SPECTRAL SIGNATURES OF MEMORIZATION BASINS IN DIFFUSION TRAJECTORIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion models are known to reproduce individual examples from their training data as a form of *memorization* with implications for generalization, privacy, and downstream use. Recent work demonstrated that memorized training samples reside inside contiguous regions of the data/latent space known as *basins of memorization* – once a denoising trajectory enters a basin, under finite perturbations it irreversibly reproduces a memorized example. Prior empirical work on such basins in trained models has primarily focused on classifier-free guidance based detection and mitigation strategies in text-to-image models, leaving basin entry in trained diffusion trajectories and the formation of such basins largely uncharacterized. In this paper, we present a geometric characterization of memorization basins for conditional, unconditional, and rectified-flow-based diffusion models. We observe a characteristic signature in the local spectrum of the function geometry that distinguishes basin entry: the emergence of a dominant singular mode followed by a sharp spectral recovery, both markedly stronger along memorized trajectories. This signature appears at two timescales: along denoising trajectories during inference, and across training. Crucially, these geometric signatures are actionable: regularizing them can suppress memorization during training, while during inference they can help identify and redirect trajectories falling into memorization basins. The resulting geometry-informed interventions remain effective under low-dimensional approximations, and geometry-informed steering extends to large-scale text-to-image diffusion models. Our results show that local spectral geometry provides a common basis for characterizing memorization and for suppressing, detecting, and steering away from it across training and generation.

1 INTRODUCTION

Generative models are increasingly deployed as components of larger computational systems, where their outputs inform subsequent predictions, optimization, or decisions. Diffusion models exemplify this shift beyond image synthesis, generating molecular and protein structures (Wu et al., 2024), candidate solutions for combinatorial optimization (Sun & Yang, 2023), and actions and trajectories for planning and control (Chi et al., 2025).

As diffusion models take on this broader role, the relationship between generated samples and training data becomes increasingly consequential. These models can closely reproduce individual training examples (Carlini et al., 2023), raising concerns around privacy, copyright, and generalization. Memorization is not inherently undesirable, however: more broadly in deep learning, retaining information about individual or atypical training examples can contribute positively to generalization (Feldman & Zhang, 2020). Its significance therefore depends on what is retained and on the role of the learned model. In diffusion models, memorization is also highly selective: only a subset of training examples are reproduced, with susceptibility to memorization depending on properties of both the data and the learned distribution (Carlini et al., 2023; Bonnaire et al., 2026).

This raises the question of whether memorization leaves a structured signature in the learned model. Retained training information can be reflected in local model geometry (Nickl et al., 2023) and admit compact spectral representations (Jung et al., 2026). In generative models, the manifold memorization hypothesis provides a complementary perspective, associating memorized samples with locally constrained regions of the learned distribution (Ross et al., 2025).

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.

If memorized samples occupy locally constrained regions of the learned distribution, local geometry provides a basis not only for characterizing memorization, but for intervening on it. In diffusion models, intermediate denoising states offer a time-resolved view of this geometry as trajectories evolve through distinct dynamical regimes (Biroli et al., 2024).

We therefore ask: *when does a denoising trajectory become committed to a memorized outcome, what geometric structure accompanies this commitment, and can it guide intervention?*

We adopt the framework of *memorization basins*, defining them as regions from which a trajectory cannot escape a memorized outcome under bounded perturbations. Basin entry therefore provides a dynamical reference for relating memorization to local score geometry.

Around this transition, the local Jacobian spectrum undergoes a characteristic reorganization associated with basin entry, accompanied by changes in perturbation sensitivity and generative dynamics. The same geometric organization develops systematically during training: intermediate-noise Jacobians form an increasingly dominant singular mode, while low-noise spectra exhibit stronger rank recovery. Local spectral geometry thus reveals memorization across two timescales: the evolution of the score field during training and of individual trajectories during generation.

These observations motivate three complementary mechanisms for acting on memorization. GeoReg-DM regularizes local Jacobian geometry during training to suppress its development; during generation, *Jacobian Effective Rank* (JER) leverages the endpoint of this spectral signature as an efficient local test for memorized generations, while GeoSteer-DM uses dominant singular directions to redirect memorized trajectories. Together, they use the same local geometry to suppress, detect, and steer away from memorization.

2 Related Work

Memorization and geometry in generative models. Memorization is a longstanding phenomenon in overparameterized neural networks, which can retain individual or atypical training examples while still generalizing (Feldman & Zhang, 2020). In generative models, memorization becomes directly observable through training-example reproduction (Carlini et al., 2023; Somepalli et al., 2023a), with subsequent work studying how it develops during diffusion training and how retained examples are represented within the model (Bonnaire et al., 2026; Hintersdorf et al., 2024; Kowalczuk et al., 2025). Local geometry provides a complementary perspective on generative distributions (Meng et al., 2021), including intrinsic dimensionality, manifold structure, and spectral regimes across diffusion time (Stanczuk et al., 2024; Humayun et al., 2025; Ventura et al., 2025). Memorization has specifically been associated with locally constrained or sharp regions of the learned distribution (Ross et al., 2025; Jeon et al., 2024), while related analyses study denoiser-Jacobian eigenstructure (Kadkhodaie et al., 2024), variance-dependent loss of tangent dimensions (Achilli et al., 2024), and log-probability anisotropy (Asthana & Belagiannis, 2026). We build on these analyses by resolving spectral geometry along individual trajectories and across training, and by using its singular directions for intervention.

Memorization mitigation and geometric control. Existing mitigation methods intervene during training through the data or optimization procedure (Dutt et al., 2025; Shah et al., 2025), or during generation through conditioning, attention, model predictions, initial noise, and internal model components (Somepalli et al., 2023a; Ren et al., 2024; Zhao et al., 2026; Wen et al., 2024; Han et al., 2026). Separately, geometric editing methods use Jacobian singular directions and learned subspaces to control generation (Chen et al., 2024; Gaintseva et al., 2026b;a). Our approach connects these perspectives by using memorization-associated local spectral geometry as an intervention target during training and as a signal for detection and control during generation.

3 The Dynamics and Geometry of Memorization

3.1 Memorization Basins

The manifold memorization hypothesis (Ross et al., 2025) associates memorization with locally constrained regions of the learned distribution that restrict the effective degrees of freedom of gener-

ated samples. Statistical-physics analyses further identify a collapse transition beyond which reverse trajectories become attracted to individual training samples (Achilli et al., 2024). Building on this picture, we characterize the denoising trajectory dynamics for a fixed learned model and sampler. Specifically, for perturbations bounded at each sampling step, we define a *memorization basin* as a region from which no admissible continuation of a denoising trajectory can avoid memorization.

Let $\Phi_{t\to 0} : \mathbb{R}^d \to \mathbb{R}^d$ denote the continuation of a fixed deterministic reverse process from timestep t to its final output, and let $\mathcal{M} \subset \mathbb{R}^d$ denote the set of outputs classified as memorized.

To formalize basin entry, we consider bounded perturbations to the remaining reverse process. Let $T = t_N > \cdots > t_0 = 0$ denote the sampling schedule, and let F_k map the state at t_k to the next state at t_{k-1}. Starting from x_t, we allow each remaining transition to be perturbed as

$$x_{k-1} = F_k(x_k) + u_k, \qquad \|u_k\|_2 \le \delta.$$

We denote the set of all such perturbation sequences by $\mathcal{U}_t(\delta)$ and the resulting terminal output by $\Phi^u_{t\to 0}(x_t)$. The per-step bound δ therefore specifies the magnitude of the admissible change at each remaining sampling step.

Definition 1 (Memorization basin)**.** *For timestep t and perturbation bound $\delta \ge 0$, the* memorization basin *is*

$$\mathcal{B}_t(\delta) = \left\{ x_t \in \mathbb{R}^d : \Phi^u_{t\to 0}(x_t) \in \mathcal{M} \quad \textit{for all } u \in \mathcal{U}_t(\delta) \right\}.$$

Thus, $x_t \in \mathcal{B}_t(\delta)$ when no continuation within the per-step bound can escape $\mathcal{M}$. Basin membership is defined relative to the fixed sampler, sampling schedule, and bound δ. Importantly, these basins are *absorbing*: if an admissible continuation carries $x_t \in \mathcal{B}_t(\delta)$ to a later state x_s in sampling order, then $x_s \in \mathcal{B}_s(\delta)$ (Appendix A.1). Otherwise, an escape from x_s could be concatenated with the preceding continuation to construct an escape from x_t. Basin entry therefore marks a persistent transition beyond which memorization cannot be avoided within bound δ.

Definition 2 (Basin entry time)**.** *For a reverse trajectory (x_t) that enters a memorization basin, its* basin entry time *is*

$$t^\dagger = \max\{t : x_t \in \mathcal{B}_t(\delta)\}$$

After $t^\dagger$ in sampling order, every δ-bounded continuation terminates within $\mathcal{M}$, so $t^\dagger$ marks the onset of basin entry. It serves as the formal notion of commitment against which we evaluate the spectral signatures below.

3.2 Spectral Geometry of Memorization

While the basin entry time $t^\dagger$ provides a rigorous condition for trajectory absorption, evaluating it directly requires exhaustive perturbations of the remaining trajectory and denoising each of the perturbed trajectories. We therefore seek an easy to observe signature of a trajectory "falling" into a memorization basin. Prior work shows that deep neural networks including diffusion models learn smooth functions which can be locally characterized using their input–output Jacobians (Balestriero & Baraniuk, 2018; Humayun et al., 2025). The Jacobian spectrum of denoisers have also previously been used to study inductive biases and generalization (Kadkhodaie et al., 2024). If $x_t \in \mathbb{R}^d$ denotes an intermediate state of a trajectory with marginal density p_t, the local geometry of p_t along the trajectory can be characterized using the Jacobian of the score function which is equivalent to the Hessian of the log-density $H_t(x_t) = \nabla_{x_t} s_t(x_t) = \nabla^2_{x_t} \log p_t(x_t)$. Jeon et al. (2024) characterized memorization using the trace of the Hessian to identify regions with highly localized density. We, however, are interested in the rate of change in the log-density landscape along different data space directions during denoising – which can be characterized using the Jacobian spectrum of the score function.

For a trained denoiser approximating the score $s_t(x_t)$ via the forward process $x_t = \alpha_t x_0 + b_t \epsilon$, Tweedie's formula expresses the clean-estimate Jacobian as an affine function of the Hessian, $\frac{\partial \hat{x}_0}{\partial x_t} = \alpha_t^{-1}\big(I + b_t^2 H_t(x_t)\big)$. In an ϵ-parameterized network, this relationship simplifies to:

$$J_t = \frac{\partial \epsilon_\theta(x_t, t)}{\partial x_t} \approx -b_t H_t(x_t).$$

with equality for the optimal model. Appendix A.2 derives the corresponding operator for rectified flow. We analyze the directional structure of J_t through its singular-value decomposition

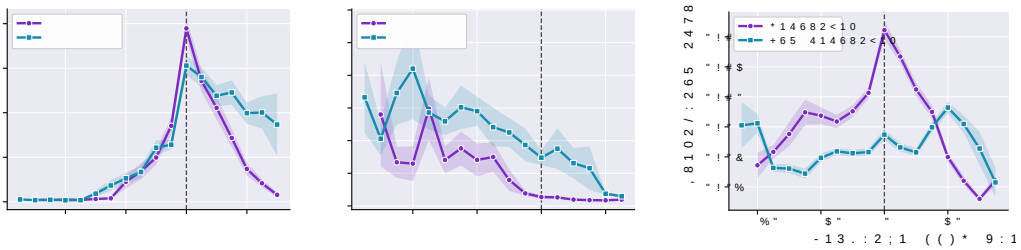


Figure 1: **Spectral and behavioral signatures of memorization during inference.** DDIM trajectories on CelebA are aligned at $\tau = t - t^*$, with t^* defined by the steepest increase in the leading spectral gap. **(a)** Both groups exhibit spectral concentration around $\tau = 0$, with stronger concentration and subsequent recovery for memorized trajectories. **(b)** Memorized trajectories become less sensitive to perturbations after the transition. **(c)** Their predicted clean samples approach the final output most rapidly around the same region.

$J_t = U_t \Sigma_t V_t^\top$, with $\sigma_1(t) \geq \cdots \geq \sigma_r(t) \geq 0$, and summarize its spectral concentration and dimensionality using relative spectral gap, stable rank, and effective rank (defined in Appendix A.4). Because J_t serves as a curvature proxy, we also verify that it is nearly symmetric and sign-coherent near the data endpoint (empirically verified in Appendix A.3).

Spectral collapse and recovery of denoising trajectories. We adopt the CelebA setting of Bonnaire et al. (2026), training a U-Net diffusion model on 4,096 images and generating samples with DDIM. We classify a generation as memorized using the nearest-neighbor distance-ratio criterion of Yoon et al. (2023). Along all denoising trajectories, the local Jacobian spectrum undergoes a phase transition, visible as a sharp rise in the relative spectral gap Δ_{12} (Figure 1a). In terms of stable rank, this transition appears as a spectral collapse followed by recovery (Appendix F.2). The dynamic is driven largely by the top singular value. During collapse, σ_1 overshoots the remaining singular values and concentrates the spectrum around a single dominant mode. During recovery, the remaining singular values catch up and the spectrum relaxes toward a more isotropic structure as generation approaches low noise. Memorized trajectories undergo a stronger version of this cycle, with a sharper overshoot of σ_1 followed by a more pronounced recovery. As a result, Δ_{12} separates memorized from non-memorized trajectories at inference time (Figure 1a).

Because individual trajectories undergo collapse–recovery at different timesteps, we define t^* as the state preceding the steepest increase in Δ_{12} and align trajectories by $\tau = t - t^*$. After t^*, memorized trajectories become substantially less sensitive to finite perturbations, and their predicted clean samples rapidly approach the final output (Figure 1b–c). Over the same region, bounded adversarial perturbations progressively lose their ability to redirect trajectories away from memorized outcomes (Appendix F.1). Together, these results suggest that the spectral transition at t^* is an observable marker of declining escapeability, i.e., of the region where trajectories enter the memorization basin.

Connecting dynamical collapse to local memorization geometry. This collapse–recovery dynamic bridges trajectory-level views of memorization (Biroli et al., 2024; Achilli et al., 2024) with local geometric accounts (Ross et al., 2025; Stanczuk et al., 2024). While the geometry near a memorized endpoint approaches an isotropic I/b_t structure (Appendix A.5), trajectories reach this regime by first passing through a phase of strong spectral concentration around basin entry. This intermediate regime isolates a dominant singular direction near t^*, providing the geometric direction that GEOSTEER-DM later exploits to redirect memorized trajectories.

Spectral concentration emerges during training. Furthermore, the geometric organization observed during denoising develops progressively during training (Figure 2). As training advances and population-level memorization increases, the local Jacobian becomes increasingly concentrated around a dominant mode at intermediate noise levels, reflected by a pronounced rise in the relative spectral gap Δ_{12} around $t \in [200, 400]$ that grows with memorization rate (Figure 2a). At lower noise levels ($t \leq 200$), the spectral gap closes and the local entropy rank $\nu(J_t)$ recovers toward a less concentrated spectrum at later checkpoints (Figure 2b). These observations provide a training-time regularization target, directly motivating our approach GEOREG-DM.

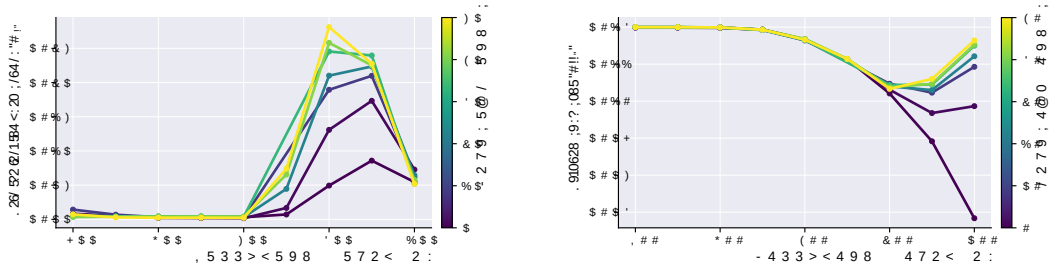


Figure 2: **Training-time evolution of spectral concentration.** Local Jacobian metrics for different training checkpoints colored by memorization rate. **(a)** The relative leading spectral gap Δ_{12} rises sharply at intermediate noise, and falls again at low noise. Its peak grows with memorization rate and shifts toward higher noise in later checkpoints. **(b)** Entropy effective rank reveals matching intermediate concentration followed by low-noise rank recovery in later checkpoints.

4 CONTROLLING MEMORIZATION THROUGH SPECTRAL GEOMETRY

Our geometric analysis highlights two key points of intervention: regularizing the dominant intermediate-noise mode during training, and using local singular directions to detect and steer trajectories at inference time. We translate these insights into three complementary mechanisms: GEOREG-DM for training, JER for detection, and GEOSTEER-DM for inference-time steering.

