Splat-Diffusion

Training-Free Contextual Steering of Diffusion Language Models via Dual-Pass Embedding Manifold Divergence

Nicole Kohm

Exobody Systems Inc. · Morrison, Colorado

ABSTRACT

We propose a method for maintaining contextual grounding in diffusion-based text generation without additional fine-tuning. The mechanism exploits a mathematical property of well-trained transformer embedding spaces: the Jacobian of the input-to-representation mapping is sufficiently smooth that representational structure learned during autoregressive pretraining remains valid under non-autoregressive generation topologies. A dual-pass embedding comparison, computing a candidate response's representation both in isolation and conditioned on conversational context, yields a scalar divergence signal that steers diffusion denoising toward contextually relevant outputs. We formalize why this works without retraining: the neuron-to-manifold type coercion inherent in transformer architectures preserves semantic continuity when Jacobians are smooth, allowing the pretrained model to serve as its own zero-shot relevance critic during parallel generation.

---

1 Problem Statement

Diffusion-based text generation offers parallel token production, global structural coherence, and iterative refinement. Removing sequential token prediction, however, eliminates the implicit planning mechanism autoregressive models exploit: each token chosen in light of all preceding tokens. The result is semantic drift. Diffusion-generated responses gravitate toward generic, context-insensitive outputs because the generation process lacks an intrinsic tether to the conversational context that prompted them.

Existing remedies (fine-tuning diffusion models on dialogue data, or adapting classifier-free guidance from image diffusion) require additional training. We argue this is unnecessary.

2 Core Thesis

The representational quality of a pretrained transformer, specifically its embedding manifold geometry, is a property of its weights, not its generation procedure. Autoregressive training optimizes for next-token prediction, but the side effect is a densely structured embedding space

where semantic similarity, topical relevance, and contextual dependency are encoded as geometric relationships: distances, angles, curvatures.

This structure survives changes in generation topology because the Jacobian of the mapping from token sequences to manifold positions is smooth.

Formally: let $\boldsymbol{f}_\theta : \boldsymbol{T}^* \to \mathbb{R}^d$ map a token sequence to its embedding representation under frozen parameters θ. If the Jacobian

$$\boldsymbol{J}_f = \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} \tag{1}$$

has bounded spectral norm and condition number across the relevant input domain, then small perturbations in input produce proportionally small perturbations in representation. The manifold geometry, and any signal derived from it, remains reliable even when the input was produced by a diffusion process rather than autoregressive sampling.

A single pretrained autoregressive model can therefore serve as both the generation substrate (its weights parameterize the diffusion denoiser) and the relevance critic (its embedding space provides the steering signal), with zero additional training.

3 Proposed Method

At each diffusion denoising step t, given candidate response tokens $\boldsymbol{r}_t$:

(a) Forward pass 1: Compute $\boldsymbol{h}_r = \boldsymbol{f}_\theta(\boldsymbol{r}_t)$. The embedding of the response in isolation.

(b) Forward pass 2: Compute $\boldsymbol{h}_{c|r} = \boldsymbol{f}_\theta(\boldsymbol{c} \oplus \boldsymbol{r}_t)$ restricted to response token positions. The embedding of the response conditioned on conversation $\boldsymbol{c}$.

(c) Compute divergence:

$$\delta_t = 1 - \cos\Big(\boldsymbol{h}_r, \boldsymbol{h}_{c|r}\Big) \tag{2}$$

(d) Steer: If δ_t falls below threshold τ (response semantics insufficiently shaped by context), inject a gradient nudge toward higher divergence, biasing the next denoising step toward candidates whose meaning is more context-dependent.

Threshold τ and nudge magnitude are hyperparameters. We propose adaptive scheduling tied to diffusion timestep: looser early in denoising when the response is still mostly noise, tightening as the output crystallizes and semantic identity stabilizes.

4 Theoretical Justification: The Type Coercion Argument

Individual neuron activations in a transformer are scalars, elements of $\mathbb{R}$. A hidden state is a vector in $\mathbb{R}^d$. Semantically meaningful representations do not fill $\mathbb{R}^d$ uniformly; they concentrate on a lower-dimensional manifold $\mathcal{M} \subset \mathbb{R}^d$ whose shape was sculpted by training data statistics.

