alternative explanation: the score may simply reflect that representations closer to the embedding condition are easier to reconstruct. We therefore compare a quadratic depth model, $y_\ell = \beta_0 + \beta_1 d_\ell + \beta_2 d_\ell^2 + \epsilon_\ell$, with the same model augmented by geometry, $y_\ell = \beta_0 + \beta_1 d_\ell + \beta_2 d_\ell^2 + \beta_G G_\ell + \epsilon_\ell$, where $y_\ell = \log \mathcal{L}_{\text{bridge}}(\ell)$, using LOOCV.

Adding geometry reduces LOOCV RMSE by 28.4% for Llama and 4.9% for Qwen, while improving MAE and out-of-sample R^2 on both, so we treat geometry as complementary to depth rather than replacing it. Against CKA, hidden-state distance, and linear reconstruction error, geometry is strongest on Llama, while CKA is slightly better in RMSE and R^2 on Qwen (Appendix F.1); thus, geometry adds information beyond depth and recoverability without dominating every heuristic.

Stability beyond one epoch. The layer sweep is a low-budget diagnostic, not a convergence study. We continue training representative shallow, middle and deep layers using the same protocol to test whether its ranking is a consequence of early training. The rank of the validation-loss normalized by evaluated layers shows a mean absolute change of less than 10% for both backbones with the geometry-selected layer being favored among candidates. Thus, the fixed-budget ranking reflects stable bridgeability and not only the speed of initial optimization.

> **Takeaway.** The geometric proxies predict diffusion difficulty on synthetic data and transfer to transformer layers, providing signal beyond depth for locating diffusion-friendly interfaces.

4.3 Does Diffusion Help Beyond Generic Hidden-State Replacement?

The preceding experiments establish where the bridge can reconstruct hidden states. We next ask whether diffusion provides an advantage over generic hidden-state replacement. We compare bridge families at the geometry-selected Llama layer 3 with sequence length $S = 1024$. All bridges use the same 512 activation samples, 300 optimization steps, target layer, frozen suffix, LM head, and evaluation protocol. This reduced-budget experiment compares reconstruction and suffix compatibility across bridge families under matched conditions.

Table 2: Held-out prediction of log bridge loss after controlling for normalized layer depth. Geometry complements, rather than replaces, the broad depth trend.

Backbone	Predictor	RMSE ↓	MAE ↓	R^2_{OOS} ↑
Llama-3.1-8B	Quadratic depth	0.7537	0.3769	0.0915
	Depth + geometry	**0.5400**	**0.1792**	**0.5336**
Qwen3-8B	Quadratic depth	1.6102	1.3248	0.6983
	Depth + geometry	**1.5316**	**1.2550**	**0.7271**

Table 3: Bridge-family comparison at the geometry-selected Llama-3.1-8B layer 3 under a shared reduced-budget protocol (512 activation samples, 300 optimization steps).

Bridge	Objective	Params	Hidden MSE ↓	Teacher KL ↓	NLL ↓	PPL ↓
DiHAL SD bridge	Latent diffusion	946.3M	**0.0067**	**5.7446**	**7.6407**	**2,081.3**
Deterministic MLP	Direct regression	939.7M	0.5665	9.0225	10.6760	43,302.7
Conditional VAE	Variational modeling	53.5M	0.4308	7.5690	9.6962	16,255.4
Flow matching	Flow matching	285.2M	0.9399	10.3059	12.3001	219,713.1
Hidden DDPM	Hidden-space diffusion	272.6M	1.5416	11.2132	13.1550	516,594.6

Table 3 shows that the latent-diffusion bridge achieves the best reconstruction and suffix-level fidelity among the tested alternatives. Relative to the parameter-matched deterministic MLP, the latent-diffusion bridge reduces hidden-state MSE by 98.8%, teacher KL by 36.3%, and NLL by 28.4%, indicating that capacity or generic hidden-state regression alone does not explain the result. Under the shared protocol, these results show that latent diffusion produces substantially more suffix-compatible hidden states.

> **Takeaway.** Among the tested bridge families, latent diffusion provides the strongest reconstruction and suffix compatibility under matched conditions.

4.4 Does Geometry-Guided Selection Transfer to Full Training?

We fully train the bridge at the geometry-selected layer on the 300K-example corpus for four epochs within 40 H100 GPU-hours. We compare against a validation-loss oracle from the fixed-budget sweep and a worst-layer control. Unlike the oracle, which requires training every candidate bridge, the geometry score uses frozen activations alone.

Table 4 shows that the geometry-selected layer outperforms the worst-layer control and remains close to the oracle, outperforming it on Llama and trailing slightly on Qwen. Compared with the source models, NLL is 4.91 versus 1.87 for Llama and 4.97 versus 2.61 for Qwen, corresponding to perplexity ratios of $20.96\times$ and $10.63\times$. Fidelity varies gradually with reconstruction error (Appendix I), while four denoising evaluations yield 28.4% of the original throughput.

> **Takeaway.** Geometry-guided selection transfers to full training, remaining near the validation-loss oracle and outperforming the worst-layer control.

4.5 Does the Choice of Diffusion Interface Matter in Practice?

We compare DiHAL with continuous diffusion language models targeting token embeddings (Diffusion-LM (Li et al., 2022); SED (Strudel et al., 2023)), learned latents (LD4LG (Lovelace et al., 2023)), and continuous-to-discrete recovery (CoDAR (Shen et al., 2026)) in the Qwen3-8B setting. All methods use 300K examples and 40 H100-hours for diffusion/recovery training. This comparison examines the practical effect of the diffusion interface under a shared training budget, while accounting for differences in retained pretrained capacity and recovery mechanisms; parameter counts are provided in Appendix G.

Table 4: Fully trained DiHAL models, averaged over WikiText-103 and held-out Dolma ($N_{\mathrm{FE}} = 20$). The original model is a reference; the best DiHAL result per backbone is in bold.

Backbone	Model	Layer	NLL ↓	PPL ↓	KL ↓
Llama-3.1-8B	Original (reference)	–	1.87	6.47	0.00
	Geometry-selected	3	**4.91**	**135.64**	0.73
	Validation-loss oracle	1	5.11	165.67	**0.62**
	Worst layer	31	5.17	175.91	1.32
Qwen3-8B	Original (reference)	–	2.61	13.55	0.00
	Geometry-selected	2	4.97	144.03	**0.53**
	Validation-loss oracle	1	**4.94**	**139.77**	0.54
	Worst layer	35	5.23	186.79	1.46

Table 5: Comparison of continuous diffusion interfaces under a matched diffusion/recovery training budget (300K examples, 40 H100-hours).

Method	Diffusion target	Recovery interface	Gen.PPL ↓	Div. ↑
DiHAL	Hidden state h_{ℓ^*}	Retained suffix + LM head	**136.02**	**0.5913**
Diffusion-LM	Token embeddings	Embedding-to-token	683.43	0.4324
SED	Self-conditioned embeddings	Embedding-to-token	778.82	0.4942
LD4LG	Learned text latent	Frozen BART + latent decoder	166.11	0.5797
CoDAR	Continuous token/latent states	Continuous-to-discrete	144.83	0.4777

Table 5 shows that DiHAL achieves the lowest Gen.PPL and highest diversity among the compared interfaces. A 341.1M SD-style bridge trained from scratch also reaches Gen.PPL 140.51 and diversity 0.5230, versus 144.83 and 0.4777 for CoDAR, suggesting that neither bridge size nor pretrained U-Net initialization alone explains the difference (Appendix G.3).

5 Related Work

Diffusion language models adapt diffusion to text, where discreteness makes generation less direct than in images (Hoogeboom et al., 2022). Prior work includes discrete token diffusion (Austin et al., 2021; Gong et al., 2023), masked diffusion with iterative refinement (Sahoo et al., 2024; Shi et al., 2024; Nie et al., 2025; Ye et al., 2025), and continuous diffusion over embeddings or learned latents (Li et al., 2022; Lovelace et al., 2023). Continuous methods avoid token corruption but require recovering tokens from denoised vectors, introducing projection or decoding errors (Wang et al., 2022; Li et al., 2022). We study transformer hidden states as a diffusion-friendly denoising space.

Diffusion behavior depends on representation geometry, including curvature, conditioning, and intrinsic dimension (Pidstrigach, 2022; Rombach et al., 2022). In parallel, efficient generation has been studied through transformer compression, distillation, layer reduction, early exiting, and hybrid modules (Fan et al., 2020; Sanh et al., 2020; Lenz et al., 2025). DiHAL connects these directions by identifying internal transformer representations suitable for diffusion-based replacement.