4.1 GEOREG-DM: SPECTRAL REGULARIZATION DURING TRAINING

We introduce GEOREG, which suppresses the dominant local mode during training by penalizing the squared leading singular value σ_1 of a low-dimensional Jacobian sketch. Its gradient is determined by the leading singular pair of the sketch. Since computing σ_1 from the full $d \times d$ Jacobian at every selected training state would be prohibitively expensive, we instead use a low-dimensional sketch $\widehat{J}_t = WJ_t$, where $W \in \mathbb{R}^{k\times d}$ has orthonormal rows. The sketch is constructed directly through vector–Jacobian products without materializing J_t, providing a low-dimensional view of its dominant structure (Halko et al., 2011; Woodruff, 2014). Since $\widehat{J}_t$ has only k rows, its leading squared singular value is obtained from the largest eigenvalue of the $k \times k$ Gram matrix $\widehat{J}_t\widehat{J}_t^\top$.

We apply the regularizer to a selected subset $\mathcal{I}$ of the minibatch,

$$\mathcal{L}_{\text{geo}} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \sigma_1\left(\widehat{J}_{t_i}^{(i)}\right)^2, \tag{1}$$

Differentiating through the Jacobian sketch introduces mixed derivatives $\partial^2\epsilon_\theta/(\partial\theta\,\partial x)$. To keep this computation tractable, we regularize only a small subset of each minibatch, selected using the input sensitivity of the ordinary denoising loss, $\pi^{(i)} = \left\|\nabla_{x_{t_i}^{(i)}}\ell_{\text{base}}^{(i)}\right\|_2$.

This quantity is obtained from the base-loss backward without constructing Jacobian sketches. We then allocate the geometric computation to the states with highest sensitivity. Full differentiation, spectral computation, and selection details are given in Appendix D.

4.2 JER: EFFICIENT DETECTION FROM LOCAL SPECTRAL GEOMETRY

The spectral transitions observed in our analysis suggest that stable and effective rank provide a direct geometric signal for memorization detection. However, both depend on the Jacobian spectrum and are expensive to evaluate at inference time. We therefore propose the *Jacobian Effective Rank* (JER), a local proxy for spectral dimensionality designed to retain this geometric signal without explicitly computing the Jacobian spectrum. To obtain an efficient estimator, we leverage a connection to Rényi-2 effective rank.

For a Jacobian J with singular values $\{\sigma_i\}$, the Rényi-2 effective rank is

$$\mathrm{ER}_2(J) = \frac{(\sum_i \sigma_i)^2}{\sum_i \sigma_i^2} = \frac{\|J\|_*^2}{\|J\|_F^2}. \tag{2}$$

This quantity provides a useful bridge between the two rank measures underlying our geometric analysis: it satisfies $\mathrm{SR}(J) \le \mathrm{ER}_2(J) \le \nu(J)$ (Appendix B.1), while depending only on the nuclear and Frobenius norms rather than the full singular-value distribution.

To approximate the remaining nuclear norm, we leverage our signed-curvature analysis (Appendix A.3), which shows that the curvature spectrum is strongly sign-coherent near the data endpoint. We exploit this structure by replacing the nuclear norm with the oriented trace, which becomes exact for a symmetric semidefinite operator. This yields JER as a local trace-to-Frobenius proxy for the spectral dimensionality captured by ER_2.

To evaluate JER without constructing the Jacobian, we leverage standard stochastic trace estimation with Rademacher probes (Hutchinson, 1989). For independent $z_1, \dots, z_m$ and vector–Jacobian products $q_\ell = J^\top z_\ell$, we estimate $\widehat{\mathrm{tr}}_m(J) = m^{-1} \sum_\ell z_\ell^\top q_\ell$ and $\widehat{F}_m(J) = m^{-1} \sum_\ell \|q_\ell\|_2^2$, giving

$$S_{\mathrm{JER}}(J) = \frac{\left[a\, \widehat{\mathrm{tr}}_m(J)\right]_+^2}{\widehat{F}_m(J) + \varepsilon}, \tag{3}$$

where $a \in \{-1, +1\}$ selects the curvature orientation. Both estimates reuse the same vector–Jacobian products, so the resulting detector requires neither explicit Jacobian construction nor an SVD. At inference, JER is evaluated once at a fixed low-noise query state and thresholded using a calibration set. Details of the estimator and query construction are given in Appendix B.2.

4.3 GeoSteer-DM: Geometry-Informed Steering During Generation

GeoSteer-DM uses the local geometry to redirect a selected trajectory during denoising. The spectral analysis above shows that memorization is accompanied by the emergence of a dominant local Jacobian mode. While the singular values measure the strength of this concentration, the corresponding right singular vectors identify its directions in state space. We therefore use the leading right singular vector as the minimal direction on which to intervene, testing whether modifying the dominant geometric mode can redirect the trajectory away from memorization.

Different generative models parameterize their updates differently: DDIM predicts noise, rectified flow predicts velocity, and classifier-free guidance combines conditional and unconditional predictions. We therefore formulate GeoSteer-DM in terms of the model-induced displacement rather than modifying the model output directly.

At sampler step n, let x_{n+1}^{base} denote the ordinary update and x_{n+1}^{0} a corresponding anchor update from the same current state. We define the model-controlled displacement as

$$\delta_n = x_{n+1}^{\mathrm{base}} - x_{n+1}^{0}. \tag{4}$$

For unconditional DDIM and rectified flow, the anchor uses zero model output; under classifier-free guidance, it uses the unconditional prediction, so that δ_n isolates the contribution of conditional guidance. This provides a common displacement on which the same steering rule can act across model parameterizations; the sampler-specific formulations are given in Appendix C.

As a simple geometry-informed steering rule, we act on the leading right singular direction of the local Jacobian. This direction captures the dominant local spectral mode identified by our geometric analysis. Hence, we use it to construct a minimal intervention that modifies only this dominant component of the model-controlled displacement. Let $J_{t_n} = U_{t_n} \Sigma_{t_n} V_{t_n}^\top$ and let $v_1(t_n)$ denote its leading right singular vector. We define the corresponding projector

$$P_n = v_1(t_n) v_1(t_n)^\top$$

and decompose the model-controlled displacement as

$$\delta_n = P_n \delta_n + (I - P_n)\delta_n.$$

The two terms isolate the component along the leading geometric direction and its orthogonal complement, respectively.

We choose a reflection as a simple norm-preserving intervention on this component,

$$\delta_n^{\text{geo}} = (I - 2P_n)\delta_n, \qquad x_{n+1}^{\text{geo}} = x_{n+1}^0 + \delta_n^{\text{geo}}. \tag{5}$$

This reflection is invariant to the sign of $v_1(t_n)$ and preserves the displacement norm, $\|\delta_n^{\text{geo}}\|_2 = \|\delta_n\|_2$, so steering changes the direction of the model-induced update without attenuating its magnitude. The local geometry is recomputed after each modified transition, allowing the steering direction to adapt to the evolving trajectory.

5 EXPERIMENTS

We evaluate our framework in two complementary settings. We first train unconditional diffusion and rectified-flow models on a controlled CelebA subset, where memorization is measured by the nearest-neighbor gap ratio to the training set. We then test JER and GeoSteer-DM on Stable Diffusion v1.4, measuring memorization by SSCD similarity to known training references, and compare against existing inference-time baselines[1]. Complete experimental details are found in Appendix E.

5.1 SUPPRESSION, DETECTION, AND STEERING IN UNCONDITIONAL DIFFUSION

In the unconditional diffusion setting, we use our geometric framework to address three core questions: whether GEOREG-DM suppresses memorization while preserving coverage, whether JER detects memorized trajectories efficiently, and whether GEOSTEER-DM redirects them.

Can GeoReg-DM suppress memorization without sacrificing coverage? GeoReg-DM regularizes the local Jacobian geometry during training to suppress emerging memorization. We first evaluate it on three synthetic populations (Gaussian, Gaussian mixture, and 8D Swiss roll), where access to the true data distribution allows us to distinguish training-set copying from population coverage. Across all three settings, GeoReg-DM achieves the strongest reduction in copying while consistently preserving or improving recall (Table 1). In contrast, the alternative training strategies either produce smaller reductions in copying or trade them against population coverage. GeoReg-DM therefore improves the memorization–coverage trade-off without restricting the learned distribution.

This effect extends directly to image generation on CelebA32, where GeoReg-DM substantially reduces copying while increasing the mean nearest-neighbor gap ratio, outperforming alternative training strategies on both measures (Table 1). This suppression comes with only a modest increase in FID, which is expected in this setting: removing exact training-set reproductions can itself increase FID even when perceptual quality and sample diversity are preserved. We investigate this effect further, together with representative generations and additional perceptual-quality analyses, in Fig. 10 and Appendix F.3. Together with the synthetic results, these findings show that GeoReg-DM suppresses training-set copying without simply contracting the learned distribution, and that this behavior transfers from controlled populations to image generation.

Table 1: **GeoReg reduces memorization across synthetic populations and CelebA32.** Results are averaged across five independent seeds for each setting. For synthetic populations, memorization is measured against the fixed training set, while precision and recall are evaluated against independent population samples. For CelebA32, r denotes the nearest-neighbor gap ratio. Full baseline definitions, training settings, and additional metrics are given in Appendix E.1.

	Gaussian80			Mixture32			SwissRoll8D128			CelebA32		
Method	Mem ↓	Prec. ↑	Rec. ↑	Mem ↓	Prec. ↑	Rec. ↑	Mem ↓	Prec. ↑	Rec. ↑	Mem ↓	Mean r ↑	FID ↓
Standard	31.85	0.978	0.807	46.85	0.997	0.757	16.10	**0.928**	0.816	8.80	0.766	54.89
Tikhonov	25.15	0.981	0.921	43.25	**0.998**	0.812	13.90	0.918	0.839	6.50	0.780	**49.03**
Reduced LR	21.15	**0.986**	0.880	41.35	0.992	0.771	10.30	0.877	**0.899**	6.80	0.785	52.21
Ambient-style	24.35	0.985	0.858	47.00	0.993	0.600	17.30	0.917	0.789	4.90	0.808	52.15
GeoReg (Ours)	**15.85**	0.983	**0.979**	**19.25**	0.980	**0.963**	**10.20**	0.904	0.878	**1.56**	**0.846**	59.77

[1]Code will be released upon acceptance.

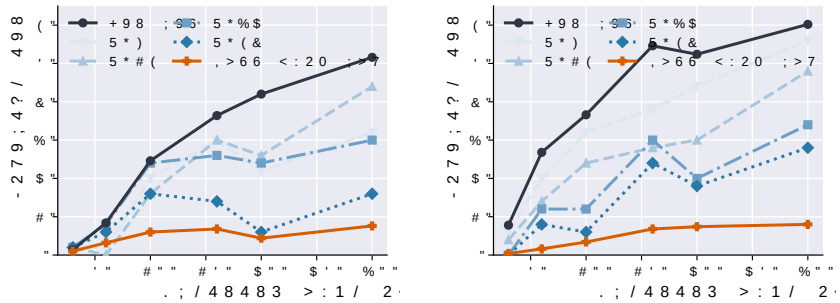


Figure 3: **GEOSTEER-DM suppresses memorization across training checkpoints and geometric resolutions.** Memorization rate of unmodified and steered sampling for (a) DDIM and (b) rectified flow. Full-Jacobian steering gives the strongest suppression. Low-dimensional Jacobian sketches retain much of this effect, with suppression generally strengthening as k increases.

Local Jacobian geometry enables efficient post-hoc detection. As a baseline, we use temporal recovery, which follows directly from our spectral analysis and scores how strongly the spectrum recovers after its collapse along a trajectory. However, it requires full Jacobian spectra at multiple points along the trajectory, which makes it expensive. Comparing it with local terminal geometry on held-out CelebA trajectories shows that JER captures a strong memorization signal at substantially lower computational cost (Table 2). For DDIM, JER achieves the strongest overall detection, outperforming both local baselines and full temporal recovery. For rectified flow, temporal recovery remains more discriminative, while JER still outperforms all other local geometric detectors. In both cases, JER requires only tens of milliseconds per sample, compared with tens of seconds for tracking spectral recovery along the full trajectory. This contrast suggests a difference in how memorization information is distributed across the sampling approaches.

Table 2: **Efficient geometry-based memorization detection.** Results on held-out CelebA trajectories. AUROC and AUPRC denote the areas under the receiver-operating-characteristic and precision–recall curves, respectively; MCC denotes the Matthews correlation coefficient. Recall, specificity, and MCC are computed at operating points selected on a disjoint calibration set and frozen before evaluation. For the unconditional score-magnitude baseline, classifier orientation is also selected on the calibration split and then frozen for held-out evaluation.

Sampler	Method	AUROC ↑	AUPRC ↑	Recall ↑	Spec. ↑	MCC ↑	Time ↓
DDIM	FLIPD Ross et al. (2025)	0.845	0.722	0.437	0.932	0.443	42.05 ms
	Normal bundle Stanczuk et al. (2024)	0.902	0.773	0.452	0.952	0.493	922.19 ms
	Unconditional score magnitude	0.860	0.785	0.542	0.934	0.537	**12.05 ms**
	Spectral temporal recovery	0.823	0.698	0.305	**0.956**	0.366	40.88 s
	JER (ours)	**0.956**	**0.895**	**0.898**	0.886	**0.763**	71.26 ms
RF	FLIPD Ross et al. (2025)	0.720	0.733	0.310	0.894	0.248	42.48 ms
	Normal bundle Stanczuk et al. (2024)	0.760	0.746	0.338	0.897	0.278	922.12 ms
	Unconditional score magnitude	0.783	0.800	0.375	0.929	0.358	**11.98 ms**
	Spectral temporal recovery	**0.912**	**0.931**	**0.789**	0.879	**0.667**	38.31 s
	JER (ours)	0.870	0.844	0.403	**0.935**	0.392	57.41 ms

Local geometry enables actionable, low-rank trajectory steering. Using the full Jacobian spectrum, GEOSTEER-DM substantially suppresses memorization across training checkpoints for both DDIM and rectified flow (Figure 3), reducing late-stage memorization from 51.6% to 7.6% and from 60.2% to 8.0%, respectively. Importantly, effective intervention does not require resolving the full Jacobian: low-rank sketches with $k \in \{8, \ldots, 64\}$ retain sufficient directional information to substantially reduce memorization, with performance improving as geometric resolution increases. At $k = 64$, aggregate memorization falls to 10.0% for DDIM and 14.0% for rectified flow. These results show that the spectral directions associated with basin entry are not merely diagnostic but provide actionable directions for intervention, and that useful steering signals can be recovered without full Jacobian construction.

5.2 Scaling Geometric Steering to Text-to-Image Diffusion

The controlled experiments establish that memorization geometry is both detectable and actionable, and that effective intervention directions can be recovered under low-rank approximations. We now test whether these properties scale to Stable Diffusion v1.4, where text conditioning and high dimensionality make full Jacobian construction infeasible. Using matrix-free estimates of the guidance-residual Jacobian (Appendix C.2), we evaluate GeoSteer-DM across 500 prompts with known memorized references (2,000 generations per setup).

Geometric steering scales to text-to-image diffusion. Using matrix-free estimates of local guidance geometry, GeoSteer-DM substantially reduces training-reference similarity and the fraction of generations exceeding the SSCD copying threshold, while generation quality and a modest reduction in prompt alignment (Figure 4; full metrics in Appendix F.12). This demonstrates that local geometric steering scales to large-scale, text-conditioned models without constructing full Jacobians. Representative samples in Figure 15 illustrate this trajectory-level behavior, showing that steering suppresses verbatim reproduction while preserving core prompt semantics.

GeoSteer-DM complements existing mitigation strategies. Unlike existing methods that intervene via text conditioning, attention, model predictions, or initial noise, GeoSteer-DM acts directly on local Jacobian geometry. As shown in Figure 4, adding GeoSteer-DM consistently shifts each baseline toward lower reference similarity with minimal impact on CLIP score or FID. Furthermore, these shifts are markedly smaller on non-memorized control prompts (Appendix F.12), suggesting that GeoSteer-DM selectively targets memorization-associated geometry rather than indiscriminately perturbing generative trajectories.

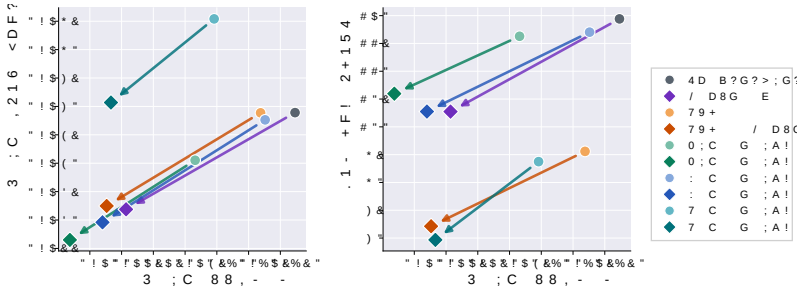


Figure 4: **GeoSteer-DM improves memorization–quality trade-offs across mitigation families.** Mean SSCD to memorized training references versus CLIP prompt alignment (left) and FID (right) on memorized prompts. Connected markers show each method, including unmodified sampling, before and after adding GeoSteer-DM.