The mapping from token sequences to points on $\mathcal{M}$ is a *type coercion*: discrete symbolic objects (tokens) are cast to continuous manifold coordinates (embeddings).

For this coercion to preserve semantic content, for nearby meanings to map to nearby points on $\mathcal{M}$, the Jacobian of the mapping must be smooth: bounded, non-degenerate, slowly varying. Autoregressive training with gradient descent inherently produces this. Exploding or collapsing Jacobians would surface as training instability (gradient explosion, vanishing gradients), so *stable convergence is indirect evidence of Jacobian regularity.*

The embedding manifold geometry and its Jacobian smoothness are properties of the frozen weights θ, not of how tokens were sampled. They remain valid when generation topology changes from autoregressive to diffusion. The dual-pass divergence signal is a geometric probe of $\mathcal{M}$. It asks: "does the conversation shift where this response sits on $\mathcal{M}$?" That question is well-posed regardless of how the response was produced.

No fine-tuning is required: the manifold is already smooth, the probe already works, and changing the generation process damages neither.

5 Experimental Design

5a Jacobian Spectral Analysis

For pretrained transformers at varying scale (1B, 7B, 70B parameters), compute the SVD of $\boldsymbol{J}_f$ at many sampled input points. Report condition numbers and spectral decay profiles.

Hypothesis: Well-trained models exhibit tightly bounded condition numbers with smooth spectral decay, confirming manifold regularity. Models that trained unstably or were undertrained should show measurably worse conditioning.

5b Signal Validation

Generate N candidate responses via diffusion for each of K conversational prompts. Compute δ for each candidate. Collect blind human relevance judgments. Report Spearman rank correlation between δ and human scores.

Hypothesis: δ is a reliable proxy for contextual relevance with no supervised training of the metric.

5c Steering Ablation

Four conditions:

(i) unsteered diffusion,
(ii) splat-diffusion with fixed τ,
(iii) splat-diffusion with adaptive τ scheduling,
(iv) conventionally fine-tuned diffusion baseline.

Evaluate on standard dialogue benchmarks measuring coherence, topical relevance, lexical diversity, and fluency.

Hypothesis: Adaptive splat-diffusion matches or closely approaches the fine-tuned baseline without any additional training.

5d Manifold Trajectory Visualization

Using UMAP or diffusion-map dimensionality reduction on the embedding space, plot the trajectory of response embeddings across denoising steps for steered versus unsteered conditions.

Visual hypothesis: Steered trajectories converge toward context-relevant manifold regions; unsteered trajectories diffuse outward into generic territory.

5e Compute Overhead Profiling

Profile wall-clock cost of dual forward passes per denoising step. Report overhead as a fraction of base diffusion generation cost. Assess feasibility of sparse stepping (running the divergence check every k-th step rather than every step) and characterize the quality-cost tradeoff curve.

6 Expected Contributions

A training-free method for contextual coherence in diffusion text generation. A formal grounding of the method in Jacobian smoothness of pretrained embedding manifolds. Empirical evidence that autoregressive representational quality transfers zero-shot to non-autoregressive generation steering. And a reframe that may be the most consequential piece:

> *Autoregressive training is not a generation strategy; it is a representational optimization whose products are topology-independent.*

7 Limitations and Open Questions

The Jacobian smoothness assumption may not hold uniformly across all manifold regions. Adversarial or far-out-of-distribution inputs could probe zones where the Jacobian degenerates. Characterizing the boundary of the safe region, the manifold subspace where the smoothness guarantee holds with confidence, is necessary follow-up work.

The dual forward pass doubles compute per steering check. Whether sparse checking at every k-th step preserves steering quality is an open empirical question.

The method inherits whatever biases exist in the pretrained model's embedding space. It creates no new representational capacity; it repurposes existing geometry. If the pretrained model has blind spots, splat-diffusion will share them.

Long-range coherence across multi-paragraph or multi-turn outputs may require more than a single scalar signal. Extensions to vector-valued or trajectory-valued steering, tracking not just "how different" but "different in which direction on $\mathcal{M}$," are a natural next step.