6 Conclusion

We studied continuous diffusion in language models as a hidden-state interface-selection problem, asking which internal representations are most compatible with diffusion rather than treating the denoising space as fixed. Across synthetic distributions and Transformer hidden states, theory-motivated geometric proxies strongly predict diffusion bridgeability, add predictive signal beyond layer depth, and guide interface selection across two 8B-scale backbones. Geometry-selected interfaces remain favorable after full training, in the other hand hidden-state diffusion compares favorably with alternative continuous-diffusion interfaces under a shared training budget. We instantiate this principle in `DiHAL`, providing a concrete realization of a broader design direction: the representation space in which diffusion operates should itself be treated as a first-class design variable for future diffusion language models.

AI USE STATEMENT

In this work, we used generative AI tools primarily for review, feedback, and implementation assistance. The authors independently developed the conceptual framework, theoretical models, hypotheses, mathematical claims, proofs, research methodology, and experimental design. Generative AI tools were used to provide feedback on these components, improve their clarity and presentation, and identify possible issues.

We also used generative AI tools to assist with method implementation and software development, synthetic benchmark data generation under author-defined specifications, translation, data cleaning and reformatting, interpretation of experimental results, literature search and summarization, reference identification, manuscript revision, and preparation of scientific figures and research artifacts. Qualitative or thematic data analysis was not part of this work.

All AI-assisted outputs were reviewed and verified by the authors. The authors made all final decisions regarding the research content and take full responsibility for the final submission.

REFERENCES

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=h7-XixPCAL.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 348 of *Grundlehren der mathematischen Wissenschaften*. Springer Cham, 2014. ISBN 978-3-319-00227-9. doi: 10.1007/978-3-319-00227-9.

Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SylO2yStDr.

Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and Lingpeng Kong. Diffuseq-v2: Bridging discrete and continuous text spaces for accelerated seq2seq diffusion models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=OLcDbSRjbx.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne,

Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman,

Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=6nbpPqUCIi7.

Emiel Hoogeboom, Alexey A. Gritsenko, Jasmijn Bastings, Ben Poole, Rianne van den Berg, and Tim Salimans. Autoregressive diffusion models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Lm8T39vLDTE.

Jaehyeong Jo and Sung Ju Hwang. Continuous diffusion model for language modeling. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=VGv5y60sXC.

M. Ledoux. *The Concentration of Measure Phenomenon*. Mathematical surveys and monographs. American Mathematical Society, 2001. ISBN 9780821837924. URL https://books.google.com.sg/books?id=mCX_cWL6rqwC.

Barak Lenz, Opher Lieber, Alan Arazi, Amir Bergman, Avshalom Manevich, Barak Peleg, Ben Aviram, Chen Almagor, Clara Fridman, Dan Padnos, Daniel Gissin, Daniel Jannai, Dor Muhlgay, Dor Zimberg, Edden M. Gerber, Elad Dolev, Eran Krakovsky, Erez Safahi, Erez Schwartz, Gal Cohen, Gal Shachaf, Haim Rozenblum, Hofit Bata, Ido Blass, Inbal Magar, Itay Dalmedigos, Jhonathan Osin, Julie Fadlon, Maria Rozman, Matan Danos, Michael Gokhman, Mor Zusman, Naama Gidron, Nir Ratner, Noam Gat, Noam Rozen, Oded Fried, Ohad Leshno, Omer Antverg, Omri Abend, Or Dagan, Orit Cohavi, Raz Alon, Ro'i Belson, Roi Cohen, Rom Gilad, Roman Glozman, Shahar Lev, Shai Shalev-Shwartz, Shaked Haim Meirom, Tal Delbari, Tal Ness, Tomer Asida, Tom Ben Gal, Tom Braude, Uriya Pumerantz, Josh Cohen, Yonatan Belinkov, Yuval Globerson, Yuval Peleg Levy, and Yoav Shoham. Jamba: Hybrid transformer-mamba language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=JFPaD7lpBD.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. Diffusion-LM improves controllable text generation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=3s9IrEsjLyk.

Justin Lovelace, Varsha Kishore, Chao Wan, Eliot Seo Shekhtman, and Kilian Q Weinberger. Latent diffusion for language generation. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=NKdtztladR.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, JUN ZHOU, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=KnqiC0znVF.

Jakiw Pidstrigach. Score-based generative models detect manifolds, 2022. URL https://proceedings.nips.cc/paper_files/paper/2022/file/e8fb575e3ede31f9b8c05d53514eb7c6-Paper-Conference.pdf.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Subham Sekhar Sahoo, Marianne Arriola, Aaron Gokaslan, Edgar Mariano Marroquin, Alexander M Rush, Yair Schiff, Justin T Chiu, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=L4uaAR4ArM.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter, 2020. URL https://www.emc2-ai.org/assets/docs/neurips-19/emc2-neurips19-paper-33.pdf.

Junzhe Shen, Jieru Zhao, Ziwei He, and Zhouhan Lin. Codar: Continuous diffusion language models are more powerful than you think, 2026. URL https://arxiv.org/abs/2603.02547.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. Simplified and generalized masked diffusion for discrete data. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=xcqSOfHt4g.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research, 2024. URL https://arxiv.org/abs/2402.00159.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=PxTIG12RRHS.

Robin Strudel, Corentin Tallec, Florent Altché, Yilun Du, Yaroslav Ganin, Arthur Mensch, Will Sussman Grathwohl, Nikolay Savinov, Sander Dieleman, Laurent Sifre, and Rémi Leblond. Self-conditioned embedding diffusion for text generation, 2023. URL https://openreview.net/forum?id=OpzV3lp3IMC.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer Berlin, Heidelberg, 2009. ISBN 978-3-540-71050-9. doi: 10.1007/978-3-540-71050-9.

Rose E Wang, Esin Durmus, Noah Goodman, and Tatsunori Hashimoto. Language modeling via stochastic processes. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=pMQwKL1yctf.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025.

Siyue Zhang, Yilun Zhao, Liyuan Geng, Arman Cohan, Anh Tuan Luu, and Chen Zhao. Diffusion vs. autoregressive language models: A text embedding perspective. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4273–4303, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.213. URL https://aclanthology.org/2025.emnlp-main.213/.

A Proofs of Theorems

Preliminaries

Law of a random variable. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}^d$ be a random variable. The law (distribution) of X, denoted by $\mathcal{L}(X)$, is the probability measure defined by

$$\mathcal{L}(X)(A) := \mathbb{P}(X \in A), \quad A \in \mathcal{B}(\mathbb{R}^d).$$

2-Wasserstein distance. Let $\mathcal{P}_2(\mathbb{R}^d)$ denote the set of probability measures with finite second moment. For $\nu, \mu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$W_2^2(\nu, \mu) := \inf_{\pi \in \Pi(\nu,\mu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} \|x - y\|^2 \, \pi(dx, dy),$$

where $\Pi(\nu, \mu)$ denotes the set of couplings of ν and μ. Equivalently,

$$W_2^2(\nu, \mu) = \inf \left\{ \mathbb{E}\|X - Y\|^2 : \mathcal{L}(X) = \nu, \mathcal{L}(Y) = \mu \right\}.$$

In $\mathbb{R}^d$, an optimal coupling attaining the infimum exists. If preferred, one may instead work with ε-optimal couplings and let $\varepsilon \downarrow 0$ at the end.

Lemma 2 (Strong convexity implies gradient monotonicity)**.** *Let $U : \mathbb{R}^d \to \mathbb{R}$ be differentiable and m-strongly convex, i.e.*

$$U(y) \geq U(x) + \langle \nabla U(x), y - x \rangle + \frac{m}{2}\|y - x\|^2 \quad \textit{for all } x, y.$$

Then

$$\langle \nabla U(x) - \nabla U(y), x - y \rangle \geq m\|x - y\|^2 \quad \textit{for all } x, y.$$

Proof. Apply strong convexity twice, swapping x and y:

$$U(y) \geq U(x) + \langle \nabla U(x), y - x \rangle + \frac{m}{2}\|y - x\|^2,$$

$$U(x) \geq U(y) + \langle \nabla U(y), x - y \rangle + \frac{m}{2}\|x - y\|^2.$$

Summing the two inequalities yields

$$\langle \nabla U(x) - \nabla U(y), x - y \rangle \geq m\|x - y\|^2.$$

□

Lemma 3 (Gibbs invariance and uniqueness)**.** *Assume that $U \in C^2(\mathbb{R}^d)$, ∇U is globally Lipschitz, and*

$$\nabla^2 U(x) \succeq mI \qquad \textit{for all } x \in \mathbb{R}^d$$

for some $m > 0$. Let

$$\mu(dx) = Z^{-1} e^{-U(x)} \, dx, \qquad Z := \int_{\mathbb{R}^d} e^{-U(x)} \, dx.$$

Then $Z < \infty$, $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, and μ is invariant for

$$dX_t = -\nabla U(X_t)\, dt + \sqrt{2}\, dW_t.$$

Moreover, if for every initial law $\nu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ the law $\nu_t := \mathcal{L}(X_t)$ satisfies

$$W_2(\nu_t, \mu) \leq e^{-mt} W_2(\nu_0, \mu),$$

then μ is the unique invariant distribution in $\mathcal{P}_2(\mathbb{R}^d)$.