6 Conclusion

We characterized memorization in diffusion models through the local spectral geometry of the generative field. Across training and generation, memorization manifests through a distinct spectral collapse–recovery dynamic: as models memorize during training, an intermediate-noise dominant mode emerges; during generation, entering a memorization basin corresponds to a transient spectral collapse around this leading direction. Local spectral geometry thus provides a unified framework for identifying when and how trajectories become committed to memorized outcomes. Acting on this geometric structure yields practical tools across the generative lifecycle: training-time spectral regularization, efficient detection, and inference-time trajectory steering, which scales to models as large as Stable Diffusion. While GeoSteer-DM employs a simple norm-preserving reflection along the leading singular mode, future work may leverage higher-order subspace projections or adaptive trajectory control. Overall, our findings demonstrate that local Jacobian geometry provides both a diagnostic and an actionable target for controlling memorization.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experiments, assist with software-code development. We have not used generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, or support qualitative or thematic data analysis; the remaining required-disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for text editing, plotting and visualization, literature search and summarization, identifying relevant literature, and editing and structuring the manuscript for clarity and readability. We have reviewed all AI-assisted work. AI-assisted code was reviewed and tested for correctness by the authors, experimental suggestions and interpretations were independently evaluated against the resulting experiments, and literature identified or summarized with AI assistance was verified against the original sources. All scientific claims and conclusions were reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Beatrice Achilli, Enrico Ventura, Gianluigi Silvestri, Bao Pham, Gabriel Raya, Dmitry Krotov, Carlo Lucibello, and Luca Ambrogioni. Losing dimensions: Geometric memorization in generative diffusion. *arXiv preprint arXiv:2410.08727*, 2024.

Rohan Asthana and Vasileios Belagiannis. Detecting and mitigating memorization in diffusion models through anisotropy of the log-probability. *arXiv preprint arXiv:2601.20642*, 2026.

Randall Balestriero and Richard G. Baraniuk. A spline theory of deep networks. In *International Conference on Machine Learning*, pp. 374–383. PMLR, 2018.

Ricardo Baptista, Agnimitra Dasgupta, Nikola B Kovachki, Assad Oberai, and Andrew M Stuart. Memorization and regularization in generative diffusion models. *arXiv preprint arXiv:2501.15785*, 2025.

Giulio Biroli, Tony Bonnaire, Valentin De Bortoli, and Marc Mézard. Dynamical regimes of diffusion models. *Nature Communications*, 15(1):9957, 2024.

Tony Bonnaire, Raphaël Urfin, Giulio Biroli, and Marc Mézard. Why diffusion models don't memorize: The role of implicit dynamical regularization in training. *Advances in Neural Information Processing Systems*, 38:141266–141286, 2026.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX security symposium (USENIX Security 23)*, pp. 5253–5270, 2023.

Siyi Chen, Huijie Zhang, Minzhe Guo, Yifu Lu, Peng Wang, and Qing Qu. Exploring low-dimensional subspace in diffusion models for controllable image editing. *Advances in neural information processing systems*, 37:27340–27371, 2024.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Giannis Daras, Kulin Shah, Yuval Dagan, Aravind Gollakota, Alex Dimakis, and Adam Klivans. Ambient diffusion: Learning clean distributions from corrupted data. *Advances in Neural Information Processing Systems*, 36:288–313, 2023.

Raman Dutt, Ondrej Bohdal, Pedro Sanchez, Sotirios Tsaftaris, and Timothy Hospedales. Memcontrol: Mitigating memorization in diffusion models via automated parameter selection. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 4491–4501, 2025.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in neural information processing systems*, 33:2881–2891, 2020.

Tatiana Gaintseva, Andreea-Maria Oncescu, Chengcheng Ma, Ziquan Liu, Martin Benning, Gregory Slabaugh, Jiankang Deng, and Ismail Elezi. Casteer: Cross-attention steering for controllable concept erasure. In *International Conference on Learning Representations*, volume 2026, pp. 40098–40110, 2026a.

Tatiana Gaintseva, Andrew Stepanov, Ziquan Liu, Martin Benning, Gregory Slabaugh, Jiankang Deng, and Ismail Elezi. Midsteer: Optimal affine framework for steering generative models. *arXiv preprint arXiv:2605.05220*, 2026b.

Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53 (2):217–288, 2011.

Hyeonggeun Han, Sehwan Kim, Hyungjun Joo, Sangwoo Hong, and Jungwoo Lee. Adjusting initial noise to mitigate memorization in text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 38:684–711, 2026.

Dominik Hintersdorf, Lukas Struppek, Kristian Kersting, Adam Dziedzic, and Franziska Boenisch. Finding nemo: Localizing neurons responsible for memorization in diffusion models. *Advances in Neural Information Processing Systems*, 37:88236–88278, 2024.

Ahmed Imtiaz Humayun, Ibtihel Amara, Cristina Nader Vasconcelos, Deepak Ramachandran, Candice Schumann, Junfeng He, Katherine Heller, Golnoosh Farnadi, Negar Rostamzadeh, and Mohammad Havaei. What secrets do your manifolds hold? understanding the local geometry of generative models. In *International Conference on Learning Representations*, volume 2025, pp. 89311–89336, 2025.

Michael F Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics-Simulation and Computation*, 18(3):1059–1076, 1989.

Dongjae Jeon, Dueun Kim, and Albert No. Understanding and mitigating memorization in generative models via sharpness of probability landscapes. *arXiv preprint arXiv:2412.04140*, 2024.

Yohan Jung, Hyungi Lee, Wenlong Chen, Thomas Möllenhoff, Yingzhen Li, Juho Lee, and Mohammad Emtiyaz Khan. Compact memory for continual logistic regression. *Advances in Neural Information Processing Systems*, 38:31171–31197, 2026.

Zahra Kadkhodaie, Florentin Guth, Eero Simoncelli, and Stéphane Mallat. Generalization in diffusion models arises from geometry-adaptive harmonic representations. In *International Conference on Learning Representations*, volume 2024, pp. 46543–46567, 2024.

Antoni Kowalczuk, Dominik Hintersdorf, Lukas Struppek, Kristian Kersting, Adam Dziedzic, and Franziska Boenisch. Finding dori: Discovery of retained images in diffusion models. *arXiv preprint arXiv:2507.16880*, 2025.

Chenlin Meng, Yang Song, Wenzhe Li, and Stefano Ermon. Estimating high order gradients of the data distribution by denoising. *Advances in Neural Information Processing Systems*, 34:25359–25369, 2021.

Peter Nickl, Lu Xu, Dharmesh Tailor, Thomas Möllenhoff, and Mohammad Emtiyaz Khan. The memory-perturbation equation: Understanding model's sensitivity to data. *Advances in Neural Information Processing Systems*, 36:26923–26949, 2023.

Jie Ren, Yaxin Li, Shenglai Zeng, Han Xu, Lingjuan Lyu, Yue Xing, and Jiliang Tang. Unveiling and mitigating memorization in text-to-image diffusion models through cross attention. In *European Conference on Computer Vision*, pp. 340–356. Springer, 2024.

Brendan Ross, Hamidreza Kamkari, Tongzi Wu, Rasa Hosseinzadeh, Zhaoyan Liu, George Stein, Jesse Cresswell, and Gabriel Loaiza-Ganem. A geometric framework for understanding memorization in generative models. In *International Conference on Learning Representations*, volume 2025, pp. 34773–34805, 2025.

Kulin Shah, Alkis Kalavasis, Adam R Klivans, and Giannis Daras. Does generation require memorization? creative diffusion models using ambient diffusion. *arXiv preprint arXiv:2502.21278*, 2025.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6048–6058, 2023a.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803, 2023b.

Jan Pawel Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Diffusion models encode the intrinsic dimension of data manifolds. In *Forty-first International Conference on Machine Learning*, 2024.

Zhiqing Sun and Yiming Yang. Difusco: Graph-based diffusion solvers for combinatorial optimization. *Advances in neural information processing systems*, 36:3706–3731, 2023.

Enrico Ventura, Beatrice Achilli, Gianluigi Silvestri, Carlo Lucibello, and Luca Ambrogioni. Manifolds, random matrices and spectral gaps: The geometric phases of generative diffusion. In *International Conference on Learning Representations*, volume 2025, pp. 88317–88338, 2025.

Yuxin Wen, Yuchen Liu, Chen Chen, and Lingjuan Lyu. Detecting, explaining, and mitigating memorization in diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.

David P Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1-2):1–157, 2014.

Kevin E Wu, Kevin K Yang, Rianne Van den Berg, Sarah Alamdari, James Y Zou, Alex X Lu, and Ava P Amini. Protein structure generation via folding diffusion. *Nature communications*, 15(1): 1059, 2024.

TaeHo Yoon, Joo Young Choi, Sehyun Kwon, and Ernest K Ryu. Diffusion probabilistic models generalize when they fail to memorize. In *ICML 2023 workshop on structured probabilistic inference {\&} generative modeling*, 2023.

Michael Zhang, James Lucas, Jimmy Ba, and Geoffrey E Hinton. Lookahead optimizer: k steps forward, 1 step back. *Advances in neural information processing systems*, 32, 2019.

Kairan Zhao, Eleni Triantafillou, and Peter Triantafillou. You don't need all that attention: Surgical memorization mitigation in text-to-image diffusion models. *arXiv preprint arXiv:2603.00133*, 2026.

APPENDIX OVERVIEW

Appendix A provides additional geometric derivations and diagnostics. Appendix B develops the JER formulation and detection protocol. Appendix C gives the formulation and implementation of GeoSteer-DM, and Appendix D provides the corresponding details for GeoReg-DM. Appendix E describes the experimental settings and evaluation protocols. Appendix F reports additional results, ablations, and controls.

A ADDITIONAL GEOMETRIC DERIVATIONS AND DIAGNOSTICS

A.1 ABSORPTION AND THE MEANING OF BASIN ENTRY

The definition of $\mathcal{B}_t(\delta)$ is intended to distinguish eventual memorization from *commitment* to memorization. For this interpretation to be meaningful along a reverse trajectory, basin membership should persist under the same class of admissible perturbations used to define it. Otherwise, a state could satisfy the basin condition at one timestep but subsequently leave it under an admissible continuation, making "basin entry" a transient event.

The definition in Sec. 3.1 has precisely this persistence property. Let $x_t \in \mathcal{B}_t(\delta)$, and suppose an admissible perturbation sequence w carries x_t to a later state in sampling order,

$$x_s = \Phi^w_{t\to s}(x_t), \qquad s < t.$$

Consider any admissible continuation $v \in \mathcal{U}_s(\delta)$ from x_s. Since both w and v satisfy the same per-step perturbation bound, their concatenation $w \oplus v$ is an admissible continuation from x_t. By the definition of the basin,

$$\Phi^{w\oplus v}_{t\to 0}(x_t) \in \mathcal{M}.$$

Moreover,

$$\Phi^v_{s\to 0}(x_s) = \Phi^{w\oplus v}_{t\to 0}(x_t),$$

and therefore

$$\Phi^v_{s\to 0}(x_s) \in \mathcal{M} \qquad \text{for every } v \in \mathcal{U}_s(\delta).$$

It follows directly that

$$x_s \in \mathcal{B}_s(\delta).$$

Equivalently, for every admissible continuation from t to s,

$$\Phi^w_{t\to s}\big(\mathcal{B}_t(\delta)\big) \subseteq \mathcal{B}_s(\delta).$$

Thus, once a trajectory reaches a memorization basin, no continuation within the prescribed perturbation class can return it to a state from which memorization is avoidable. This is what makes basin entry a persistent dynamical event and motivates the basin entry time $t^\dagger$ in Definition 2.

The dependence on δ is essential. The family $\{\mathcal{B}_t(\delta)\}_{\delta\geq 0}$ describes commitment at different levels of intervention strength: increasing δ enlarges the admissible set of continuations and therefore imposes a stronger requirement for basin membership. In particular, for $\delta_1 \leq \delta_2$,

$$\mathcal{B}_t(\delta_2) \subseteq \mathcal{B}_t(\delta_1).$$

A trajectory may therefore be committed to a memorized data point against small perturbations while remaining escapable under stronger interventions. We examine this δ-dependent loss of escapeability empirically in Appendix F.1.

A.2 CURVATURE-EQUIVALENT JACOBIAN FOR RECTIFIED FLOW

In Sec. 3.2, we characterize local geometry through the curvature of the time-dependent marginal density. For an ϵ-parameterized diffusion model, the input Jacobian of the model prediction is directly proportional to the score Hessian. Rectified flow requires an additional correction: its velocity field contains a linear transport component whose input Jacobian is unrelated to density curvature. We derive this correction here and show that, after removing it, the resulting operator is proportional to the same log-density Hessian used in the diffusion analysis.

Consider the standard linear rectified-flow path

$$x_s = sx_0 + (1-s)z, \qquad x_0 \sim p_{\text{data}}, \qquad z \sim \mathcal{N}(0, I), \qquad s \in [0, 1],$$

where sampling proceeds from noise at $s = 0$ to data at $s = 1$. Let p_s denote the marginal density of x_s and

$$\mathbf{s}_s(x) = \nabla_x \log p_s(x)$$

its score. The population-optimal rectified-flow velocity is

$$v^*(x, s) = \mathbb{E}[x_0 - z \mid x_s = x].$$

For this Gaussian interpolation, Tweedie's identity gives

$$\mathbb{E}[x_0 \mid x_s = x] = \frac{x}{s} + \frac{(1-s)^2}{s}\mathbf{s}_s(x).$$

Using

$$v^*(x, s) = \frac{\mathbb{E}[x_0 \mid x_s = x] - x}{1-s},$$

we obtain

$$v^*(x, s) = \frac{x}{s} + \frac{1-s}{s}\mathbf{s}_s(x),$$

and therefore

$$\mathbf{s}_s(x) = \frac{s\, v^*(x, s) - x}{1-s}.$$

Differentiating with respect to the state yields

$$H_s(x) := \nabla_x^2 \log p_s(x) = \frac{sJ_v^*(x, s) - I}{1-s},$$

where

$$J_v^*(x, s) = \frac{\partial v^*(x, s)}{\partial x}.$$

This relation identifies the RF operator corresponding to density curvature. For a learned velocity field v_θ, we define

$$J_s^{\text{RF}} = s\frac{\partial v_\theta(x, s)}{\partial x} - I. \tag{6}$$

At the population optimum,

$$J_s^{\text{RF}} = (1-s)H_s(x).$$

Thus, up to a timestep-dependent scalar, equation 6 recovers the same log-density Hessian that motivates the ϵ-Jacobian in diffusion. This is why the two model parameterizations can be analyzed using the common notation J_t in the main text.

Why the identity correction matters. The raw velocity Jacobian itself is not curvature-equivalent. Rearranging the expression above gives

$$J_v^*(x, s) = \frac{1}{s}I + \frac{1-s}{s}H_s(x).$$

Its first term is induced by the linear transport path and is present independently of the local curvature of p_s. Using J_v directly would therefore mix this known parameterization-dependent component with the geometric structure we intend to measure.

This distinction matters for the spectral quantities used throughout the paper. Stable rank, relative spectral gap, effective rank, and right-singular subspaces are invariant to multiplication by a nonzero scalar, but not to an additive identity shift. Removing the identity component is therefore necessary before comparing RF geometry with the corresponding diffusion quantities.

For this reason, we do not additionally divide equation 6 by $1 - s$. The omitted factor changes only the overall scale of the operator and leaves the spectral-shape statistics and singular subspaces used in our analysis unchanged. At a sampler state x_t with physical RF time s_t, the notation J_t in the main text therefore refers to

$$J_t = s_t\frac{\partial v_\theta(x_t, s_t)}{\partial x_t} - I. \tag{7}$$

Unless stated otherwise, RF spectral quantities are computed from this curvature-equivalent operator.

Efficient implementation. The correction does not require constructing either the full Jacobian or an explicit $d \times d$ identity matrix. For the left sketch W used by GeoReg,

$$WJ_t = s_t W \frac{\partial v_\theta(x_t, s_t)}{\partial x_t} - W.$$

The transport component can therefore be removed directly in the sketched space after computing the corresponding vector–Jacobian products. The same principle applies to stochastic Jacobian contractions used by the JER detector: for a probe z,

$$J_t^\top z = s_t \left(\frac{\partial v_\theta(x_t, s_t)}{\partial x_t} \right)^\top z - z.$$

Thus, GeoReg and JER can use the curvature-equivalent RF operator without materializing J_t. We evaluate this operator only at interior flow times $s_t \in (0, 1)$.

A.3 Signed Curvature and Jacobian Symmetry

The geometric analysis in Sec. 3.2 follows from the connection between the denoising field and the time-dependent data density. For an ϵ-parameterized diffusion model, the model Jacobian used in the main text satisfies

$$J_t = \frac{\partial \epsilon_\theta(x_t, t)}{\partial x_t} \approx -b_t H_t(x_t).$$

Its singular values therefore characterize the magnitude and organization of local curvature, but do not retain its sign.

For the signed-curvature diagnostic in this section, we orient the operator toward the Hessian and define

$$\widetilde{J}_t = -J_t = -\frac{\partial \epsilon_\theta(x_t, t)}{\partial x_t}.$$

At the population optimum, $\widetilde{J}_t \approx b_t H_t$, so it is proportional to the log-density Hessian by a positive timestep-dependent scalar. We decompose

$$\widetilde{J}_t = S_t + A_t, \qquad S_t = \frac{\widetilde{J}_t + \widetilde{J}_t^\top}{2}, \qquad A_t = \frac{\widetilde{J}_t - \widetilde{J}_t^\top}{2},$$

and measure the relative asymmetric component as

$$\rho_{\mathrm{asym}}(t) = \frac{\|A_t\|_F}{\|\widetilde{J}_t\|_F}.$$

Figure 5 shows that the learned operator remains close to symmetric. More importantly, symmetrization preserves the spectral organization underlying our analysis. Across evaluated states, the stable and effective ranks of S_t correlate with their full-Jacobian counterparts at 0.989 and 0.987, respectively, and retain the characteristic spectral collapse and subsequent recovery. Although asymmetry increases around the collapse, this increase is similar for memorized and non-memorized trajectories, with no clear between-group separation.

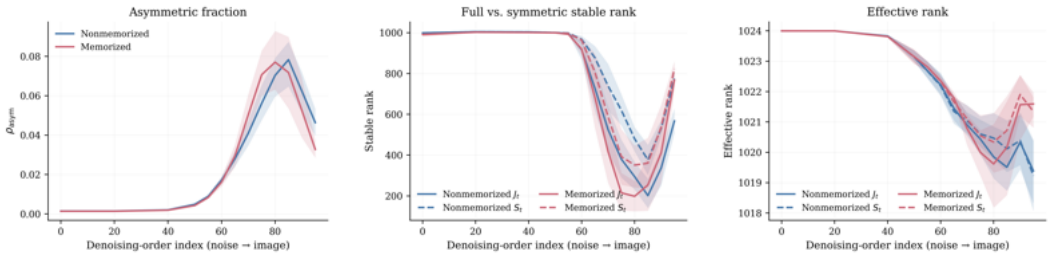
The eigenvalues of S_t additionally recover the curvature sign discarded by the singular-value representation. Across the evaluated states, the spectrum is overwhelmingly dominated by negative curvature, with negligible spectral mass in positive directions. The observed transition therefore reflects primarily a reorganization of negative-curvature magnitudes rather than a change in curvature sign. Together, these results support the curvature interpretation of the spectral organization used throughout the main analysis.

A.4 Spectral and Behavioral Diagnostics

We define the complementary spectral and behavioral quantities used in Sec. 3.2.

Our primary whole-spectrum summary is the stable rank

$$\mathrm{SR}(J_t) = \frac{\|J_t\|_F^2}{\sigma_1(t)^2} = \sum_i \left(\frac{\sigma_i(t)}{\sigma_1(t)} \right)^2.$$



(A) Asymmetric fraction **(B) Stable rank** **(C) Effective rank**

Figure 5: **Symmetry of the learned Jacobian and robustness of the spectral transition.** **(A)** Relative Frobenius norm of the antisymmetric component of the oriented curvature operator $\widetilde{J}_t = -J_t = -\partial\epsilon_\theta/\partial x_t$. **(B–C)** Stable rank and effective rank of the full operator J_t (solid) and its symmetric component S_t (dashed) along memorized and non-memorized trajectories. The learned operator remains close to symmetric, while symmetrization preserves the characteristic spectral collapse and subsequent recovery.

Low stable rank indicates concentration in a few dominant directions, while higher values indicate a more broadly distributed spectrum. Its invariance under $J_t \mapsto cJ_t$ separates this directional organization from overall curvature scale. We additionally track entropy effective rank and the relative leading spectral gap $\Delta_{12}(J_t)$, which respectively measure spectral participation and dominant-direction separation.

Complementary spectral measures. For singular values $\{\sigma_i\}$ of J_t, let $p_i = \sigma_i / \sum_j \sigma_j$. The entropy effective rank and relative leading spectral gap are

$$\nu(J_t) = \exp\left(-\sum_i p_i \log p_i\right), \qquad \Delta_{12}(J_t) = \frac{\sigma_1 - \sigma_2}{\sigma_1}.$$

They respectively measure participation across the full spectrum and separation of the dominant singular direction.

Perturbation sensitivity. At DDIM sampler step s, we perturb each trajectory state $x_s^{(i)}$ with one isotropic Gaussian perturbation calibrated to 10% of its norm,

$$\widetilde{x}_s^{(i)} = x_s^{(i)} + 0.1\,\frac{\|x_s^{(i)}\|_2}{\|z_s^{(i)}\|_2 + 10^{-12}} z_s^{(i)}, \qquad z_s^{(i)} \sim \mathcal{N}(0, I).$$

Let Φ_s denote deterministic DDIM continuation from step s to the terminal output. We define

$$S_s^{(i)} = \left\| \Phi_s(x_s^{(i)}) - \Phi_s(\widetilde{x}_s^{(i)}) \right\|_2,$$

which measures how strongly the perturbation changes the eventual generated output. We evaluate $s \in 40, 50, 60, 70, 80, 90$ using one perturbation per trajectory and report group means with standard-error bands.

Prediction-improvement rate. At sampler step s, the predicted clean sample is

$$\widehat{x}_{0,s}^{(i)} = \frac{x_s^{(i)} - \sqrt{1-\bar{\alpha}_{t_s}}\,\epsilon_\theta(x_s^{(i)}, t_s)}{\sqrt{\bar{\alpha}_{t_s}}}, \qquad d_s^{(i)} = \left\| \widehat{x}_{0,s}^{(i)} - x_{\text{final}}^{(i)} \right\|_2.$$

If s^- denotes the preceding cached sampler step, we define

$$R_s^{(i)} = \frac{d_{s^-}^{(i)} - d_s^{(i)}}{s - s^-}.$$

Predictions are cached at every sampler step in the reported experiment, so $s^- = s - 1$. Positive values indicate that the predicted clean sample has moved closer to the trajectory's eventual output. We report rates after step 50, averaged within each memorization group with standard-error bands.

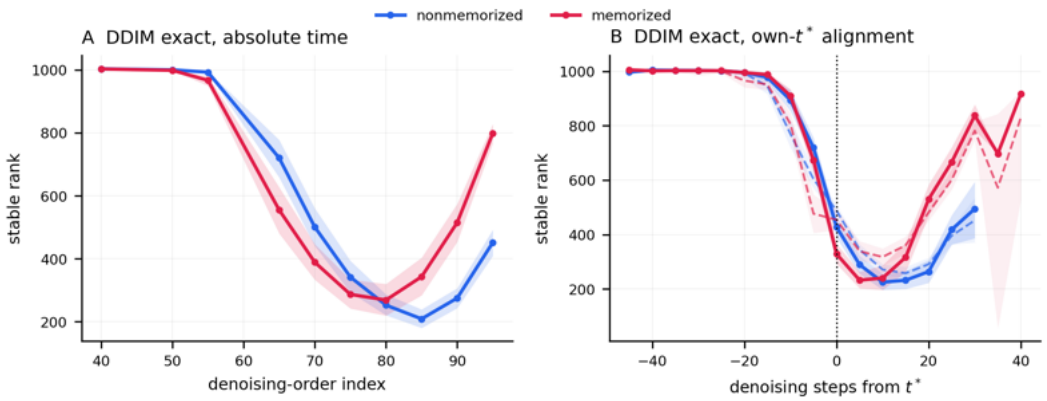


Figure 6: **Control for trajectory alignment.** **(a)** Stable-rank evolution of memorized and non-memorized DDIM trajectories in absolute sampling time, showing distinct late-stage evolution without alignment at t^*. **(b)** The same trajectories aligned at their individual t^*; dashed curves show the shuffled-anchor control obtained by randomly reassigning t^* values within each group. The transition becomes easier to resolve after alignment, while the distinct post-transition recovery is not explained by trajectory-specific alignment.

Neither behavioral diagnostic depends on the Jacobian spectrum or similarity to a training example, providing independent measurements of the change in trajectory dynamics around t^*.

Alignment control. Because t^* is estimated separately for each trajectory, alignment at this data-dependent landmark can sharpen features in the averaged spectral curves. Our analysis does not rely on this sharpness: t^* provides a trajectory-specific reference for resolving the evolution around the spectral transition. The distinct late-stage behavior of memorized and non-memorized trajectories is already visible in absolute sampling time (Fig. 6a). We additionally randomly reassign the observed t^* values across trajectories within each memorization group, preserving the distribution of alignment times while breaking their trajectory-specific correspondence (Fig. 6b).

We additionally assess alignment using 10,000 within-group permutations of t^*, preserving each group's distribution of alignment times while breaking their trajectory-specific correspondence. Using a window fixed a priori at $k = 10$ sampler steps, we measure the change in prediction-improvement rate from the k steps preceding t^* to the k steps following it, and use the difference in this change between memorized and non-memorized trajectories as the test statistic. The observed contrast is 0.0160, compared with a shuffled mean of -0.0050 (95% null interval $[-0.0208, 0.0109]$; permutation $p = 0.0043$). The observed contrast exceeds the shuffled null, showing that the change in prediction dynamics is localized to each trajectory's own t^*. We therefore use t^* as an observable spectral landmark associated with trajectory-specific prediction dynamics, while evaluating the δ-dependent loss of escapeability separately in Appendix F.1.

A.5 Posterior-Uncertainty Interpretation of Local Curvature

The curvature operator used in the main text also admits a posterior-uncertainty interpretation. For the Gaussian perturbation path

$$x_t = \alpha_t x_0 + b_t z, \qquad z \sim \mathcal{N}(0, I),$$

the second-order Tweedie identity gives

$$\mathrm{Cov}[x_0 \mid x_t] = \frac{b_t^2}{\alpha_t^2}\left(I + b_t^2 H_t(x_t)\right).$$

Thus, the eigendirections of H_t also describe how uncertainty about the underlying clean sample is distributed locally across directions. The identity is exact for the population score; replacing it by the learned score gives the corresponding model-based approximation Meng et al. (2021).

This provides a complementary interpretation of the spectral quantities used in Sec. 3.2: changes in the directional organization of curvature can equivalently be viewed as changes in the directional organization of posterior uncertainty along the denoising trajectory.

Connection to local intrinsic dimension. The posterior-covariance identity also clarifies the high-rank recovery observed for memorized trajectories. Consider a neighborhood in which the posterior concentrates on a single training example $x^{(i)}$. The corresponding conditional noise prediction is

$$\epsilon^*(x_t, t) = \frac{x_t - \sqrt{\bar{\alpha}_t}\, x^{(i)}}{b_t},$$

and therefore

$$\frac{\partial \epsilon^*(x_t, t)}{\partial x_t} = \frac{1}{b_t} I.$$

All singular values are equal, so

$$\mathrm{SR}(J_t) = d.$$

Equivalently, the posterior covariance of the underlying clean sample vanishes in every direction. Thus, although the local data distribution is effectively zero-dimensional, the noise-prediction Jacobian becomes full-rank and isotropic. The late spectral recovery observed along memorized trajectories is therefore consistent with the endpoint geometry expected near an isolated memorized example. This provides the connection between our trajectory-level spectral analysis and static local-intrinsic-dimension characterizations of memorization (Ross et al., 2025; Stanczuk et al., 2024).

B Efficient Memorization Detection

B.1 Temporal Recovery and Jacobian Effective Rank

The temporal diagnostic in Sec. 5.1 measures the joint recovery of stable rank and entropy effective rank following spectral collapse. Let J_n denote the Jacobian at monitored index n, and define

$$\mathrm{SR}(J) = \frac{\sum_i \sigma_i^2}{\sigma_1^2}, \qquad \nu(J) = \exp\left(-\sum_i p_i \log p_i\right), \qquad p_i = \frac{\sigma_i}{\sum_j \sigma_j}.$$

If

$$m_*(n) = \arg\min_{m \le n} \mathrm{SR}(J_m),$$

the corresponding relative recoveries are

$$d_{\mathrm{SR}}(n) = \left[\frac{\mathrm{SR}(J_n) - \mathrm{SR}(J_{m_*(n)})}{\mathrm{SR}(J_{m_*(n)}) + \varepsilon}\right]_+,$$

and

$$d_\nu(n) = \left[\frac{\nu(J_n) - \nu(J_{m_*(n)})}{\nu(J_{m_*(n)}) + \varepsilon}\right]_+.$$

The temporal recovery score is then

$$R_{\mathrm{sample}} = \max_{n \in \mathcal{E}} d_{\mathrm{SR}}(n) d_\nu(n). \tag{8}$$

Evaluating equation 8 requires resolving the Jacobian spectrum at multiple points along the trajectory. JER instead targets a local spectral quantity that can be accessed without an explicit spectral decomposition.

Effective-rank connection. For a nonzero matrix J, consider the Rényi-2 effective rank

$$\mathrm{ER}_2(J) = \frac{\|J\|_*^2}{\|J\|_F^2} = \frac{\left(\sum_i \sigma_i\right)^2}{\sum_i \sigma_i^2} = \frac{1}{\sum_i p_i^2}. \tag{9}$$

This quantity is bounded by the two spectral summaries entering equation 8:

$$\mathrm{SR}(J) \le \mathrm{ER}_2(J) \le \nu(J). \tag{10}$$

For the first inequality, letting $M = \sigma_1$, $S = \sum_i \sigma_i$, and $Q = \sum_i \sigma_i^2$ gives $Q \le MS$, and hence

$$\frac{Q}{M^2} \le \frac{S^2}{Q}.$$

The second follows from the ordering of Rényi entropies, $H_2(p) \le H_1(p)$, since $\mathrm{ER}_2(J) = \exp(H_2(p))$ and $\nu(J) = \exp(H_1(p))$.

Signed-trace surrogate. Computing ER_2 directly still requires the nuclear norm. For a symmetric semidefinite operator, however,

$$\|J\|_* = |\mathrm{tr}(J)|\,.$$

Motivated by the predominantly sign-coherent curvature observed in Appendix A.3, we replace the nuclear norm by an oriented trace and define

$$\mathrm{JER}^{(a)}(J) = \frac{[a\ \mathrm{tr}(J)]_+^2}{\|J\|_F^2 + \varepsilon_F}, \qquad a \in \{-1, +1\}. \tag{11}$$

Since $|\mathrm{tr}(J)| \le \|J\|_*$, the trace-based quantity provides a conservative proxy for ER_2 when the denominator is matched, becoming exact in the symmetric sign-coherent limit. The orientation a and stochastic estimates of the trace and Frobenius energy used in practice are specified in Appendix B.2.

JER should therefore be viewed as a computationally efficient local readout of the spectral dimensionality implicated by the temporal diagnostic, rather than an estimator of R_{sample} itself.

B.2 Stochastic Estimation and Implementation

The trace and Frobenius energy in equation 11 can be estimated from the same stochastic Jacobian contractions. Let $z_1, \dots, z_m \in \{-1, +1\}^d$ be independent Rademacher probes and define

$$q_\ell = J^\top z_\ell, \qquad \ell = 1, \dots, m. \tag{12}$$

We estimate

$$\widehat{\mathrm{tr}}_m(J) = \frac{1}{m}\sum_{\ell=1}^{m} z_\ell^\top q_\ell, \qquad \widehat{F}_m(J) = \frac{1}{m}\sum_{\ell=1}^{m} \|q_\ell\|_2^2. \tag{13}$$

Since $\mathbb{E}[z_\ell z_\ell^\top] = I$, these satisfy

$$\mathbb{E}\big[\widehat{\mathrm{tr}}_m(J)\big] = \mathrm{tr}(J), \qquad \mathbb{E}\Big[\widehat{F}_m(J)\Big] = \|J\|_F^2.$$

Crucially, both estimates reuse the same vector–Jacobian products, so no explicit Jacobian or spectral decomposition is required.

While $\widehat{\mathrm{tr}}_m(J)$ is unbiased for $\mathrm{tr}(J)$, squaring its positive part introduces a variance-dependent bias. We therefore treat stochastic JER as a computational proxy rather than an unbiased estimator of the deterministic quantity, and calibrate its score empirically on a disjoint calibration set.

The resulting stochastic JER score is

$$\widehat{\mathrm{JER}}_m(J) = \frac{\big[a\,\widehat{\mathrm{tr}}_m(J)\big]_+^2}{\widehat{F}_m(J) + \varepsilon_m}, \qquad \varepsilon_m = 10^{-8}\max\Big(1, \widehat{F}_m(J)\Big)\,, \tag{14}$$

where a fixes the orientation of the signed curvature. The sampler-specific operators and configurations used in our experiments are

$$(J, a, m) = \begin{cases} \left(\dfrac{\partial \epsilon_\theta}{\partial x}, +1,\ 2\right), & \text{DDIM}, \\ \left(s\dfrac{\partial v_\theta}{\partial x} - I, -1,\ 1\right), & \text{rectified flow}. \end{cases} \tag{15}$$

For rectified flow, the identity correction is the curvature-equivalent operator derived in Appendix A.2. Multiple probes are evaluated jointly in a single batched autodiff call.