Proof. By strong convexity, for every $x \in \mathbb{R}^d$,

$$U(x) \geq U(0) + \langle \nabla U(0), x \rangle + \frac{m}{2}\|x\|^2.$$

Completing the square, this implies

$$e^{-U(x)} \leq C \exp\Big(-\frac{m}{4}\|x\|^2\Big)$$

for some constant $C < \infty$. Hence $Z < \infty$ and μ has finite second moment.

Let

$$L\varphi(x) := \Delta\varphi(x) - \langle \nabla U(x), \nabla\varphi(x) \rangle$$

be the generator. For $\varphi \in C_c^\infty(\mathbb{R}^d)$, integration by parts gives

$$\int_{\mathbb{R}^d} L\varphi(x)\, \mu(dx) = Z^{-1} \int_{\mathbb{R}^d} \big(\Delta\varphi(x) - \langle \nabla U(x), \nabla\varphi(x) \rangle\big) e^{-U(x)}\, dx = 0.$$

Since this holds for all $\varphi \in C_c^\infty(\mathbb{R}^d)$, the stationary equation $L^*\mu = 0$ holds in the weak sense. Therefore μ is invariant.

Finally, if $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ is any invariant distribution, then applying the contraction estimate with $\nu_0 = \nu$ yields

$$W_2(\nu, \mu) = W_2(\nu_t, \mu) \leq e^{-mt} W_2(\nu, \mu) \qquad \text{for all } t \geq 0.$$

Fix any $t > 0$. Since $e^{-mt} < 1$, this implies $W_2(\nu, \mu) = 0$, hence $\nu = \mu$. □

PROOF OF THEOREM 1

By Lemma 3, μ is an invariant distribution for

$$dX_t = -\nabla U(X_t)\, dt + \sqrt{2}\, dW_t.$$

Let $\nu_0 := \mathcal{L}(X_0)$. Choose an optimal coupling (X_0, Y_0) between ν_0 and μ such that

$$\mathbb{E}\|X_0 - Y_0\|^2 = W_2^2(\nu_0, \mu).$$

(Alternatively, choose an ε-optimal coupling and let $\varepsilon \downarrow 0$.) We take this initial coupling to be independent of the Brownian motion below.

Define the synchronous coupling:

$$dX_t = -\nabla U(X_t)\, dt + \sqrt{2}\, dW_t,$$

$$dY_t = -\nabla U(Y_t)\, dt + \sqrt{2}\, dW_t,$$

driven by the same Brownian motion W_t.

Since $Y_0 \sim \mu$ and μ is invariant, we have

$$\mathcal{L}(Y_t) = \mu \qquad \text{for all } t \geq 0.$$

Let $Z_t := X_t - Y_t$. Subtracting the SDEs gives

$$dZ_t = -(\nabla U(X_t) - \nabla U(Y_t))\, dt.$$

Since the stochastic terms cancel, Z_t has absolutely continuous paths and satisfies the pathwise ODE

$$\dot{Z}_t = -(\nabla U(X_t) - \nabla U(Y_t)) \qquad \text{for a.e. } t \geq 0.$$

Therefore the ordinary chain rule applies, and

$$\frac{d}{dt}\|Z_t\|^2 = 2\langle Z_t, \dot{Z}_t \rangle = -2\langle Z_t, \nabla U(X_t) - \nabla U(Y_t) \rangle.$$

By Lemma 2,

$$\frac{d}{dt}\|Z_t\|^2 \leq -2m\|Z_t\|^2.$$

By Grönwall's inequality,

$$\|Z_t\|^2 \le e^{-2mt}\|Z_0\|^2.$$

Taking expectation yields

$$\mathbb{E}\|X_t - Y_t\|^2 \le e^{-2mt}\mathbb{E}\|X_0 - Y_0\|^2.$$

Since (X_t, Y_t) is a coupling of $\mathcal{L}(X_t)$ and μ, by definition of W_2,

$$W_2^2(\mathcal{L}(X_t), \mu) \le \mathbb{E}\|X_t - Y_t\|^2.$$

Therefore,

$$W_2^2(\mathcal{L}(X_t), \mu) \le e^{-2mt}W_2^2(\nu_0, \mu).$$

Taking square roots gives

$$W_2(\mathcal{L}(X_t), \mu) \le e^{-mt}W_2(\nu_0, \mu).$$

It remains to prove uniqueness of the invariant distribution in $P_2(\mathbb{R}^d)$. Let $\pi \in P_2(\mathbb{R}^d)$ be any invariant distribution of the same Langevin SDE. Applying the contraction estimate above with $\nu_0 = \pi$ gives, for every $t > 0$,

$$W_2(\pi, \mu) = W_2(\mathcal{L}(X_t), \mu) \le e^{-mt}W_2(\pi, \mu),$$

where we used the invariance of π to obtain $\mathcal{L}(X_t) = \pi$. Since $m > 0$ and $e^{-mt} < 1$ for $t > 0$, this implies

$$W_2(\pi, \mu) = 0.$$

Hence $\pi = \mu$. Therefore μ is the unique invariant distribution in $P_2(\mathbb{R}^d)$. □

PROOF OF THEOREM 2

We prove stability of an invariant measure under a uniformly bounded perturbation of the score function.

Let $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ have density p and define

$$U(x) := -\log p(x).$$

Assume that $U \in C^2(\mathbb{R}^d)$ is m-strongly convex, i.e.

$$\nabla^2 U(x) \succeq mI \quad \text{for all } x \in \mathbb{R}^d,$$

for some $m > 0$, and that ∇U is globally Lipschitz.

Define the true score

$$s(x) := \nabla \log p(x) = -\nabla U(x).$$

Let $\hat{s}$ be a globally Lipschitz function satisfying

$$\sup_{x \in \mathbb{R}^d} \|\hat{s}(x) - s(x)\| \le \varepsilon.$$

Consider the SDEs

$$dX_t = s(X_t)\,dt + \sqrt{2}\,dW_t, \qquad d\hat{X}_t = \hat{s}(\hat{X}_t)\,dt + \sqrt{2}\,dW_t,$$

and assume that the second SDE admits an invariant distribution $\hat{\mu}$. By Lemma 3, μ is an invariant distribution of the first SDE.

We first show that $\hat{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$. Since $s = -\nabla U$ and U is m-strongly convex, Lemma 2 applied with $y = 0$ gives

$$\langle x, s(x) - s(0)\rangle \le -m\|x\|^2 \qquad \text{for all } x \in \mathbb{R}^d.$$

Hence

$$\langle x, s(x)\rangle = \langle x, s(x) - s(0)\rangle + \langle x, s(0)\rangle \le -m\|x\|^2 + \|s(0)\|\,\|x\|.$$

Using $\|\hat{s}(x) - s(x)\| \le \varepsilon$, we obtain

$$\langle x, \hat{s}(x)\rangle \le \langle x, s(x)\rangle + \varepsilon\|x\| \le -m\|x\|^2 + (\|s(0)\| + \varepsilon)\|x\|.$$

By Young's inequality, there exists a constant $C_0 < \infty$ such that

$$\langle x, \hat{s}(x)\rangle \le -\frac{m}{2}\|x\|^2 + C_0 \qquad \text{for all } x \in \mathbb{R}^d.$$

Let

$$\hat{L}f(x) := \langle \hat{s}(x), \nabla f(x)\rangle + \Delta f(x)$$

be the generator of the perturbed SDE, and set

$$V(x) := \|x\|^2.$$

Then

$$\hat{L}V(x) = 2\langle x, \hat{s}(x)\rangle + 2d \le -m\|x\|^2 + C_1 = -mV(x) + C_1$$

for some constant $C_1 < \infty$.