Static query. JER is evaluated once after ordinary generation. From the terminal sample, we construct a deterministic low-noise query state at nominal noise time $t_0 = 0.1$ and evaluate equation 14 at that state. No additional random forward noise is introduced. The resulting score is used with the threshold rule

$$D(x) = \mathbb{1}\left\{\widehat{\mathrm{JER}}_m(J) \geq \tau_{\mathrm{JER}}\right\}, \tag{16}$$

where τ_{JER} is determined exclusively from the calibration split as described next.

B.3 Detection Protocol and Baselines

Calibration and evaluation. We evaluate detection using disjoint calibration and held-out sets for each sampler. The calibration set contains 1,000 trajectories from ten sampling seeds, with 100 trajectories per seed; evaluation uses a separate 1,000 trajectories constructed in the same way. Terminal memorization labels are defined by the nearest-neighbor criterion in Eq. 38 and are used only for calibration and evaluation.

For each detector, the operating threshold is selected on the calibration set to maximize F1 subject to

$$\mathrm{Spec}_{\mathrm{pooled}} \geq 0.90, \qquad \min_b \mathrm{Spec}_b \geq 0.85,$$

where b indexes sampling seeds. The resulting threshold is fixed before evaluation on the held-out trajectories. We report AUROC and AUPRC as threshold-independent measures, together with recall, specificity, and Matthews correlation coefficient (MCC) at the selected operating point.

Detection baselines. We compare JER with the full temporal recovery diagnostic in equation 8 and three local geometric baselines. FLIPD Ross et al. (2025) estimates local intrinsic dimensionality from the score norm and a stochastic estimate of the score-Jacobian trace. The normal-bundle estimator Stanczuk et al. (2024) constructs a local geometric representation from independently perturbed score evaluations and estimates its dimension through an SVD; in our $d = 1024$ controlled setting, we use the prescribed $4d = 4096$ perturbed evaluations. Unless stated otherwise, all baselines are implemented and evaluated following the protocols and hyperparameter settings of their respective original works.

We additionally evaluate unconditional score magnitude as a local sharpness baseline, motivated by Jeon et al. (2024) but distinct from their conditional memorization detector. Since the predictive orientation of this scalar statistic is not fixed a priori, its orientation and operating threshold are selected on the calibration set and then frozen for held-out evaluation.

Detector runtime is measured independently of sampling and includes the model and derivative evaluations required to obtain each score. Temporal recovery requires nine full-Jacobian evaluations along the trajectory, whereas JER uses a single local query with two probe directions for DDIM and one for rectified flow. Full detection and runtime results are reported in Table 2.

JER is closely related to FLIPD (Ross et al., 2025) and the normal-bundle estimator (Stanczuk et al., 2024): for $J = c\,P_N$ projecting onto a $(d-k)$-dimensional normal space, $\mathrm{tr}(J)^2/\|J\|_F^2 = d - k$, so JER recovers the codimension $d - \mathrm{LID}$. Unlike FLIPD, which converts a trace estimate into LID using the noise scale and a score-norm term, JER normalizes by Jacobian energy and is invariant to $J \mapsto cJ$, separating the organization of curvature from its magnitude. We evaluate this distinction empirically in Sec. 5.1, comparing both detection performance and computational cost.

C GeoSteer-DM: Steering Formulation and Implementation

This appendix provides the general formulation of GeoSteer-DM and its specialization to the samplers used in our experiments. The central object is the model-controlled state-space displacement introduced in Sec. 4.3. GeoSteer identifies a local Jacobian subspace at the current sampler state and modifies only the component of this displacement lying in that subspace.

Throughout, n indexes sampler steps and t_n denotes the corresponding diffusion or flow time, following the notation of the main text. The formulation below applies to both unconditional sampling and classifier-free guidance through an appropriate choice of anchor.

Algorithm 1 GEOSTEER-DM: geometric steering along a sampling trajectory

Require: Sampler $\mathcal{S}$; initial state x_{n_0}; active steering steps $\mathcal{A}$

for each remaining sampler step $n \geq n_0$ **do**
  Evaluate the ordinary model output u_n^{base} at the current state x_n.
  if $n \in \mathcal{A}$ **then**
    Form the anchor transition x_{n+1}^0 and ordinary transition x_{n+1}^{base}.
    $\delta_n \leftarrow x_{n+1}^{\text{base}} - x_{n+1}^0$.
    Compute the current Jacobian $J_{t_n} = U_{t_n} \Sigma_{t_n} V_{t_n}^\top$.
    $P_n \leftarrow v_1(t_n) v_1(t_n)^\top$.
    $x_{n+1} \leftarrow x_{n+1}^0 + (I - 2P_n)\delta_n$.
  else
    $x_{n+1} \leftarrow \mathcal{S}_{t_n}(x_n, u_n^{\text{base}})$.
  end if
end for
return the steered terminal sample.

C.1 SAMPLER AND ANCHOR FORMULATION

We write a generic sampler transition as

$$x_{n+1} = \mathcal{S}_{t_n}(x_n, u_n), \tag{17}$$

where u_n denotes the native model input used by the sampler update. At the same current state x_n, we evaluate an anchor input u_n^0 and the ordinary model input u_n^{base}, giving

$$x_{n+1}^0 = \mathcal{S}_{t_n}(x_n, u_n^0), \qquad x_{n+1}^{\text{base}} = \mathcal{S}_{t_n}(x_n, u_n^{\text{base}}).$$

Their difference

$$\delta_n = x_{n+1}^{\text{base}} - x_{n+1}^0 \tag{18}$$

is the model-controlled displacement on which GeoSteer acts. The anchor is therefore a local decomposition of a single transition; it does not define a separate sampling trajectory.

For unconditional DDIM and rectified flow, we use

$$u_n^0 = 0,$$

so that δ_n isolates the displacement induced by the model prediction. Under classifier-free guidance,

$$u_n^{\text{base}} = \epsilon_u(x_n, t_n) + s_{\text{cfg}} \left[\epsilon_c(x_n, t_n) - \epsilon_u(x_n, t_n)\right], \qquad u_n^0 = \epsilon_u(x_n, t_n),$$

such that δ_n isolates the additional displacement induced by conditional guidance while preserving the unconditional transition.

C.2 LOCAL JACOBIAN GEOMETRY

At each active steering step, GeoSteer constructs a local state-space subspace from the right singular vectors of the Jacobian associated with the current sampler state. We write

$$J_{t_n} = U_{t_n} \Sigma_{t_n} V_{t_n}^\top, \tag{19}$$

and use the leading columns of V_{t_n} as the steering directions. Because these right singular vectors lie in the sampler-state space, they can act directly on the displacement δ_n defined in equation 18.

For the controlled DDIM experiments,

$$J_{t_n}^{\text{DDIM}} = \frac{\partial \epsilon_\theta(x_n, t_n)}{\partial x_n}. \tag{20}$$

For rectified flow, we use the curvature-equivalent operator derived in Appendix A.2,

$$J_{t_n}^{\text{RF}} = s_n \frac{\partial v_\theta(x_n, s_n)}{\partial x_n} - I, \tag{21}$$

where s_n denotes the physical flow time corresponding to sampler step n. Thus, the controlled DDIM and RF experiments use the same curvature-based geometric object up to their respective model parameterizations.

At Stable Diffusion scale, we define the local geometry directly from the classifier-free guidance residual. Let

$$g(z_n, t_n) = \epsilon_c(z_n, t_n) - \epsilon_u(z_n, t_n),$$

where ϵ_c and ϵ_u denote the conditional and unconditional noise predictions, respectively. We use the guidance-residual Jacobian

$$J_{t_n}^{\mathrm{SD}} = J_g(z_n, t_n) = \frac{\partial g(z_n, t_n)}{\partial z_n} = \frac{\partial \left[\epsilon_c(z_n, t_n) - \epsilon_u(z_n, t_n)\right]}{\partial z_n}. \tag{22}$$

Its leading right singular vector identifies the latent direction along which the conditional–unconditional guidance displacement changes most strongly. This aligns the geometric operator with the quantity that GeoSteer modifies under classifier-free guidance.

The operator also admits an approximate conditional-curvature interpretation. Under the usual noise-prediction/score correspondence,

$$\epsilon(z_n, t_n) \approx -\sigma_{t_n} \nabla_{z_n} \log p_{t_n}(z_n),$$

the guidance residual satisfies approximately

$$g(z_n, t_n) \approx -\sigma_{t_n} \left[\nabla_{z_n} \log p_{t_n}(z_n \mid c) - \nabla_{z_n} \log p_{t_n}(z_n)\right] = -\sigma_{t_n} \nabla_{z_n} \log p_{t_n}(c \mid z_n),$$

and hence

$$J_g(z_n, t_n) \approx -\sigma_{t_n} \nabla_{z_n}^2 \log p_{t_n}(c \mid z_n).$$

We use this relation only as an interpretation: the learned conditional and unconditional predictors are approximate, and J_g need not be exactly symmetric or conservative. GeoSteer therefore operates on the leading right singular direction of the guidance-residual Jacobian rather than assuming an exact Hessian structure.

C.3 Subspace Steering Operator

Let

$$V_{n,K} = [v_1(t_n), \ldots, v_K(t_n)]$$

contain the leading K right singular vectors of J_{t_n}, and define the corresponding projector

$$P_n = V_{n,K} V_{n,K}^\top.$$

For a sketched Jacobian, $V_{n,K}$ is obtained from the right singular vectors of the sketch and still lies in the original sampler-state space.

We decompose the model-controlled displacement as

$$\delta_n = P_n \delta_n + (I - P_n)\delta_n.$$

GeoSteer applies separate gains to the selected geometric subspace and its orthogonal complement. For subspace gain ρ_n and global displacement gain $a_n \geq 0$, we define

$$\delta_n^{\mathrm{geo}} = a_n \left[(I - P_n)\delta_n + \rho_n P_n \delta_n\right] = a_n \left[I + (\rho_n - 1)P_n\right] \delta_n, \qquad x_{n+1}^{\mathrm{geo}} = x_{n+1}^0 + \delta_n^{\mathrm{geo}}. \tag{23}$$

The canonical GeoSteer operator used in the controlled experiments takes $K = 1$, $a_n = 1$, and $\rho_n = -1$, yielding

$$\delta_n^{\mathrm{geo}} = (I - 2P_n)\delta_n, \qquad P_n = v_1(t_n) v_1(t_n)^\top. \tag{24}$$

Thus, the component aligned with the dominant local geometric direction is reversed while its orthogonal complement is preserved.

The more general form in equation 23 also contains suppression ($\rho_n = 0$), partial reversal ($-1 < \rho_n < 0$), and global attenuation ($a_n < 1$) as special cases. These variants are useful for describing the Stable Diffusion implementation below, but the canonical controlled-model intervention is the norm-preserving reflection in equation 24.

Operator properties. Because P_n depends only on the selected subspace, the intervention is invariant to sign flips and orthogonal changes of basis within that subspace. Moreover, since $P_n\delta_n \perp (I - P_n)\delta_n$,

$$\|\delta_n^{\text{geo}}\|_2^2 = a_n^2 \left[\|(I - P_n)\delta_n\|_2^2 + \rho_n^2 \|P_n\delta_n\|_2^2 \right]. \tag{25}$$

For the canonical choice $a_n = 1$ and $\rho_n = -1$,

$$\|\delta_n^{\text{geo}}\|_2 = \|\delta_n\|_2 .$$

The intervention therefore changes the direction of the model-controlled displacement without changing its magnitude.

For this canonical case, the transformation

$$H_n = I - 2P_n$$

is a Householder reflection satisfying

$$H_n^\top = H_n, \qquad H_n^2 = I, \qquad H_n^\top H_n = I.$$

It reverses the selected component exactly while leaving the orthogonal component unchanged. Reflection provides a simple norm-preserving intervention for testing whether the identified geometric direction is actionable, while more sophisticated control rules could further exploit this structure.

C.4 Sampler-Specific Forms

The state-space formulation in equation 23 is independent of how the generative model parameterizes its sampler update. For the deterministic samplers used in our experiments, the same transformation can also be written directly in terms of the corresponding model-controlled quantity.

DDIM. Let $s < t$ denote consecutive diffusion times. A deterministic ϵ-parameterized DDIM transition can be written as

$$\mathcal{S}_{t\to s}(x_t, \epsilon) = A_{t\to s} x_t + B_{t\to s}\epsilon,$$

where

$$A_{t\to s} = \sqrt{\frac{\bar\alpha_s}{\bar\alpha_t}}, \qquad B_{t\to s} = \sqrt{1-\bar\alpha_s} - \sqrt{\frac{\bar\alpha_s}{\bar\alpha_t}}\sqrt{1-\bar\alpha_t}.$$

With the zero-prediction anchor,

$$\delta_t = B_{t\to s}\epsilon_\theta(x_t, t).$$

Since $B_{t\to s}$ is scalar, applying equation 23 to δ_t is equivalent to transforming the noise prediction as

$$\epsilon_\theta^{\text{geo}}(x_t, t) = a_t \left[I + (\rho_t - 1)P_t \right] \epsilon_\theta(x_t, t). \tag{26}$$

For the canonical reflection, this reduces to $\epsilon_\theta^{\text{geo}} = (I - 2P_t)\epsilon_\theta$.

Rectified flow. For explicit Euler sampling at physical flow time s_n,

$$x_{n+1} = x_n + \Delta s_n \, v_\theta(x_n, s_n).$$

The zero-velocity anchor gives

$$x_{n+1}^0 = x_n, \qquad \delta_n = \Delta s_n \, v_\theta(x_n, s_n).$$

The corresponding velocity-space form is therefore

$$v_\theta^{\text{geo}}(x_n, s_n) = a_n \left[I + (\rho_n - 1)P_n \right] v_\theta(x_n, s_n). \tag{27}$$

Here, P_n is constructed from the curvature-equivalent RF Jacobian in equation 21. Thus, the sampler displacement is transformed in the same way as for DDIM, while the geometry defining the steering subspace accounts for the linear transport component of the rectified-flow parameterization.

Classifier-free guidance. Under classifier-free guidance, let

$$g_n = \epsilon_c(x_n, t_n) - \epsilon_u(x_n, t_n), \qquad u_n^{\text{base}} = \epsilon_u(x_n, t_n) + s_{\text{cfg}} g_n.$$

Using the unconditional prediction as the anchor isolates the guidance-induced displacement. For deterministic DDIM, the GeoSteer transformation is therefore equivalent to

$$g_n^{\text{geo}} = a_n \left[I + (\rho_n - 1)P_n\right] g_n, \qquad u_n^{\text{geo}} = \epsilon_u(x_n, t_n) + s_{\text{cfg}} g_n^{\text{geo}}. \tag{28}$$

The unconditional component is preserved, while GeoSteer acts only on the additional displacement introduced by conditional guidance.

C.5 Experimental Configuration

Controlled experiments. For CelebA32, GeoSteer uses the full 1024×1024 Jacobian and exact SVD at each active steering step. Both DDIM and rectified flow use the rank-one reflection in equation 24, with $K = 1$, $a_n = 1$, and $\rho_n = -1$. The steering geometry is recomputed at the current edited state after every modified transition, so subsequent interventions follow the trajectory produced by GeoSteer rather than the original unmodified path. We evaluate the intervention at training checkpoints 30k, 60k, 100k, 160k, 200k, and 300k using 500 paired generations per checkpoint and sampler. The baseline and steered samples in each pair share the same initial noise. Sampler-specific rewind and active steering windows are fixed across checkpoints and are reported with the complete evaluation protocol in Appendix E.

Stable Diffusion. For Stable Diffusion v1.4, GeoSteer uses the guidance-residual Jacobian J_g defined in Eq. 22. Because explicitly constructing

$$J_g = \frac{\partial(\epsilon_c - \epsilon_u)}{\partial z}$$

in latent space is impractical, we estimate its leading right singular direction matrix-free using power iteration on $J_g^\top J_g$.

For a unit direction v, each iteration first approximates the Jacobian–vector product using a central finite difference,

$$J_g v \approx \frac{g(z_n + hv, t_n) - g(z_n - hv, t_n)}{2h}, \qquad h = 0.3, \tag{29}$$

and then evaluates

$$J_g^\top (J_g v)$$

with an autograd vector–Jacobian product. Normalizing the result gives the next power-iteration direction. We use five iterations at the first active steering step and two warm-started iterations at subsequent steps, initializing each estimate from the direction obtained at the preceding active step. The computation is performed in FP32 with TF32 disabled.

The resulting unit direction $v_1(t_n)$ defines $P_n = v_1(t_n)v_1(t_n)^\top$ and is recomputed at the current, already-steered latent at every active step. Under classifier-free guidance, GeoSteer acts on the guidance residual according to Eq. 28. We use $K = 1$, $\rho_n = -1$, and $a_n = 0.7$, giving

$$g_n^{\text{geo}} = 0.7\,(I - 2P_n)\,g_n, \qquad u_n^{\text{geo}} = \epsilon_u(z_n, t_n) + 7.5\,g_n^{\text{geo}}.$$

Steering is applied over denoising-order indices 0–17 of deterministic DDIM-50 sampling.

The Stable Diffusion experiment applies GeoSteer over this prescribed steering window without JER gating or rewind, allowing the steering operator to be evaluated independently at this scale. Dataset construction, sampling parameters, and evaluation metrics are given in Appendix E.

D GeoReg-DM: Objective and Implementation

This appendix provides implementation details for the training-time regularizer introduced in Sec. 4.1. GeoReg selectively regularizes the local Jacobian within each minibatch using an energy-weighted spectral objective and a low-dimensional differentiable sketch to keep the additional computation tractable.

D.1 Training Objective and Selection

For each training example, we sample the model-specific training time and construct the corresponding noisy or interpolated state x_t. Let $\ell_{\text{base}}^{(i)}$ denote the per-sample diffusion or rectified-flow training loss and

$$\mathcal{L}_{\text{base}} = \frac{1}{B}\sum_{i=1}^{B} \ell_{\text{base}}^{(i)}$$

the ordinary minibatch objective.

GeoReg does not evaluate the geometric penalty on every training state. Instead, we prioritize examples using the input sensitivity of the base loss. For the mean-reduced diffusion objective, the original per-example half-squared-error priority is recovered directly from the input gradient obtained during the ordinary base backward,

$$\pi^{(i)} = \frac{Bd}{2}\left\|\nabla_{x_{t_i}^{(i)}}\mathcal{L}_{\text{base}}\right\|_2 .$$

The common scale does not affect the resulting ranking. Among timestep-eligible examples, we select the m highest-priority states for geometric regularization. This reuses the base-loss backward and avoids an additional network pass for priority computation.

For each $i \in \mathcal{I}$, GeoReg penalizes the squared leading singular value of the sketched Jacobian,

$$\Phi_{\text{GeoReg}}\left(\widehat{J}^{(i)}\right) = \sigma_1\left(\widehat{J}^{(i)}\right)^2,$$

that is, the squared gain of the dominant direction of the sketched Jacobian. Since $\widehat{J}^{(i)} \in \mathbb{R}^{k\times d}$, this quantity is obtained directly as the largest eigenvalue of the small Gram matrix

$$G^{(i)} = \widehat{J}^{(i)}\widehat{J}^{(i)\top} \in \mathbb{R}^{k\times k}.$$

The geometric loss is averaged over the selected states,

$$\mathcal{L}_{\text{geo}} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \Phi_{\text{GeoReg}}\left(\widehat{J}^{(i)}\right), \tag{30}$$

and the complete training objective is

$$\mathcal{L} = \mathcal{L}_{\text{base}} + \lambda_{\text{geo}}\mathcal{L}_{\text{geo}}. \tag{31}$$

Selection depends only on quantities available at the current training update. In particular, terminal nearest-neighbor distances and memorization labels are never used by either the objective or the selection rule.

D.2 Differentiable Jacobian Sketch

Computing and differentiating through the full input Jacobian at every regularized training state would be prohibitively expensive. GeoReg instead uses a low-dimensional left sketch. Let $W \in \mathbb{R}^{k\times d}$ have orthonormal rows and define

$$\widehat{J} = WJ \in \mathbb{R}^{k\times d}. \tag{32}$$

The geometric objective is evaluated on $\widehat{J}$ rather than on the full $d \times d$ operator.

For a uniformly random row-orthonormal sketch,

$$\mathbb{E}_W\left[\|WJ\|_F^2\right] = \frac{k}{d}\|J\|_F^2. \tag{33}$$

The sketch therefore provides a low-dimensional view of the local Jacobian while reducing the spectral computation to the $k \times k$ Gram matrix $\widehat{J}\widehat{J}^\top$. The objective is defined directly on the sketched operator; we do not interpret its spectral quantities as exact estimates of those of the full Jacobian.

Algorithm 2 GEOREG-DM: geometry-regularized training step

Require: Minibatch of B training examples; fixed $W \in \mathbb{R}^{k\times d}$; target count m; λ_{geo}

1: Sample training times t_i and construct corrupted states $x_{t_i}^{(i)}$ and per-example losses $\ell_{\text{base}}^{(i)}$.
2: $\mathcal{L}_{\text{base}} \leftarrow \dfrac{1}{B}\displaystyle\sum_{i=1}^{B} \ell_{\text{base}}^{(i)}$.
3: Backpropagate $\mathcal{L}_{\text{base}}$, accumulating parameter gradients and retaining $\nabla_{x_{t_i}^{(i)}} \mathcal{L}_{\text{base}}$.
4: $\pi^{(i)} \leftarrow \dfrac{Bd}{2}\left\|\nabla_{x_{t_i}^{(i)}} \mathcal{L}_{\text{base}}\right\|_2$.
5: $\mathcal{I} \leftarrow \text{TopM}\big(\{i : \text{Eligible}(t_i)\}, \pi, m\big)$.
6: Detach the selected realized states $\{x_{t_i}^{(i)} : i \in \mathcal{I}\}$ and re-enable input gradients.
7: Construct all $m \times k$ state–projection pairs jointly and compute the packed Jacobian sketches

$$\widehat{J}_{t_i}^{(i)} = W J_{t_i}^{(i)}, \qquad i \in \mathcal{I}.$$

8: **for** each $i \in \mathcal{I}$ **do**
9: $\quad G^{(i)} \leftarrow \widehat{J}_{t_i}^{(i)} \widehat{J}_{t_i}^{(i)\top}$.
10: $\quad \Phi^{(i)} \leftarrow \lambda_{\max}\Big(G^{(i)}\Big)$.
11: **end for**
12: $\mathcal{L}_{\text{geo}} \leftarrow \dfrac{1}{|\mathcal{I}|}\displaystyle\sum_{i\in\mathcal{I}} \Phi^{(i)}$.
13: Backpropagate $\lambda_{\text{geo}}\mathcal{L}_{\text{geo}}$ and accumulate its parameter gradients with those from $\mathcal{L}_{\text{base}}$.
14: $\theta \leftarrow \text{OptimizerStep}(\theta)$.

Differentiable construction. The sketch can be obtained without explicitly materializing J. For a selected state $x_t^{(i)}$, we detach the realized state and re-enable gradients,

$$\widehat{x}^{(i)} = \text{detach}(x_t^{(i)}),$$

and let $w_a^\top$ denote row a of W. Each row of the sketch is computed as

$$\widehat{J}_{a,:}^{(i)} = \nabla_{\widehat{x}^{(i)}} \left\langle w_a, f_\theta(\widehat{x}^{(i)}, t_i) \right\rangle, \qquad a = 1, \ldots, k, \tag{34}$$

where f_θ denotes the model prediction associated with the training objective: the noise prediction for diffusion and the velocity field for rectified flow.

These vector–Jacobian products are evaluated with graph construction enabled, so differentiating the geometric objective with respect to θ introduces the mixed derivatives

$$\frac{\partial^2 f_\theta}{\partial\theta\,\partial x}.$$

Detaching the realized state prevents this gradient from propagating through the forward corruption or interpolation process.

For the selected states, the $m \times k$ state–projection pairs are evaluated jointly in a packed autodiff computation and reshaped into one $k \times d$ Jacobian sketch per state. The geometric objective is then obtained from the largest eigenvalue of each $k \times k$ Gram matrix $\widehat{J}\widehat{J}^\top$.

D.3 CELEBA TRAINING AND CHECKPOINT STABILIZATION

For the CelebA experiments, we train an unconditional noise-predicting diffusion model on the 4,096-image CelebA32 subset described in Appendix E.3. GeoReg uses a fixed row-orthonormal sketch $W \in \mathbb{R}^{8\times d}$. Among training states with raw diffusion timestep $t \in \{90, \ldots, 500\}$, we select the $m = 32$ states with highest input-loss-gradient priority and construct their Jacobian sketches using the packed left-VJP implementation described above. The geometric objective is the mean over the selected states, with no additional minibatch-size rescaling. The 8×8 Gram matrices are diagonalized exactly, and the geometry-dependent computation is performed in FP32.

Table 3: **CelebA GeoReg training and stabilization settings.**

Setting	Value
Batch size	512
Adam learning rate	10^{-4}
Geometry timestep window	$t \in \{90, \dots, 500\}$
Selected states per update	$m = 32$
Sketch rank	$k = 8$
Initial interpolation	$\alpha_0 = 0.25$
Synchronization interval	$H = 500$ updates
Periodic interpolation	$\alpha_s = 0.5$
Geometry computation	FP32

Periodic checkpoint stabilization. To stabilize continued optimization under the geometric objective, we maintain fast and slow parameter copies initialized from a warm-up checkpoint.

The fast parameters are optimized using the GeoReg objective. After every $H = 500$ updates, the two copies are synchronized according to

$$\theta_s \leftarrow (1 - \alpha_s)\theta_s + \alpha_s \theta_f, \qquad \theta_f \leftarrow \theta_s, \qquad \alpha_s = 0.5. \tag{35}$$

The slow parameters continue to evolve after every synchronization and are used for evaluation at the reported major checkpoints. Optimizer moments and counters of the fast optimizer are retained across synchronization.

This procedure is related to the fast–slow parameter interpolation used by Lookahead (Zhang et al., 2019), but here it is used specifically to stabilize the GeoReg continuation. The initial coefficient α_0 and periodic coefficient α_s have distinct roles: α_0 controls the contribution of the regularized branch to the initial checkpoint, whereas α_s controls how accumulated fast updates are incorporated into the evolving slow parameters. The stabilization procedure is separate from the GeoReg objective itself and uses one actively optimized model together with a stored slow parameter copy. The stabilization was introduced to prevent rapid degradation in generation quality during continued regularization.

D.4 Computational cost

Table 4 reports the cost of the optimized GeoReg implementation. Under the same benchmark, an ordinary DDIM training update takes 272.36 ms, compared with 1045.89 ms for GeoReg, corresponding to a $3.84\times$ training-time cost. Peak allocated memory increases from 16.95 to 19.63 GiB.

Nearly all of the additional computation comes from constructing and differentiating through the Jacobian sketch. The packed sketch construction takes 237.09 ms and the geometry backward 519.54 ms, whereas the exact 8×8 Gram eigensolve and spectral objective require only 1.02 ms. Thus, the spectral computation itself is negligible relative to the Jacobian-dependent differentiation required by GeoReg.

E Experimental Details

E.1 GeoReg Training Baselines

We compare GeoReg with several training-time baselines chosen to distinguish its spectral intervention from simpler mechanisms that can reduce memorization. Unless stated otherwise, all methods use the same model, training schedule, minibatches, and evaluation protocol as GeoReg.

Standard training. The unregularized baseline continues optimization using only $\mathcal{L}_{\text{base}}$.

Tikhonov score regularization. Following the score regularization studied by Baptista et al. (2025), we penalize the squared magnitude of the estimated score. In the noise-prediction parameterization this gives

$$\Phi_{\text{Tik}} = \frac{\|\epsilon_\theta(x_t, t)\|_2^2}{d\, b_t^2}, \tag{36}$$

Table 4: **Computational cost of GEOREG-DM.** CUDA-synchronized median breakdown over ten warmed DDIM training updates on one NVIDIA A100-SXM4-40GB, using $B = 512$, $m = 32$ selected states, and a fixed rank-8 left Jacobian sketch.

Component	Time (ms)	Fraction
Batch setup, base forward + loss	88.66	8.48%
Base backward	176.11	16.84%
State selection + geometry-mode entry	1.98	0.19%
Packed left-Jacobian sketch	237.09	22.67%
Exact Gram eigensolve + spectral objective	1.02	0.10%
Geometry backward	519.54	49.67%
Adam optimizer step	16.74	1.60%
Scalar/host closure and residual	4.75	0.45%
Total	**1045.89**	**100.00%**

where b_t denotes the diffusion noise scale. This baseline regularizes the magnitude of the learned score without explicitly constraining its local derivatives.

Reduced learning rate. As an optimization control, we halve the Adam learning rate after `10k` training checkpoint. All methods share the same training history with learning rate 10^{-4} through `10k`; the reduced-LR baseline then continues ordinary diffusion training using only the base noise-prediction objective at 5×10^{-5} through `60k`. This tests whether reduced memorization can be explained by slower continued optimization alone.

Ambient-style noisy training. We additionally consider a fixed-noisy-bank baseline inspired by Ambient Diffusion Daras et al. (2023). At the training fork, each training example is corrupted once at diffusion index 50, producing a fixed noisy training set used throughout the continuation. For timesteps up to this corruption level, training uses the ordinary noise-prediction objective; at larger timesteps, we use the corresponding noise-aware objective without restoring clean supervision. This changes the information supplied during training rather than directly regularizing the learned score or its Jacobian. Because the construction is adapted to our uniform-time noise-prediction setting, we refer to it as an Ambient-style baseline rather than a reproduction of the original method.

Regularization strengths. For the synthetic experiments, regularization coefficients are calibrated at the common training fork so that the initial auxiliary-loss parameter-gradient norm is approximately 10% of the base-loss gradient norm. The reduced-learning-rate and Ambient-style controls do not use this calibration because they modify the optimization or training-data procedure rather than adding a regularization term.

E.2 SYNTHETIC DATA EXPERIMENTS

We evaluate GeoReg on three synthetic populations of increasing geometric complexity. In each setting, the training set is sampled once and then held fixed throughout optimization; only diffusion noise and timesteps are resampled during training. Evaluation uses independent samples from the known population distribution. The model is trained only on the fixed training coordinates and is not provided with component identities, latent coordinates, population parameters, or other information used to construct the datasets.

Gaussian80. The training set consists of $N = 80$ independent samples from a two-dimensional standard Gaussian,

$$x \sim \mathcal{N}(0, I_2).$$

No standardization is applied. This provides a full-dimensional unimodal setting in which attraction to individual training coordinates can be separated from multimodal or manifold structure.

Mixture32. We construct an eight-component Gaussian mixture in two dimensions with uniformly weighted component means

$$\mu_j = 1.8\left(\cos\frac{2\pi j}{8}, \sin\frac{2\pi j}{8}\right), \qquad j = 0, \ldots, 7,$$

and covariance $0.2^2 I_2$ within each component. The fixed training set contains four independently sampled points from each component, giving $N = 32$. Population samples instead draw the component identity uniformly before sampling from the corresponding Gaussian. The stratified training set allows us to examine copying together with mode coverage, component balance, and within-component variation.

SwissRoll8D128. The final synthetic setting is a noisy two-dimensional curved surface embedded in $\mathbb{R}^8$. We sample

$$u \sim \mathrm{Uniform}(1.5\pi, 4.5\pi), \qquad v \sim \mathrm{Uniform}(-10, 10),$$

and define

$$y = \left(\frac{u\cos u}{10}, \frac{v}{10}, \frac{u\sin u}{10}, 0, \ldots, 0\right).$$

Samples are then generated as

$$x = Qy + 0.03z, \qquad z \sim \mathcal{N}(0, I_8),$$

where Q is a fixed orthogonal 8×8 matrix generated with seed 601. The training set contains $N = 128$ samples. We do not whiten the thin directions, allowing the evaluation to distinguish variation along the surface from variation away from it.

Training and sampling. All synthetic experiments use a width-128 noise-prediction MLP trained with Adam at learning rate 10^{-3} and batch size 512 for `30k` updates, with methods sharing the same initial `5k`-update unregularized prefix. The low dimensionality permits exact input-Jacobian computation. Geometric regularization is applied to at most 32 eligible states per minibatch at diffusion timesteps $t \in \{5, \ldots, 100\}$, using the selection procedure specified for the corresponding experiment. Generation uses deterministic DDIM with 100 sampling steps.

For each dataset and data seed, we fix 2,000 validation population samples, 10,000 final population samples, and a separate 10,000-sample population-null set. The latter provides a population baseline for nearest-training-point statistics: population samples can naturally lie near the finite training set.

Evaluation. We distinguish copying from population-level distributional agreement. Copying is measured relative to the fixed finite training set. Sliced Wasserstein distance (SWD), distributional precision, and distributional recall are instead computed against independent samples from the underlying population. SWD measures agreement between the generated and population distributions through one-dimensional projections, while precision measures the fraction of generated samples lying within the estimated support of the population and recall measures the fraction of the population support covered by the generated distribution.

These quantities are complementary. A model can produce individually plausible samples while covering only part of the population, resulting in high precision but lower recall. Conversely, increased coverage can be accompanied by additional mass outside the population support. SWD is sensitive to broader distributional discrepancies, including changes in relative mass across mixture components. None of these population metrics measures copying directly, which is why we report them together with nearest-training-point statistics.