We now justify the use of this unbounded Lyapunov function by truncation. Let $\psi_R \in C^2([0,\infty))$ be nondecreasing and concave, with

$$0 \le \psi_R' \le 1, \qquad \psi_R'' \le 0,$$

such that

$$\psi_R(r) = r \quad \text{for } r \le R, \qquad \psi_R(r) = \text{constant} \quad \text{for } r \ge 2R.$$

Define

$$V_R(x) := \psi_R(V(x)).$$

Since $V_R \in C_b^2(\mathbb{R}^d)$ and the perturbed SDE has globally Lipschitz drift, the associated Markov semigroup $(\widehat{P}_t)_{t\ge 0}$ has generator $\hat{L}$. Hence

$$\frac{d}{dt}\widehat{P}_t V_R\Big|_{t=0} = \hat{L}V_R.$$

By invariance of $\hat{\mu}$,

$$\int_{\mathbb{R}^d} \widehat{P}_t V_R(x)\,\hat{\mu}(dx) = \int_{\mathbb{R}^d} V_R(x)\,\hat{\mu}(dx) \quad \text{for all } t \ge 0.$$

Differentiating at $t = 0$ gives

$$\int_{\mathbb{R}^d} \hat{L}V_R(x)\,\hat{\mu}(dx) = 0.$$

By the chain rule for $\hat{L}$,

$$\hat{L}V_R = \psi_R'(V)\hat{L}V + \psi_R''(V)\|\nabla V\|^2.$$

Since $\psi_R'' \le 0$, we have

$$\hat{L}V_R \le \psi_R'(V)\hat{L}V \le \psi_R'(V)(-mV + C_1).$$

Therefore

$$0 = \int \hat{L}V_R\, d\hat{\mu} \le -m\int \psi_R'(V)V\, d\hat{\mu} + C_1\int \psi_R'(V)\, d\hat{\mu}.$$

Since $0 \le \psi_R' \le 1$, this implies

$$m\int \psi_R'(V)V\, d\hat{\mu} \le C_1.$$

Moreover, $\psi_R'(V) = 1$ whenever $V \le R$, and hence

$$m\int_{\{V\le R\}} V\, d\hat{\mu} \le C_1.$$

Letting $R \to \infty$ and using monotone convergence gives

$$\int \|x\|^2\,\hat{\mu}(dx) = \int V\, d\hat{\mu} \le \frac{C_1}{m} < \infty.$$

Thus $\hat{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$.

Now choose an optimal coupling $(X_0, \hat{X}_0)$ of μ and $\hat{\mu}$ such that

$$\mathbb{E}\|X_0 - \hat{X}_0\|^2 = W_2^2(\mu, \hat{\mu}).$$

We take this initial coupling to be independent of the Brownian motion below. Drive both SDEs by the same Brownian motion $(W_t)_{t\geq 0}$ and define

$$Z_t := \hat{X}_t - X_t.$$

Subtracting the two SDEs yields

$$dZ_t = \big(\hat{s}(\hat{X}_t) - s(X_t)\big)\, dt.$$

Since the stochastic terms cancel, Z_t has absolutely continuous paths and satisfies the pathwise ODE

$$\dot{Z}_t = \hat{s}(\hat{X}_t) - s(X_t) \qquad \text{for a.e. } t \geq 0.$$

Therefore the ordinary chain rule applies, and

$$\frac{d}{dt}\|Z_t\|^2 = 2\langle Z_t, \hat{s}(\hat{X}_t) - s(X_t)\rangle.$$

We decompose

$$\hat{s}(\hat{X}_t) - s(X_t) = \big(\hat{s}(\hat{X}_t) - s(\hat{X}_t)\big) + \big(s(\hat{X}_t) - s(X_t)\big).$$

Hence

$$\frac{d}{dt}\|Z_t\|^2 = 2\langle Z_t, \hat{s}(\hat{X}_t) - s(\hat{X}_t)\rangle + 2\langle Z_t, s(\hat{X}_t) - s(X_t)\rangle.$$

For the first term, using the uniform bound on the score error,

$$2\langle Z_t, \hat{s}(\hat{X}_t) - s(\hat{X}_t)\rangle \leq 2\varepsilon\|Z_t\|.$$

For the second term, Lemma 2 implies

$$\langle \hat{X}_t - X_t, \nabla U(\hat{X}_t) - \nabla U(X_t)\rangle \geq m\|\hat{X}_t - X_t\|^2.$$

Since $s = -\nabla U$, we get

$$\langle Z_t, s(\hat{X}_t) - s(X_t)\rangle \leq -m\|Z_t\|^2.$$

Combining the two bounds gives

$$\frac{d}{dt}\|Z_t\|^2 \leq -2m\|Z_t\|^2 + 2\varepsilon\|Z_t\|.$$

By Young's inequality,

$$2\varepsilon\|Z_t\| \leq m\|Z_t\|^2 + \frac{\varepsilon^2}{m}.$$

Substituting this into the differential inequality, we obtain

$$\frac{d}{dt}\|Z_t\|^2 \leq -m\|Z_t\|^2 + \frac{\varepsilon^2}{m}.$$

Multiplying both sides by e^{mt} and integrating from 0 to t yields

$$\|Z_t\|^2 \leq e^{-mt}\|Z_0\|^2 + \frac{\varepsilon^2}{m^2}\big(1 - e^{-mt}\big) \qquad \text{a.s.}$$

Taking expectations gives

$$\mathbb{E}\|Z_t\|^2 \leq e^{-mt}\mathbb{E}\|Z_0\|^2 + \frac{\varepsilon^2}{m^2}\big(1 - e^{-mt}\big).$$

Since $X_0 \sim \mu$ and $\hat{X}_0 \sim \hat{\mu}$ are invariant initial laws, we have

$$X_t \sim \mu, \qquad \hat{X}_t \sim \hat{\mu} \qquad \text{for all } t \geq 0.$$

Therefore $(X_t, \hat{X}_t)$ is a coupling of μ and $\hat{\mu}$, so

$$W_2^2(\hat{\mu}, \mu) \leq \mathbb{E}\|\hat{X}_t - X_t\|^2 = \mathbb{E}\|Z_t\|^2.$$

Combining this with the previous estimate gives

$$W_2^2(\hat{\mu}, \mu) \leq e^{-mt} W_2^2(\hat{\mu}, \mu) + \frac{\varepsilon^2}{m^2}\big(1 - e^{-mt}\big).$$

Letting $t \to \infty$, we obtain

$$W_2^2(\hat{\mu}, \mu) \leq \frac{\varepsilon^2}{m^2}.$$

Taking square roots gives

$$W_2(\hat{\mu}, \mu) \leq \frac{\varepsilon}{m}.$$

□

PROOF OF LEMMA 1

Let

$$\Sigma := \mathrm{Cov}(X).$$

Define

$$R := X - \Pi(X).$$

Since

$$\mathbb{E}X = \mathbb{E}\Pi(X) + \mathbb{E}R,$$

we may write

$$X - \mathbb{E}X = \big(\Pi(X) - \mathbb{E}\Pi(X)\big) + \big(R - \mathbb{E}R\big).$$

Therefore, using the inequality $\|u + v\|^2 \leq 2\|u\|^2 + 2\|v\|^2$, we obtain

$$\mathrm{tr}(\Sigma) = \mathbb{E}\|X - \mathbb{E}X\|^2 \leq 2\,\mathbb{E}\|\Pi(X) - \mathbb{E}\Pi(X)\|^2 + 2\,\mathbb{E}\|R - \mathbb{E}R\|^2.$$

For the first term, by assumption,

$$\mathbb{E}\|\Pi(X) - \mathbb{E}\Pi(X)\|^2 = \mathrm{tr}(\mathrm{Cov}(\Pi(X))) \leq C_1 k.$$

For the second term, since covariance is dominated by the second moment,

$$\mathbb{E}\|R - \mathbb{E}R\|^2 = \mathrm{tr}(\mathrm{Cov}(R)) \leq \mathbb{E}\|R\|^2.$$

Using the approximation assumption,

$$\mathbb{E}\|R\|^2 = \mathbb{E}\|X - \Pi(X)\|^2 \leq C_2(\delta^2 + \eta).$$

Combining the two bounds yields

$$\mathrm{tr}(\Sigma) \leq 2C_1 k + 2C_2(\delta^2 + \eta).$$

By definition,

$$r_{\mathrm{eff}}(\Sigma) = \frac{\mathrm{tr}(\Sigma)}{\|\Sigma\|}.$$

Using the non-degeneracy assumption $\|\Sigma\| \geq c > 0$, we obtain

$$r_{\mathrm{eff}}(\Sigma) \leq \frac{2C_1 k + 2C_2(\delta^2 + \eta)}{c} = \frac{2C_1}{c}\, k + \frac{2C_2}{c}\,(\delta^2 + \eta).$$

In particular, if $\delta, \eta = O(1)$ and $\|\Sigma\|$ is bounded below by a positive constant, then

$$r_{\mathrm{eff}}(\Sigma) = O(k + 1).$$

If additionally $k \geq 1$, this simplifies to

$$r_{\mathrm{eff}}(\Sigma) = O(k).$$

□

PROOF OF THEOREM 3

Let $X \sim \mu$, and denote

$$\bar{x} := \mathbb{E}X, \qquad \Sigma := \mathrm{Cov}(X) = \mathbb{E}\big[(X-\bar{x})(X-\bar{x})^\top\big].$$

By the trace identity,

$$\mathbb{E}\|X-\bar{x}\|^2 = \mathbb{E}\,\mathrm{tr}\big((X-\bar{x})(X-\bar{x})^\top\big) = \mathrm{tr}\Big(\mathbb{E}(X-\bar{x})(X-\bar{x})^\top\Big) = \mathrm{tr}(\Sigma).$$

Next, by the assumptions of Lemma 1, we have

$$\mathrm{tr}(\mathrm{Cov}(\Pi(X))) \le C_1 k, \qquad \mathbb{E}\|X-\Pi(X)\|^2 \le C_2(\delta^2+\eta), \qquad \|\Sigma\| \ge c_0 > 0.$$