E.3 Controlled CelebA Setting

We use a controlled CelebA setting in which full input Jacobians can be computed during sampling. The training set contains 4,096 images resized and center-cropped to 32×32, converted to grayscale, and scaled to $[0, 1]$. We subtract the mean image intensity computed over this subset and apply no variance normalization.

Both generative models use the same time-conditioned U-Net with base width 64, channel multipliers $(1, 2, 4)$, and two residual blocks per resolution. Attention is applied at 16×16, 8×8, and the bottleneck. The network uses four-head attention, GroupNorm, SiLU activations, dropout 0.1, and sinusoidal time embeddings.

Diffusion. The diffusion model is trained with the standard noise-prediction objective

$$\mathcal{L}_{\mathrm{DM}} = \mathbb{E}_{x_0,\epsilon,t}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t\right)\right\|_2^2\right],$$

using $T = 1000$ noise levels and a linear schedule $\beta_t \in [10^{-4}, 2 \times 10^{-2}]$. Training timesteps are sampled uniformly. Generation uses deterministic DDIM ($\eta = 0$) with 100 linearly spaced schedule points from timestep 999 to 0, corresponding to 99 transitions.

Rectified flow. The rectified-flow model is trained on the linear interpolation

$$x_t = (1-t)x_{\mathrm{noise}} + tx_{\mathrm{data}}, \qquad v^\star = x_{\mathrm{data}} - x_{\mathrm{noise}},$$

with uniformly sampled discretized times over $[0, 1]$ and velocity-matching objective

$$\mathcal{L}_{\mathrm{RF}} = \mathbb{E}\left[\|v_\theta(x_t, t) - v^\star\|_2^2\right].$$

Sampling uses explicit Euler integration on 100 uniformly spaced points from $t = 0$ to 1, again giving 99 numerical updates.

The remaining regularizer-specific settings are given in Appendix D.3. Although training runs permit a longer optimization horizon, the results in the main text use checkpoints through 300k updates.

E.4 Memorization and Generation-Quality Metrics

For the controlled models, memorization is measured directly in the mean-centered grayscale pixel space used for training. Given a generated sample y, let $d_1(y)$ and $d_2(y)$ denote its smallest and second-smallest Euclidean distances to the 4,096 training examples. We define the nearest-neighbor gap ratio

$$r(y) = \frac{d_1(y)}{d_2(y) + 10^{-10}}, \tag{37}$$

and classify a generation as memorized when

$$r(y) < \frac{1}{3}. \tag{38}$$

We report memorization rate together with the corresponding nearest-neighbor gap statistics.

Generation quality is evaluated using FID with 2,048-dimensional Inception-v3 pool-3 features. For GeoReg, FID is computed from 50,000 generations against the full 4,096-image CelebA training set; for GeoSteer, it is computed from 500 generations per arm and checkpoint against a fixed 1,024-image CelebA reference set. We treat FID as a secondary metric in this controlled setting because direct reproduction of images in the finite reference set can itself improve FID.

E.5 Controlled-Model Evaluation

GeoReg. We evaluate all training methods using the same ordinary deterministic DDIM sampling procedure. We report results through `60k` updates, by which point the training objective has largely saturated, and do not continue these runs into the deliberately overtrained regime used separately to study memorization dynamics and GeoSteer. Training configurations and computational-cost measurements are given in Appendix D.4.

Memorization detection. JER is evaluated at the 160k checkpoint for both generative dynamics in the main body. We report additional checkpoints in the appendix. Calibration uses 1,000 trajectories per sampler, obtained from ten sampling seeds with 100 generations each. A disjoint set of another 1,000 trajectories per sampler is reserved for evaluation. Threshold selection and baseline definitions are given in Appendix B.3. Throughout our analyses, memorization rates are estimated from the generation cohorts used by each experiment; the differing evaluation sizes therefore produce small variations in the reported control rates across sections.

GeoSteer. We evaluate steering at 30k, 60k, 100k, 160k, 200k, 300k updates. Each sampler–checkpoint pair contains 500 paired baseline and steered generations, obtained from five sampling seeds with 100 generations each. Within each pair, both trajectories begin from the same Gaussian initial state.

For DDIM, steering resumes from sampler state 60 and is applied to transitions 60–98. For rectified flow, it resumes from state 20 and is applied to Euler updates 20–89. The geometric operator and sampler-specific formulation are given in Appendix C.

E.6 Stable Diffusion Evaluation

We evaluate text-to-image mitigation using Stable Diffusion v1.4 at 512×512 resolution. Generation uses deterministic DDIM with 50 steps, $\eta = 0$, and classifier-free guidance scale 7.5.

Evaluation cohorts. The memorized cohort contains 500 prompts with known memorized reference images from the Wen et al. benchmark. We generate four paired deterministic samples per prompt, giving 2,000 generations per method. To evaluate behavior away from known memorization, we additionally use 100 prompts labeled as non-memorized in the corresponding ground-truth collection and generate one paired sample per prompt and method.

Metrics. Training-image similarity is measured using SSCD cosine similarity. For the memorized cohort, each generation is compared against the complete bank of 500 memorized references and assigned its maximum similarity; we additionally report the fraction satisfying

$$\mathrm{SSCD} \geq 0.50.$$

For the non-memorized control cohort, nearest-reference similarity is computed against the common Webster/LAION reference bank.

Prompt fidelity is measured as the cosine similarity between normalized CLIP ViT-B/32 prompt and image embeddings. We additionally report FID using Inception-v3 pool-3 features against a fixed reference set of 10,000 LAION-derived images center-cropped and resized to 512×512.

Baselines. We compare GeoSteer with the unmodified sampler and four inference-time mitigation methods: Ren et al. (Ren et al., 2024), RTA (Somepalli et al., 2023b), Wen et al. (Wen et al., 2024), and Han et al. (Han et al., 2026). We additionally compose each method with GeoSteer to evaluate whether geometric steering provides complementary mitigation. All methods use the configurations reported for the corresponding experiments. To isolate the effect of the interventions, all methods are applied to every trajectory independently of detection, and each baseline–GeoSteer pair shares the same initial noise.

F Additional Results and Ablations

F.1 Adversarial Evidence for Memorization Basin Entry

We test the basin interpretation directly by optimizing bounded perturbations to the remaining DDIM trajectory around the geometry-defined transition t^*. Starting from offset $\tau = t - t^*$, we introduce a control u_k before each remaining sampler transition with

$$\|u_k\|_2 \leq \delta, \qquad \delta_{\mathrm{RMS}} = \frac{\delta}{\sqrt{d}}.$$

We evaluate $\delta_{\mathrm{RMS}} \in \{0.001, 0.003, 0.005, 0.007\}$ and $\tau \in \{-10, -5, 0, 5, 10\}$ on 5000 held-out memorized trajectories. An intervention escapes when the terminal sample satisfies $r(x_0^{\mathrm{adv}}) \geq 1/3$.

We compare two strategies for finding an escaping continuation: end-to-end projected gradient ascent over the remaining controls and a greedy state-by-state optimization of the same nearest-neighbor gap objective. Matched random perturbations provide a magnitude control.

Figure 7 shows a sharp loss of escapeability around t^*. At $\delta_{\mathrm{RMS}} = 0.003$, optimized escape probability falls from 0.96 at $\tau = -10$ to 0.54 at t^* and 0.07 at $\tau = 10$. Increasing the perturbation bound

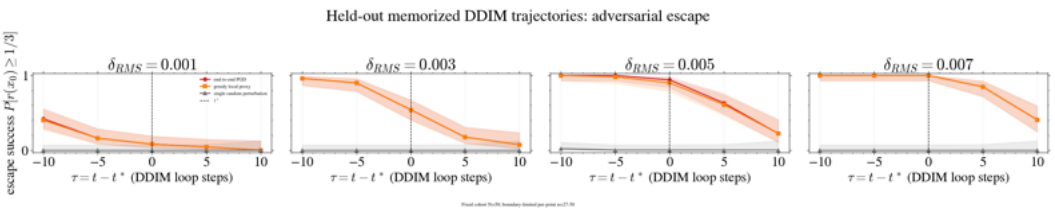


Figure 7: **Adversarial evidence for memorization basin entry.** Escape probability under bounded interventions relative to t^*. Escapeability decreases sharply through the transition, with larger perturbation bounds remaining effective later in denoising. End-to-end and greedy optimization yield similar profiles, while matched random perturbations almost never escape. Shaded regions show 95% Wilson confidence intervals.

shifts this transition later in sampling, as expected from the dependence of $B_t(\delta)$ on the admissible continuation set. End-to-end and greedy optimization agree on more than 97% of trajectory–offset pairs, whereas matched random perturbations produce almost no escapes. As an additional control, we observe that successful escapes largely satisfy an independent endpoint-quality criterion.

These results support t^* as an empirical proxy for the basin entry region: bounded escape becomes rapidly more difficult around the spectral transition, with stronger interventions remaining effective for longer. A finite adversarial search cannot certify the universal quantification defining $t^\dagger$, so we do not identify the two quantities exactly.

F.2 Singular-Value Dynamics of the Spectral Transition

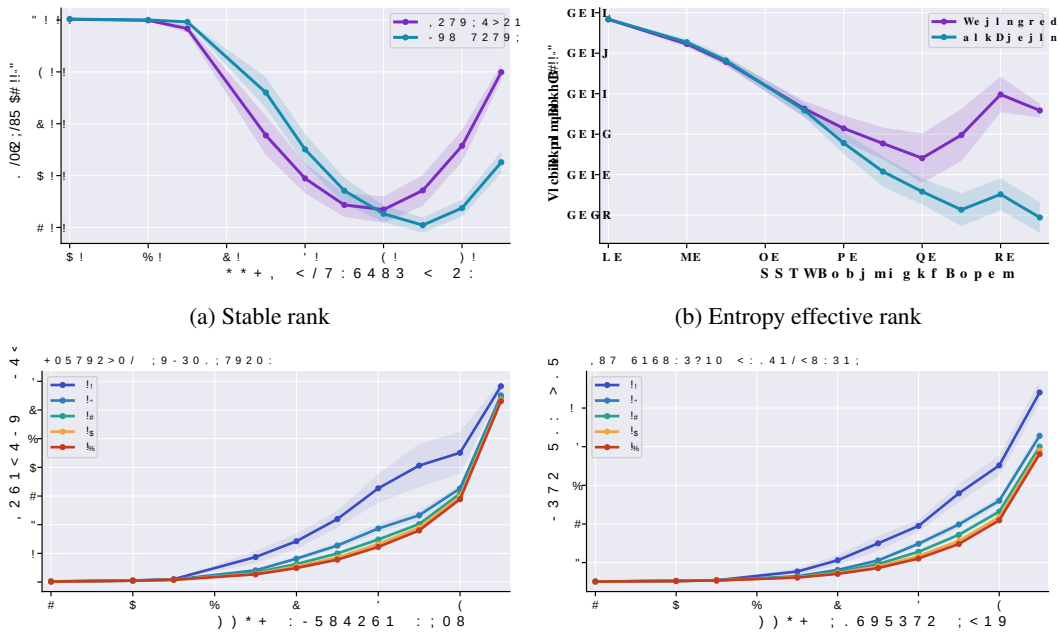
Section 3.2 summarizes the spectral transition using the relative spectral gap Δ_{12}. Here, we further investigate the underlying singular-value dynamics and demonstrate how this transition manifests in stable rank.

The spectral collapse is driven by the leading singular value. At high noise, the spectrum remains nearly isotropic in both cohorts, with singular values near unity and stable rank approaching the ambient dimension. As denoising enters intermediate noise levels, all singular values increase, but σ_1 grows faster than the bulk spectrum (Figure 8c–d). Because $\mathrm{SR}(J_t) = \sum_i \sigma_i^2/\sigma_1^2$, this relative overshoot of σ_1 induces the collapse in stable rank (Figure 8a). Spectral recovery occurs when secondary singular values subsequently catch up to σ_1.

Memorized trajectories exhibit earlier separation and stronger recovery. Along memorized trajectories, σ_1 decouples from σ_2 earlier in the sampling process and by a wider margin, causing stable rank to fall sooner. The two cohorts diverge most prominently at low noise: for memorized samples, the leading singular values reconverge toward a nearly common scale, consistent with the isotropic $1/b_t$ scaling expected near an isolated memorized state. Consequently, their stable rank recovers substantially. For non-memorized trajectories, σ_1 remains separated from the remaining spectrum, leaving the recovery incomplete. In absolute time, minimum stable rank is comparable across groups; the key distinguishing features are the earlier separation of σ_1 and its subsequent recovery, which Δ_{12} captures more directly upon alignment at t^* (Figure 1a).

Local entropy rank. Local entropy rank $\nu(J_t)$ displays the same qualitative collapse–recovery profile across a narrower dynamic range (Figure 8b). Because ν integrates over the full singular spectrum, it is dominated by bulk modes rather than a single leading singular value. Nevertheless, memorized trajectories systematically recover toward a higher-rank spectrum at low noise, whereas non-memorized trajectories continue to decay.

Rectified flow. Rectified-flow trajectories show a qualitatively similar picture (Figure 9). Along memorized trajectories, σ_1 separates from the secondary singular values earlier and by a wider margin, then declines back toward them, producing a pronounced recovery in stable rank. Along non-memorized trajectories, σ_1 continues to grow and remains separated, so stable rank stays low.



(a) Stable rank

(b) Entropy effective rank

(c) Leading singular values, memorized

(d) Leading singular values, non-memorized

Figure 8: **Singular-value dynamics of the spectral transition.** Exact full-SVD Jacobians along DDIM trajectories at the 160k checkpoint in absolute sampling time. **(a–b)** Both cohorts exhibit collapse–recovery dynamics in stable and entropy effective rank, with substantially stronger low-noise recovery for memorized trajectories. **(c–d)** Spectral collapse is driven by σ_1 separating from lower singular modes. At low noise, secondary singular values approach σ_1 for memorized trajectories, whereas σ_1 remains more strongly separated for non-memorized trajectories.

Table 5: **Training-set proximity under GeoReg.**

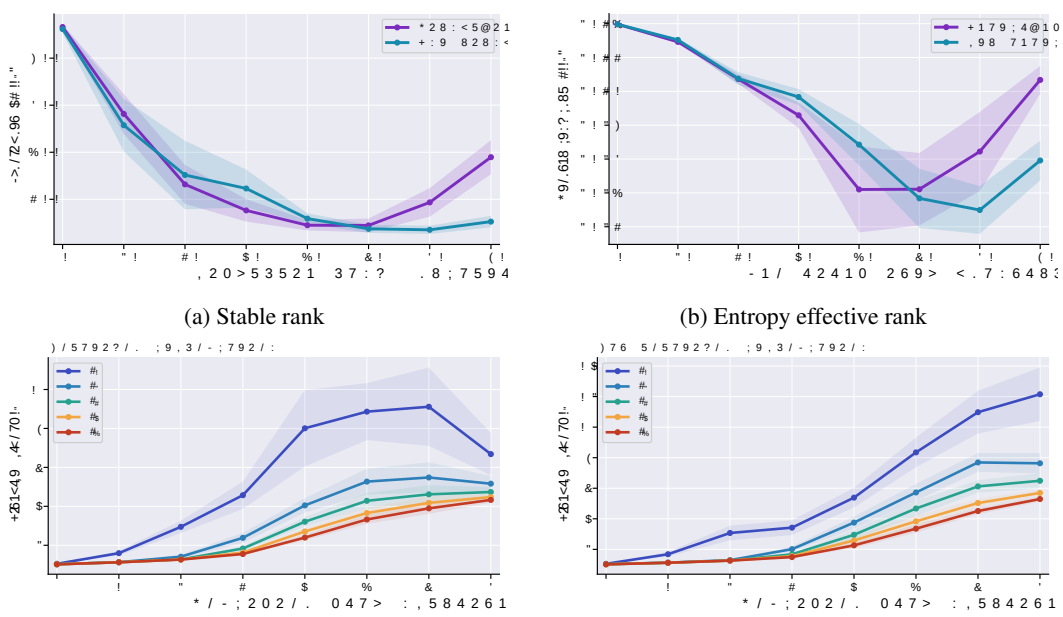
Measure	Standard	GeoReg
Copy rate ($r < 1/3$)	8.80%	1.56%
Mean r	0.766	0.846
Median r	0.861	0.910
r q05	0.273	0.470
Pixel NN	3.087	3.833
LPIPS NN	0.0216	0.0238

F.3 Generation Quality and Perceptual Effects of GeoReg-DM

We further examine how the reduction in copying affects the generated distribution. On CelebA32, we compare standard training with the final GeoReg configuration using the evaluation protocol of Sec. 5.1. Figure 10 shows representative generations from the two distributions. Despite the substantial reduction in training-set reproduction, GeoReg continues to produce coherent CelebA samples.

The change is not confined to samples close to the copying threshold. As shown in Table 5, the mean nearest-neighbor gap ratio increases from 0.766 to 0.846, while its 5th percentile increases from 0.273 to 0.470. Nearest-training distances also increase in both pixel and LPIPS space. GeoReg therefore shifts the distribution more broadly away from individual training examples rather than merely moving borderline generations across the $r = 1/3$ threshold.