Therefore Lemma 1 yields

$$r_{\mathrm{eff}}(\Sigma) = \frac{\mathrm{tr}(\Sigma)}{\|\Sigma\|} \le \frac{2C_1}{c_0}\,k + \frac{2C_2}{c_0}\,(\delta^2+\eta).$$

Since

$$\mathrm{tr}(\Sigma) = \|\Sigma\|\, r_{\mathrm{eff}}(\Sigma),$$

it follows that

$$\mathrm{tr}(\Sigma) \le \|\Sigma\| \left(\frac{2C_1}{c_0}\,k + \frac{2C_2}{c_0}\,(\delta^2+\eta)\right).$$

Using $\mathbb{E}\|X-\bar{x}\|^2 = \mathrm{tr}(\Sigma)$, we obtain

$$\big(\mathbb{E}\|X-\bar{x}\|^2\big)^{1/2} = \sqrt{\mathrm{tr}(\Sigma)} \le \|\Sigma\|^{1/2} \left(\frac{2C_1}{c_0}\,k + \frac{2C_2}{c_0}\,(\delta^2+\eta)\right)^{1/2}.$$

It remains to prove the concentration statement. Since

$$p(x) = Z^{-1}e^{-U(x)}$$

and

$$\nabla^2 U(x) \succeq mI \qquad \text{for all } x \in \mathbb{R}^d,$$

the Bakry–Émery criterion implies that μ satisfies a logarithmic Sobolev inequality with constant of order $1/m$, up to the standard normalization convention. Consequently, by the Herbst argument, every 1-Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ satisfies a Gaussian concentration bound of the form

$$\mathbb{P}(|f(X) - \mathbb{E}f(X)| \ge t) \le 2\exp(-cmt^2) \qquad \text{for all } t \ge 0,$$

where $c > 0$ is an absolute constant.

Now define

$$f(x) := \|x-\bar{x}\|.$$

For any $x, y \in \mathbb{R}^d$,

$$|f(x)-f(y)| = \big|\|x-\bar{x}\| - \|y-\bar{x}\|\big| \le \|x-y\|,$$

so f is 1-Lipschitz. Applying the preceding concentration bound to this f gives

$$\mathbb{P}(|\|X-\bar{x}\| - \mathbb{E}\|X-\bar{x}\|| \ge t) \le 2\exp(-cmt^2) \qquad \text{for all } t \ge 0.$$

This proves the claimed concentration inequality and completes the proof. □

Corollary 1. *By Jensen's inequality,*

$$\mathbb{E}\|X-\mathbb{E}X\| \le \sqrt{\mathbb{E}\|X-\mathbb{E}X\|^2} = \sqrt{\mathrm{tr}(\Sigma)}.$$

Hence

$$\mathbb{P}\Big(\|X-\mathbb{E}X\| \ge \sqrt{\mathrm{tr}(\Sigma)} + t\Big) \le 2\exp(-cmt^2).$$

In particular, if δ, η are bounded by constants and $\|\Sigma\| \asymp 1$, then there exists a constant $C > 0$ such that

$$\sqrt{\mathrm{tr}(\Sigma)} \le C\sqrt{k},$$

and therefore

$$\mathbb{P}\Big(\|X-\mathbb{E}X\| \ge C\sqrt{k} + t\Big) \le 2\exp(-cmt^2).$$

B Interpretation of the Geometric Proxies

In this appendix, we explain why the empirical quantities used in our layer-selection procedure may be viewed as proxies for the theoretical geometric terms introduced in Section 3.1. Our goal is not to recover the true strong-convexity constant or the true dimension of the manifold of the distribution of layerwise representations. We instead seek observables that have the same *functional roles* in diffusion-friendly geometry.

Theoretical role of m. In our theory, the quantity m appears as the strong-convexity constant of the potential $U(x) = -\log p(x)$, namely

$$\nabla^2 U(x) \succeq mI.$$

Which means that in every direction the potential has at least curvature m. A larger m implies stronger contraction of the matching Langevin dynamics and greater robustness to score perturbations. For this reason, in the diffusion perspective m measures how strongly the representation distribution exhibits restoring geometry toward high-density regions.

In practice, however, the exact density $p(x)$ of layer activations is unknown, and only a finite sample of activation vectors is available. As a result, neither $U(x)$ nor its Hessian $\nabla^2 U(x)$ can be computed exactly. This motivates the use of empirical curvature-related proxies.

Why $\hat{m}_{\mathrm{mono}}$ is an m-like quantity. Our monotonicity proxy is defined from the empirical covariance Σ and its precision matrix $P = \Sigma^{-1}$. For sampled pairs (x_i, x_j), we compute

$$m_{ij} = \frac{(x_i - x_j)^\top P (x_i - x_j)}{\|x_i - x_j\|^2},$$

and summarize these values by a robust statistic such as the median.

This quantity is closely related to the role of m in the Gaussian case. Indeed, if the layerwise representation distribution were exactly Gaussian with mean μ and precision matrix P, then

$$U(x) = \frac{1}{2}(x - \mu)^\top P (x - \mu) + \mathrm{const},$$

and hence

$$\nabla^2 U(x) = P.$$

In that setting, the strong-convexity constant is precisely

$$m = \lambda_{\min}(P).$$

Moreover, for any displacement vector δ, the Rayleigh quotient

$$\frac{\delta^\top P \delta}{\|\delta\|^2}$$

measures directional curvature under the quadratic potential. Therefore, $\hat{m}_{\mathrm{mono}}$ can be interpreted as an empirical summary of the typical directional stiffness of the representation space. Although it is not an exact lower bound on the Hessian, it captures the same global geometric intuition: layers with larger $\hat{m}_{\mathrm{mono}}$ behave as if they are embedded in a more strongly restoring global geometry.

Why $\hat{m}_{\mathrm{curv}}$ is also an m-like quantity. Our local curvature proxy is based on the covariance of local neighborhoods. For each anchor point, we compute the covariance matrix Σ_{local} of its k-nearest neighbors and define a local score proportional to

$$\frac{1}{\lambda_{\max}(\Sigma_{\mathrm{local}})}.$$

The layer-level statistic $\hat{m}_{\mathrm{curv}}$ is then obtained by aggregating these local values across anchors.

This proxy is motivated by a local quadratic approximation. If a small neighborhood of the representation distribution is approximately Gaussian, then its local density may be written as

$$p_{\mathrm{local}}(x) \propto \exp\left(-\frac{1}{2}(x - \mu_{\mathrm{local}})^\top P_{\mathrm{local}}(x - \mu_{\mathrm{local}})\right),$$

with $P_{\text{local}} \approx \Sigma_{\text{local}}^{-1}$. Under this approximation, the local Hessian of the negative log-density is given by P_{local}, and the corresponding local strong-convexity scale is

$$\lambda_{\min}(P_{\text{local}}) = \lambda_{\min}(\Sigma_{\text{local}}^{-1}) = \frac{1}{\lambda_{\max}(\Sigma_{\text{local}})}.$$

Thus, $\hat{m}_{\text{curv}}$ measures how compact and sharply curved the local representation neighborhoods are. Larger values indicate smaller local spread along the most variable direction, which is consistent with the intuition of stronger local restoring geometry.

Why $\hat{k}$ reflects intrinsic dimension. The theoretical quantity k is intended to capture the intrinsic complexity of the representation distribution, rather than its ambient dimension d. In our empirical procedure, we measure this using the effective rank of the covariance:

$$\hat{k} = r_{\text{eff}}(\Sigma) = \frac{\text{tr}(\Sigma)}{\|\Sigma\|}.$$

This quantity has a direct interpretation in idealized low-dimensional settings. Suppose the data are supported exactly on a k-dimensional isotropic subspace with covariance eigenvalues

$$\lambda_1 = \cdots = \lambda_k = \sigma^2, \qquad \lambda_{k+1} = \cdots = \lambda_d = 0.$$

Then

$$r_{\text{eff}}(\Sigma) = \frac{k\sigma^2}{\sigma^2} = k.$$

Hence, in this ideal case, the effective rank exactly recovers the intrinsic dimension.

More generally, when the representation distribution is concentrated near a low-dimensional manifold or subspace, the covariance spectrum typically contains a small number of dominant eigenvalues and many small residual ones. In such cases, $r_{\text{eff}}(\Sigma)$ measures the effective number of active variation directions. For our purposes, this is the relevant notion of intrinsic dimension, since diffusion complexity depends not on the nominal ambient dimension but on how many directions carry meaningful variation.

Interpretation and limitation. Taken together, $\hat{m}_{\text{mono}}$ and $\hat{m}_{\text{curv}}$ serve as global and local proxies for curvature-related restoring geometry, while $\hat{k}$ serves as an operational measure of intrinsic dimension. These quantities are not exact estimators of the theoretical constants in Section 3.1. Instead, they are empirical surrogates designed to preserve the same geometric design principle: diffusion-friendly layers should be locally compact, globally stable, and effectively low-dimensional.

This distinction is important. Our empirical layer-selection score should therefore be interpreted as an operational criterion motivated by theory, rather than as a direct numerical estimate of the exact constants appearing in the theorems.

C Geometric Proxy Estimation and Layer Selection

This appendix describes how the empirical geometric proxies used in the *Locate* stage are computed and combined into a layer-selection score. Appendix B explains why these quantities can be interpreted as theory-motivated surrogates for curvature and intrinsic dimension; here we focus on their computation.

C.1 Representation Extraction

For each Transformer layer, we begin from hidden activations

$$\text{acts} \in \mathbb{R}^{N \times S \times H},$$

where N is the number of sequences, S the sequence length, and H the hidden dimension (e.g., $[32, 1024, 4096]$). These activations are converted into a set of representation vectors

$$x \in \mathbb{R}^{M \times H}$$

using one of the following extraction modes.

Mean pooling. Each sequence is mapped to one vector by masked averaging over valid tokens,

$$x_n = \frac{\sum_{t=1}^{S} a_{nt}\, h_{nt}}{\sum_{t=1}^{S} a_{nt}},$$

where $a_{nt} \in \{0, 1\}$ is the attention mask and $h_{nt} \in \mathbb{R}^H$ is the hidden state at position t.

Last-token pooling. Each sequence is represented by the hidden state of the last token.

Tokenwise extraction. All valid token representations are flattened across sequences and positions, and a random subset is retained if needed to control memory and computation.

Mask handling and default setting. Padding positions (`attention_mask`$= 0$) are excluded from all computations, and special tokens are kept whenever the attention mask marks them as valid. For reproducibility, we saved the number of sequences, sequence length, hidden dimension, and number and dimension of vectors before and after projection for every layer. The main experiments use `mean` pooling and retain at most `max_seqs`$= 3{,}000$ sequences and `max_tokens`$= 200{,}000$ valid tokens per layer-wise geometry run.

C.2 Random Projection

Because hidden representations are high-dimensional, the main geometry pipeline applies a fixed random projection before estimating the proxies. We construct a Gaussian matrix $R \in \mathbb{R}^{H \times D}$, orthonormalize its columns with a QR decomposition, and use the same matrix for every layer within a backbone. The projected representations are

$$z = xR \in \mathbb{R}^{M \times D},$$

with $D = 64$ in the main experiments. This projection is a computational sketch that reduces the cost and improves the numerical stability of repeated covariance-based estimation; it does not assert that the original hidden states are intrinsically 64-dimensional. All proxies in the main layer-selection experiments are computed from these projected representations, and Appendix F.2 shows that the resulting layer ranking is stable for larger projection dimensions and in the full hidden space.

C.3 Local Curvature Proxy

To capture local compactness, we sample anchor points from the projected representations and, for each anchor z_i, identify its k nearest neighbors $\mathcal{N}_i$ under Euclidean distance. We compute the ridge-regularized local covariance

$$\Sigma_{\text{local}}^{(i)} = \text{Cov}(\mathcal{N}_i) + \lambda I$$

and define

$$m_{\text{curv}}^{(i)} = \frac{1}{\lambda_{\max}\left(\Sigma_{\text{local}}^{(i)}\right)}.$$

This quantity is large when a neighborhood is compact even along its most variable direction. The layer-level statistic is the median over anchors,

$$\hat{m}_{\text{curv}} = \text{median}_i\big(m_{\text{curv}}^{(i)}\big),$$

and we additionally record the 25th and 75th percentiles for the uncertainty bands in Figure 2.

C.4 Monotonicity Proxy

To capture global directional stiffness, let

$$\Sigma = \text{Cov}(z) + \lambda I, \qquad P = \Sigma^{-1}$$

be the ridge-regularized covariance and its precision matrix. For sampled pairs (z_i, z_j), we compute the Rayleigh-quotient-type quantity

$$m_{ij} = \frac{(z_i - z_j)^\top P (z_i - z_j)}{\|z_i - z_j\|^2},$$

which measures how strongly a pairwise displacement is penalized by the global precision geometry. The layer-level statistic is

$$\hat{m}_{\text{mono}} = \text{median}_{(i,j)}(m_{ij}),$$

for which we also report quantiles and 95% bootstrap percentile intervals.

C.5 Effective Rank

We estimate effective intrinsic dimension with the effective rank of the projected covariance,

$$\hat{k} = r_{\text{eff}}(\Sigma) = \frac{\text{tr}(\Sigma)}{\|\Sigma\|},$$

where $\|\Sigma\|$ is the largest eigenvalue of Σ. Smaller values indicate that variation is concentrated along fewer dominant directions. Vectors are centered but not individually ℓ_2-normalized, the covariance is regularized by $10^{-3}I$, and covariance construction and eigendecomposition are performed in float64, with small negative eigenvalues from numerical error clamped to zero. Effective-rank values near one therefore indicate that the projected covariance is dominated by its leading direction; they do not imply that the original hidden-state space is one-dimensional. We additionally record the participation ratio and the 25th, 50th, and 75th percentiles of $\hat{k}$ as supplementary diagnostics.

D Diffusion Bridge

D.1 Architecture and Conditioning

The conditioning signal $c(x)$ is the embedding output $\text{embedding_out} \in \mathbb{R}^{B\times S\times H}$, i.e., the hidden state immediately before the first Transformer block of the source model. A learned condition projection maps it to the 768-dimensional `encoder_hidden_states` interface of the pretrained U-Net. No CLIP or other pretrained text encoder is used; the U-Net conditioning interface is repurposed to receive the embedding-derived representation.

The bridge follows a Stable-Diffusion-style latent-denoising design. The target hidden state h_{ℓ^*} is mapped to an image-like tensor (Appendix D.2), encoded by a frozen VAE into a latent z_{ℓ^*}, noised to z_t at timestep t with a fixed scheduler, and denoised by the U-Net, which predicts $\hat{\epsilon}_\theta(z_t, t, c)$. The denoised latent is decoded and projected back to the hidden space to obtain $\hat{h}_{\ell^*}$.

During training, we update the U-Net together with the bridge-specific projections (hidden-to-image, image-to-hidden, and condition projections), while the VAE and diffusion scheduler remain fixed. The Replace stage is thus trained as conditional hidden-state reconstruction rather than text-to-image generation. For the alignment losses in Section 3.3, the language-modeling term is $\mathcal{L}_{\text{LM}} = \text{CE}\big(p_{\text{bridge}}(x_{i+1} \mid x_{\le i}), x_{i+1}\big)$, and the temperature-scaled distillation term is $\mathcal{L}_{\text{KD}} = T^2\,\text{KL}\big(\text{softmax}(o_{\text{teacher}}/T) \,\|\, \text{softmax}(o_{\text{bridge}}/T)\big)$, where o_{teacher} and o_{bridge} are the teacher and bridge logits.

D.2 Hidden-to-Image Interface

For a boundary activation $H^{(\ell^*)} \in \mathbb{R}^{B\times S\times 4096}$, a learned 1×1 projection maps the hidden dimension to the three input channels of the SD-v1.5 VAE, and the sequence positions are rearranged bijectively in row-major order into a grid whose size depends on the sequence length: 32×32 for $S = 1024$ and 16×32 for $S = 512$. No positions are removed or doubled. After latent denoising, the inverse rearrangement restores the original token order, and a learned projection maps the three channels back to the hidden dimension.

The grid is only an architectural adapter for the pretrained latent-diffusion backbone; it does not assume that language representations have a natural two-dimensional topology or RGB semantics. Likewise, the three-channel projection is imposed by the VAE interface and is not a statement about the intrinsic dimensionality of language representations. Positions are processed jointly because the targets are contextual sequence representations, and processing each token independently would impose a factorized bridge. Our evaluation criterion is therefore functional rather than coordinate-wise: the reconstructed state should remain usable by the frozen suffix and LM head.

Table 6: Validation bridge loss of denoising backbones at layer 16 ($S = 512$, $N = 500$).

Bridge backbone	Mean validation loss ↓
Stable-Diffusion-style U-Net	**1163.59**
Hidden DDPM (MLP)	1411.36
Hidden DDPM (Transformer)	2799.28
Hidden DDPM (1D convolution)	3672.30

D.3 Autoregressive Inference

The VAE and U-Net are spatially non-causal within their inputs, but no future token information is provided to the bridge. At generation step t, the bridge condition contains the embeddings of the observed prefix $x_{\leq t}$; the remaining positions of the fixed-length grid are masked and carry no token information. Starting from diffusion noise ξ_t that is independent of the text, the bridge produces a boundary representation for the observed prefix,

$$\widehat{H}_t^{(\ell^*)} = D_\theta(C_t, \xi_t), \qquad C_t = [E(x_1), \ldots, E(x_t)].$$

The reconstructed prefix is passed through the unchanged suffix with its standard causal attention mask, and only the final observed position is used to predict the next token,

$$p_\theta(x_{t+1} \mid x_{\leq t}, \xi_t) = \mathrm{softmax}\Big(W_{\mathrm{LM}}\, F_{>\ell^*}^{\mathrm{causal}}\big(\widehat{H}_t^{(\ell^*)}\big)_t\Big).$$

The predicted token is appended, and the bridge is rerun on the extended prefix with fresh noise. The bridge can therefore mix information among observed positions but cannot access unobserved future text, so DiHAL remains autoregressive at the token-generation interface. Rerunning the bridge at every step is also a major source of the inference overhead in Appendix H.

D.4 Backbone Ablation

We use a Stable-Diffusion-style U-Net because the replacement module must solve a high-dimensional conditional denoising problem over continuous representations rather than a next-token prediction task. Table 6 compares denoising backbones under the same validation setting at layer 16 ($S = 512$, $N = 500$), a layer chosen independently of the geometry score. The SD-style U-Net achieves the lowest validation loss: 1163.59, a 17.6% reduction relative to the MLP-based hidden DDPM (1411.36), and reductions of 58.4% and 68.3% relative to the Transformer-based and 1D-convolutional DDPM bridges. Deterministic and other generative bridges are compared at the selected layer in Section 4.3. We do not claim that this architecture is generally optimal but it is the best performing backbones within the small-scale ablation.

E Experimental Setup Details

Data sampling and activation dumps. We construct the activation corpus from a local raw-text copy of *Dolma v1.7*. The pipeline samples up to 300,000 sequences with stratified round-robin sampling over 700 source files, with up to 500 samples per file and random seed 42. Each sequence is tokenized with the source-model tokenizer and passed through the backbone to store the embedding output, all decoder-layer outputs, input IDs, and attention masks as float16 shards. Special tokens are retained under the tokenizer's default handling, padding is tracked by the attention mask, and special tokens are removed only when token IDs are decoded back into text.

Geometry estimation. Geometry is estimated from mask-aware mean-pooled representations projected to $D = 64$ with a fixed random projection shared across layers (Appendix C). The default pipeline uses ridge coefficient 10^{-3}, $k = 64$ nearest neighbors and 512 anchor points for local curvature, 200,000 sampled pairs for monotonicity, and bootstrap resampling with 95% confidence intervals. Proxies are estimated from repeated subsamples as described in Section 4.1.

Fixed-budget layer sweep. Bridge training uses a deterministic shard-level split with validation ratio 0.1 and split seed 42. Each candidate layer uses a 150,000-example subset, split into up to

Table 7: Depth-controlled prediction of log bridge loss. Each auxiliary predictor is added to the same quadratic normalized-depth model and evaluated with leave-one-layer-out cross-validation. The best result for each backbone is in bold.

Backbone	Predictor	RMSE ↓	MAE ↓	R^2_{OOS} ↑
Llama-3.1-8B	Quadratic depth	0.7537	0.3769	0.0915
	Depth + CKA	0.6566	0.3011	0.3106
	Depth + hidden distance	0.8903	0.4311	−0.2676
	Depth + linear reconstruction	0.7626	0.3739	0.0700
	Depth + geometry	**0.5400**	**0.1792**	**0.5336**
Qwen3-8B	Quadratic depth	1.6102	1.3248	0.6983
	Depth + CKA	**1.5024**	1.2684	**0.7373**
	Depth + hidden distance	1.7306	1.4733	0.6515
	Depth + linear reconstruction	1.6539	1.3689	0.6817
	Depth + geometry	1.5316	**1.2550**	0.7271

135,000 training and 15,000 validation examples, and is trained for one epoch with batch size 4, learning rate 3×10^{-5}, AdamW, and FP16 mixed precision, for at most 33,750 optimization steps. Bridgeability is the validation bridge loss under this protocol; training losses are recorded only as optimization diagnostics.

Full training. The final bridge at each selected layer is trained on the full 300K-example corpus for four epochs within 40 H100 GPU-hours, which is also the diffusion/recovery training budget given to each baseline in Section 4.5. The auxiliary losses $\mathcal{L}_{\mathrm{LM}}$ and $\mathcal{L}_{\mathrm{KD}}$ are applied only in the final epoch.

Compute resources. All bridge training, including the layer sweep and full training, is conducted on NVIDIA H100 GPUs. Because activation dumps store float16 hidden states for every layer, storage grows approximately linearly with the number of layers and sampled sequences.

F Additional Layer-Selection Analyses

F.1 Depth and Embedding-Proximity Controls

The full layer sweep exhibits a strong shallow-to-deep trend, which raises the possibility that the geometry score mainly reflects proximity to the embedding representation. For each layer ℓ, we predict $y_\ell = \log \mathcal{L}_{\mathrm{bridge}}(\ell)$ with a base model in normalized depth d_ℓ,

$$y_\ell = \beta_0 + \beta_1 d_\ell + \beta_2 d_\ell^2 + \epsilon_\ell,$$

and augment it with one auxiliary predictor B_ℓ,

$$y_\ell = \beta_0 + \beta_1 d_\ell + \beta_2 d_\ell^2 + \beta_B B_\ell + \epsilon_\ell,$$

where B_ℓ is embedding-to-layer linear CKA, hidden-state distance from the embedding output, held-out linear reconstruction error from the embedding output, or the geometry score. All results use leave-one-layer-out cross-validation.

For Llama-3.1-8B, geometry gives the best held-out prediction on all three metrics, reducing RMSE by 28.4% relative to quadratic depth alone. For Qwen3-8B, geometry performs better only with depth and provides the lowest MAE at the same time, while CKA is slightly better in RMSE and out-of-sample R^2. Geometry therefore provides predictive information beyond layer index and simple linear recoverability, but does not uniformly dominate every embedding-proximity heuristic.

We also repeat the analysis on the first eight layers only, where the range of depth is much narrower. Adding geometry reduces RMSE by 14.1% for Llama but increases it by 3.9% for Qwen. This restricted analysis supports a conservative interpretation: geometry is informative beyond the global depth trend, particularly for Llama, but shallow depth remains an important predictor of bridgeability.

Table 8: Robustness of the geometry-score layer ranking to the projection dimension, measured against the default $D = 64$ ranking.

Projection dimension	Spearman ρ
128	0.9894
256	0.9876
512	0.9866
4096 (full space)	0.9874

Table 9: Trainable and active parameter counts for the diffusion or recovery module of each method. For CoDAR, counts are estimated from the reported MDLM-style diffusion backbone and the GPT-2-small-style autoregressive decoder with cross-attention under the Qwen tokenizer vocabulary.

Method	Trainable params	Active bridge/module params
DiHAL	862.7M	946.3M
DiHAL (small, from scratch)	257.4M	341.1M
Diffusion-LM	93.7M	93.7M
SED	73.7M	73.7M
LD4LG	25.6M AE + 188M diffusion	25.6M AE + 188M diffusion
CoDAR	130M diffusion + 230.9M AR decoder	360.9M

F.2 Projection-Dimension Robustness

To test whether the $D = 64$ sketch (Appendix C.2) determines the layer ranking, we recompute the complete geometry score with larger projection dimensions and in the full hidden space. Table 8 reports the Spearman correlation of each ranking with the default $D = 64$ ranking. Although the numerical values of individual covariance statistics change with the projection dimension, the layer ranking used for selection is highly stable.

G Baselines and Generation Metrics

G.1 Parameter Accounting

Table 9 reports parameter counts for the comparison in Section 4.5.

Parameters updated by the optimizer are called trainable, while those used during the forward pass of diffusion and kept frozen during training are called bridge/ module parameters. For DiHAL, the retained pretrained Transformer layers and LM head are not counted, although they are part of the active language-modeling interface.

G.2 Generation Metrics

Generative perplexity (Gen.PPL) is the perplexity that a pretrained GPT-2 evaluator assigns to generated samples. Diversity is defined as

$$\mathrm{Div} = \prod_{n=2}^{4} \frac{|\text{unique } n\text{-grams}|}{|\text{total } n\text{-grams}|},$$

the product of Distinct-2, Distinct-3, and Distinct-4. We follow the unconditional-generation protocol of LD4LG, also adopted by CoDAR, so that results are comparable with prior continuous diffusion language models. A GPT-2 evaluator can favor text close to the GPT-2 distribution and can reward predictable or repetitive outputs, so neither metric alone establishes overall generation quality for 8B-scale backbones.

G.3 Bridge-Capacity and Initialization Control

To test whether the result in Table 5 is driven by the size or pretrained initialization of the DiHAL bridge, we construct a smaller SD-style U-Net from scratch by reducing `layers_per_block`

Table 10: Bridge-capacity and initialization control (Qwen3-8B setting). The comparison matches the scale of the added bridge or recovery module, but not total active system capacity, because DiHAL retains the pretrained suffix and LM head.

Method	U-Net initialization	Bridge/recovery params	Gen.PPL ↓	Div. ↑
DiHAL (full)	Pretrained SD-v1.5	946.3M	**136.02**	**0.5913**
DiHAL (small)	From scratch	341.1M	140.51	0.5230
CoDAR	Method-specific	360.9M	144.83	0.4777

from 2 to 1 and using `block_out_channels`$= (224, 448, 832, 832)$. The resulting U-Net has 254.3M parameters. Together with the frozen 83.7M VAE and approximately 3.1M conditioning and projection parameters, the complete active bridge has 341.1M parameters, close to the 360.9M recovery module of CoDAR.

Despite a 64% reduction in bridge parameters and no pretrained U-Net initialization, the smaller bridge still outperforms CoDAR on both Gen.PPL and diversity. For the previous reason the result in Table 5 is not explained only by the bridge size or pretrained Stable-Diffusion weights. This remains a bridge-capacity control rather than a total-system-capacity or inference-FLOP-matched comparison.

H Inference Cost

H.1 Cost across Insertion Depths

The end-to-end inference cost of DiHAL depends on the insertion depth and on the number of denoiser evaluations (NFEs). Table 11 reports latency, throughput, and peak memory across insertion layers, bridge denoising, Transformer layers that are kept, and the LM head are reported as well.

At the geometry-selected shallow layers, DiHAL is slower than the original backbone even with a single denoising step, and latency grows substantially with the number of NFEs, which indicates that bridge denoising dominates the overhead. Deeper insertion layers can reduce latency at one NFE by skipping more Transformer layers, but these layers are not selected by the geometry criterion and are much harder to reconstruct. Peak memory is higher in every DiHAL setting because the bridge adds active modules. The current implementation should therefore not be interpreted as an acceleration method: deeper replacement removes more Transformer computation, but denoising overhead and reconstruction difficulty prevent practical speedups in the present setting.

H.2 NFE–Quality–Throughput Trade-off

Inference cost alone does not show how many denoising evaluations are needed for usable reconstruction. Table 12 therefore reports output fidelity and inference cost jointly at the geometry-selected Llama layer 3 for $N_{\mathrm{FE}} \in \{1, 4, 20\}$.

Increasing the denoising budget does not monotonically improve fidelity. Four NFEs give the best KL, NLL, and perplexity among the evaluated settings, but perplexity remains $18.94\times$ that of the original model while throughput falls to 28.4% of the original. At 20 NFEs, throughput falls to 8.5% and fidelity is worse than at four NFEs. The current shallow DiHAL configuration therefore does not offer a favorable end-to-end quality–efficiency trade-off relative to its source model.

I Sensitivity of the Retained Suffix

The final output depends both on boundary reconstruction error and on how the retained suffix responds to that error. We analyze this sensitivity with the trained Llama-3.1-8B layer-3 bridge, without additional training. Let h_{ℓ^*} be the original boundary representation and $\hat{h}_{\ell^*}$ the bridge reconstruction. We interpolate along the observed error direction,

$$h_\alpha = h_{\ell^*} + \alpha\big(\hat{h}_{\ell^*} - h_{\ell^*}\big), \qquad \alpha \in \{0, 0.25, 0.5, 0.75, 1, 1.5, 2\},$$

where $\alpha = 0$ recovers the original state, $\alpha = 1$ is the actual bridge output, and $\alpha > 1$ amplifies the error in the same direction. Each h_α is passed through the same frozen suffix and LM head.

Table 11: End-to-end inference cost. Latency is in milliseconds per token, throughput in tokens per second, and peak memory in GB. DiHAL cost includes bridge denoising, the retained Transformer layers, and the LM head.

Model	Method	Insert. layer	NFEs	Latency ↓	Throughput ↑	Peak mem. ↓
Llama-3.1-8B	Original	–	–	0.043	23,509	16.2
Llama-3.1-8B	DiHAL	3	1	0.074	13,481	16.8
Llama-3.1-8B	DiHAL	3	4	0.150	6,681	16.8
Llama-3.1-8B	DiHAL	3	20	0.503	1,989	16.8
Llama-3.1-8B	DiHAL	10	1	0.065	15,314	13.9
Llama-3.1-8B	DiHAL	10	4	0.146	6,844	13.9
Llama-3.1-8B	DiHAL	10	20	0.515	1,941	13.9
Llama-3.1-8B	DiHAL	20	1	0.051	19,467	9.9
Llama-3.1-8B	DiHAL	20	4	0.116	8,597	9.9
Llama-3.1-8B	DiHAL	20	20	0.503	1,988	9.9
Llama-3.1-8B	DiHAL	30	1	0.035	28,676	5.8
Llama-3.1-8B	DiHAL	30	4	0.114	8,798	5.8
Llama-3.1-8B	DiHAL	30	20	0.502	1,993	5.8
Qwen3-8B	Original	–	–	0.059	17,033	15.9
Qwen3-8B	DiHAL	2	1	0.081	12,375	16.9
Qwen3-8B	DiHAL	2	4	0.130	7,668	16.9
Qwen3-8B	DiHAL	2	20	0.503	1,989	16.9
Qwen3-8B	DiHAL	12	1	0.066	15,222	13.3
Qwen3-8B	DiHAL	12	4	0.132	7,554	13.3
Qwen3-8B	DiHAL	12	20	0.506	1,977	13.3
Qwen3-8B	DiHAL	22	1	0.047	21,156	9.7
Qwen3-8B	DiHAL	22	4	0.119	8,428	9.7
Qwen3-8B	DiHAL	22	20	0.500	2,002	9.7
Qwen3-8B	DiHAL	32	1	0.036	27,581	6.1
Qwen3-8B	DiHAL	32	4	0.093	10,769	6.1
Qwen3-8B	DiHAL	32	20	0.500	2,001	6.1

Table 12: Quality–efficiency trade-off across denoising budgets at the selected Llama-3.1-8B layer 3. The best DiHAL result is in bold and the original model is shown as a reference. The $N_{\mathrm{FE}} = 20$ row corresponds to Table 4.

Model	NFE	KL ↓	NLL ↓	PPL ↓	PPL ratio ↓	Latency ↓	Throughput ↑	Rel. throughput ↑
Original (reference)	–	0.00	1.87	6.47	1.00×	0.043	23,509	100.0%
DiHAL, L3	1	0.73	4.89	133.20	20.59×	**0.074**	**13,481**	**57.3%**
DiHAL, L3	4	**0.71**	**4.81**	**122.52**	**18.94**×	0.150	6,681	28.4%
DiHAL, L3	20	0.73	4.91	135.64	20.96×	0.503	1,989	8.5%

The suffix is sensitive to boundary perturbations: at $\alpha = 0.25$, teacher KL already rises to 0.21 and top-token agreement falls to 73.4%. The response is nonetheless gradual rather than showing a sharp failure threshold, with most of the degradation occurring before $\alpha \approx 0.75$. At the actual operating point ($\alpha = 1$), KL, NLL, and PPL are 0.73, 4.91, and 135.64, and doubling the same error direction increases PPL to 212.52.

We also track the discrepancy through the retained layers. For retained layer j, let

$$D_j = \frac{\left\|\hat{h}^{(j)} - h^{(j)}\right\|_2}{\left\|h^{(j)}\right\|_2},$$

where $h^{(j)}$ and $\hat{h}^{(j)}$ are computed from the original and reconstructed boundary states. The ratio D_j/D_{ℓ^*} increases gradually from $1.00\times$ at the insertion layer to about $1.44\times$ at the final block, so the suffix moderately amplifies reconstruction error without an abrupt layer-wise explosion.

J Existing Assets and Licenses

Our experiments use publicly available pretrained backbones, datasets, and model components. We use Llama-3.1-8B-Instruct under the Llama 3.1 Community License and Qwen3-8B under the Apache 2.0 license. We use Dolma v1.7 for activation extraction and bridge training under the ODC-BY license, and WikiText-103 for evaluation under its Creative Commons/GFDL licensing terms. The

Table 13: Sensitivity of the retained Llama suffix to interpolation along the observed reconstruction-error direction. $\alpha = 0$ corresponds to the original model and $\alpha = 1$ to the DiHAL result in Table 4.

α	Teacher KL ↓	NLL ↓	PPL ↓	Top-token agreement ↑
0.00	0.00	1.87	6.47	1.0000
0.25	0.21	3.07	21.52	0.7341
0.50	0.51	4.12	61.39	0.6079
0.75	0.70	4.77	118.21	0.6006
1.00	0.73	4.91	135.64	0.5913
1.50	0.77	5.05	155.72	0.5851
2.00	0.86	5.36	212.52	0.5809

diffusion bridge reuses Stable Diffusion v1.5 components under the CreativeML OpenRAIL-M license, and generative perplexity is computed with a GPT-2 evaluator under the GPT-2 license terms. We cite the original papers and model or dataset sources for all existing assets and use them only for research evaluation consistent with their respective terms.