This stronger separation from the training set comes with a distributional trade-off: FID against the full 4,096-image CelebA training set is 54.89 for standard training and 59.77 for GeoReg. Re-



(a) Stable rank

(b) Entropy effective rank

(c) Leading singular values, memorized

(d) Leading singular values, non-memorized

Figure 9: **Singular-value dynamics for rectified flow.** Exact full-SVD Jacobians along rectified-flow trajectories at the 160k checkpoint in absolute sampling time. (a–b) Both cohorts exhibit spectral collapse, with substantially stronger recovery for memorized trajectories. (c–d) The collapse is driven by σ_1 separating from lower modes; for memorized trajectories, σ_1 subsequently declines toward the secondary singular values, whereas for non-memorized trajectories it continues to grow.

moving memorized generations from the standard model increases its FID to 59.60, indicating that reproduction memorization contributes partially to the FID difference.

The synthetic experiments allow us to examine the quality-memorization trade-off more directly because samples can be compared with the known population rather than only with a finite image reference set. Figure 11 shows the population, standard-model generations, and GeoReg generations for Gaussian80, Mixture32, and SwissRoll8D128. Standard training places substantial probability mass around the finite training examples, most visibly on Gaussian80 and Mixture32. GeoReg spreads generations more broadly over the population while preserving its overall structure.

F.4 Spectral Effect of GeoReg-DM

Section 4.1 motivates GeoReg-DM through the spectral organization observed as memorization develops during training. To directly examine how GeoReg-DM alters this geometry, we evaluate GeoReg-DM and standard training across different fixed noised training states per timestep and checkpoint. Table 6 reports paired differences (GeoReg − Standard), and Figure 12 shows the corresponding local entropy rank profiles across diffusion timesteps.

GeoReg-DM reduces intermediate-noise spectral concentration. The clearest change appears in stable rank at intermediate noise levels. At $t = 300$, GeoReg-DM systematically increases stable rank across all evaluated checkpoints, with mean paired differences ranging from $+34.6$ to $+95.8$ (Table 6). At $t = 200$, this effect peaks at the 28k checkpoint ($+122.2$) and remains positive through 60k steps. The intervention thus counteracts spectral concentration in the exact intermediate-noise regime where standard training forms a dominant spectral mode.

GeoReg-DM attenuates low-noise rank recovery. At lower noise ($t = 100$), GeoReg-DM yields lower entropy effective rank than standard training across every checkpoint, with paired shifts



Figure 10: **Generation quality under GeoReg.** Representative samples from standard training and the final GeoReg configuration, together with their pixel-L_2 nearest training examples. GeoReg substantially reduces copying while retaining coherent CelebA generations.

Table 6: **Paired spectral differences under GEOREG-DM.** Mean paired differences (GeoReg − Standard) evaluated on 32 fixed noised training states per cell; brackets report 95% bootstrap intervals. Bold entries denote intervals excluding zero. SR denotes stable rank, ν entropy effective rank, and Δ_{12} the relative leading spectral gap.

	ΔSR				$\Delta\nu$	$\Delta\Delta_{12}$
Ckpt	$t = 100$	$t = 200$	$t = 300$	$t = 400$	$t = 100$	$t = 200$
28k	$\mathbf{+77.4}$ [63.7, 92.1]	$\mathbf{+122.2}$ [93.0, 153.1]	$\mathbf{+95.8}$ [62.0, 127.1]	$\mathbf{+15.9}$ [10.3, 22.2]	$\mathbf{-0.73}$ [−0.86, −0.60]	$\mathbf{-0.058}$ [−0.090, −0.028]
30k	$\mathbf{+14.2}$ [3.3, 25.0]	$\mathbf{+25.4}$ [9.2, 42.4]	$\mathbf{+34.6}$ [18.1, 52.2]	$+2.3$ [−3.6, 8.5]	$\mathbf{-0.46}$ [−0.52, −0.39]	-0.005 [−0.025, 0.014]
40k	$+3.3$ [−8.5, 15.0]	$\mathbf{+25.6}$ [7.9, 42.0]	$\mathbf{+43.3}$ [25.8, 61.8]	$+3.7$ [−3.7, 11.6]	$\mathbf{-0.45}$ [−0.52, −0.39]	-0.003 [−0.028, 0.022]
60k	-6.8 [−17.4, 4.8]	$+8.4$ [−10.6, 28.0]	$\mathbf{+44.6}$ [12.8, 78.0]	$+20.3$ [−1.0, 46.8]	$\mathbf{-0.58}$ [−0.65, −0.52]	$+0.003$ [−0.026, 0.032]

between -0.45 and -0.73. Whereas standard training recovers toward a high-rank spectrum at low noise (Figure 12), GEOREG-DM systematically tempers this endpoint rank recovery.

Spectral changes are localized in diffusion time. At higher noise ($t \geq 500$), spectral geometry remains closely matched between both models: stable rank differences stay within 11.4 (in ambient dimension $d = 1024$), and entropy rank differences remain under 0.05. Geometric modifications from GEOREG-DM are therefore tightly localized to the intermediate- and low-noise regimes implicated by memorization dynamics rather than applied uniformly across diffusion time.

In summary, GEOREG-DM weakens the intermediate-noise spectral concentration and attenuates subsequent low-noise rank recovery which empirically coincides with lower memorization.

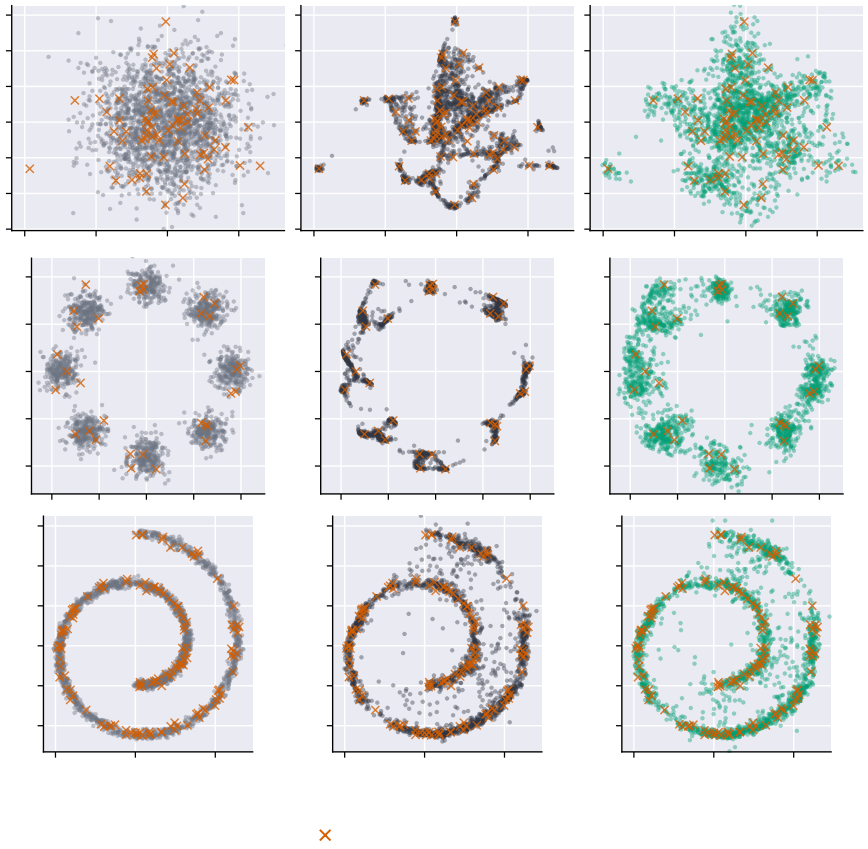


Figure 11: **GeoReg improves population coverage on synthetic data.** Rows: Gaussian80, Mixture32, and SwissRoll8D128 (shown in a 2D projection). Columns: samples from the underlying population (left), generations from standard training (middle), and generations from GeoReg (right); orange crosses mark the fixed training points. Standard training concentrates mass around individual training points, whereas GeoReg spreads generations over the population while preserving its structure.

F.5 Validation of the JER Spectral Proxy

JER replaces the nuclear norm in the Rényi-2 effective rank with an oriented trace, yielding a quantity that can be estimated efficiently from Jacobian contractions. This approximation is useful only if it preserves the spectral-dimensionality signal of the exact quantity. We therefore test this directly, independently of memorization labels and detector performance, on 128 frozen query states for which the full Jacobian spectrum can be computed.

We compare the oriented trace-to-Frobenius proxy with the exact Rényi-2 effective rank defined in Appendix B.1.

As shown in Fig. 13, the oriented trace-to-Frobenius proxy preserves the ordering of the exact Rényi-2 effective rank across query states, with Spearman correlations of 0.994 for DDIM and 0.978 for rectified flow. The proxy is conservative in magnitude, as expected from replacing the nuclear norm by the oriented trace, but the relationship remains strongly monotonic across both memorized and non-memorized states. Memorized states lie toward the high-rank end of the distribution and particularly close to the equality line. Thus, although JER is not intended as an unbiased numerical

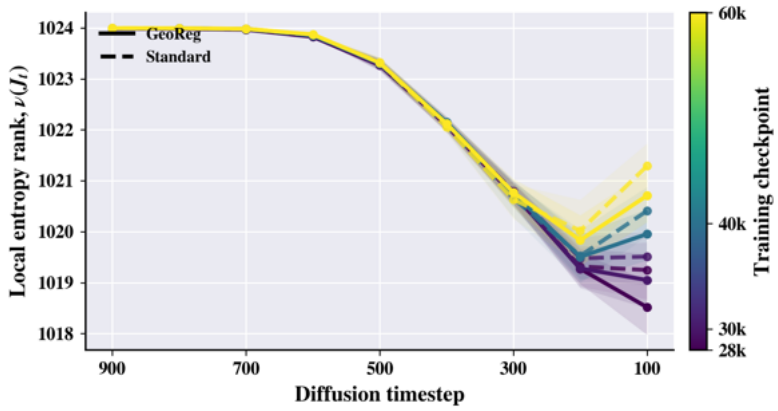


Figure 12: **Entropy effective rank under GEOREG-DM.** Local entropy effective rank $\nu(J_t)$ across diffusion time for GEOREG-DM (solid) and standard training (dashed) across checkpoints. Profiles align at higher noise and diverge near low-noise endpoints, where standard training exhibits stronger rank recovery. Intermediate-noise changes are reported in Table 6.

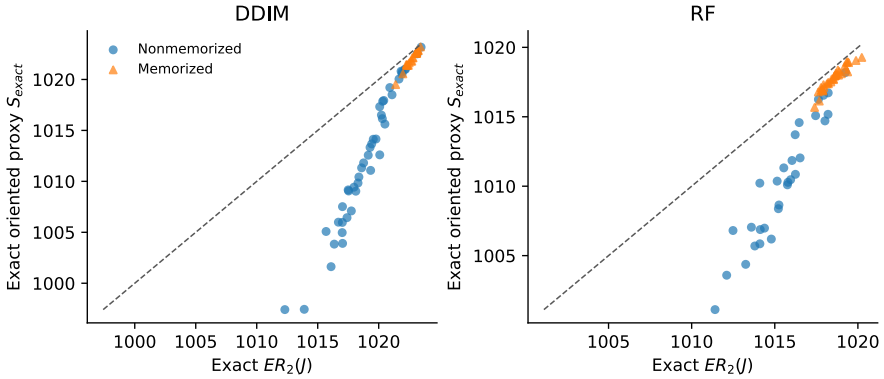


Figure 13: **Validation of the spectral proxy underlying JER.** Exact Rényi-2 effective rank (horizontal axis) versus the deterministic oriented trace-to-Frobenius proxy (vertical axis) on frozen DDIM and rectified-flow query states. Colors distinguish memorized and non-memorized generations, and the dashed line denotes equality. The proxy is conservative in magnitude but closely preserves the ordering of the exact spectral quantity across states.

estimator of ER_2, it retains the spectral ordering relevant for discrimination while avoiding an explicit spectral decomposition.

F.6 DETECTION ACROSS TRAINING CHECKPOINTS

We additionally evaluate whether JER remains informative as memorization develops during training. We repeat the DDIM detection protocol of Appendix B.3 at checkpoints from `60k` to `260k`, using disjoint calibration and held-out sets of 1,000 trajectories at each checkpoint. Across this interval, held-out memorization prevalence increases from 7.7% to 52.9%.

As shown in Table 7, JER maintains consistently high AUROC across checkpoints and outperforms FLIPD and score magnitude throughout training. The low-prevalence setting provides a particularly stringent test: at `60k`, JER reaches 0.504 AUPRC against a memorization prevalence of only 0.077, compared with 0.252 for FLIPD and 0.371 for score magnitude. AUPRC naturally increases as memorized generations become more prevalent, whereas the consistently high AUROC shows that JER retains its discriminative signal across stages of training.

Table 7: **DDIM memorization detection across training.** Held-out AUROC / AUPRC at successive model checkpoints. Prevalence is the fraction of memorized generations in the held-out set; `160k` corresponds to the DDIM evaluation in Table 2.

Ckpt	Prevalence	JER	FLIPD	Score magnitude
60k	7.7%	**0.940 / 0.504**	0.802 / 0.252	0.834 / 0.371
100k	21.9%	**0.943 / 0.793**	0.819 / 0.571	0.811 / 0.591
160k	33.4%	**0.956 / 0.895**	0.845 / 0.722	0.860 / 0.785
260k	52.9%	**0.957 / 0.951**	0.922 / 0.921	0.907 / 0.923

Table 8: **Stable Diffusion prompt-level memorization detection.** Results are evaluated on a frozen balanced set of 800 prompts. Confidence intervals are highlighted in the brackets. TPR@1% is the maximum TPR at FPR $\leq 1\%$. In the masked setting, padding positions are excluded from the CLIP text encoding, whereas in the no-mask setting they remain visible, allowing the post-EOS token tail to influence the conditioning representation.

	Masked		No mask	
Detector	AUROC (95% CI)	TPR@1%	AUROC (95% CI)	TPR@1%
Conditional JER (ours)	**0.712** [0.677, 0.747]	30.3%	0.974 [0.967, 0.979]	75.5%
Unconditional JER (ours)	0.671 [0.633, 0.706]	23.8%	0.914 [0.908, 0.921]	51.8%
Conditional score norm	0.513 [0.474, 0.555]	1.8%	0.102 [0.080, 0.125]	0.0%
Unconditional score norm	0.701 [0.695, 0.708]	16.3%	0.807 [0.777, 0.837]	10.3%
FLIPD Ross et al. (2025)	0.494 [0.454, 0.534]	1.0%	0.802 [0.770, 0.832]	6.5%
Score-difference norm Wen et al. (2024)	0.668 [0.630, 0.705]	27.5%	0.985 [0.976, 0.992]	90.8%
Hessian score difference Jeon et al. (2024)	0.664 [0.624, 0.703]	34.8%	0.990 [0.983, 0.995]	91.5%
Anisotropy Asthana & Belagiannis (2026)	0.680 [0.643, 0.716]	21.3%	0.989 [0.982, 0.994]	91.3%

F.7 Stable Diffusion Detection and Conditioning Sensitivity

We further evaluate JER on Stable Diffusion v1.4 using released collections of memorized and non-memorized prompts. Unlike the controlled experiments, the target in this benchmark is membership in these prompt collections rather than whether an individual generation reproduces a training example. This distinction is important because several diffusion memorization detectors exploit atypical conditional responses associated with memorized prompts.

Protocol. The evaluation contains 1,000 prompts, equally divided between the released memorized and non-memorized collections. We freeze 200 prompts for calibration and evaluate on the remaining balanced set of 800 prompts. All detectors use identical Stable Diffusion v1.4 and DDIM generation settings. We report held-out AUROC with 95% confidence intervals obtained from 10,000 stratified bootstrap resamples over prompts, together with the maximum true-positive rate at FPR $\leq 1\%$.

We evaluate two conditioning conventions. With *masked* conditioning, the tokenizer padding attention mask is passed to the CLIP text encoder. With *no-mask* conditioning, padding positions remain visible and the post-EOS token tail can contribute to the conditioning representation. Prompts, labels, splits, sampling schedules, and detector definitions are otherwise unchanged.

Detection at Stable Diffusion scale. Under masked conditioning, conditional JER reaches 0.712 AUROC and 30.3% TPR at 1% FPR, compared with 0.671 for unconditional JER, and near-chance performance for the corresponding conditional score norm (0.513). Thus, the local spectral statistic remains informative at Stable Diffusion scale and captures structure beyond prediction magnitude alone. At the stringent operating point, its 30.3% TPR is also close to the strongest detector in this setting, the Hessian score-difference statistic at 34.8%.

What does the prompt benchmark measure? Performance changes markedly when the post-EOS conditioning tail is exposed. Conditional JER rises to 0.974 AUROC, while score-difference, Hessian, and anisotropy statistics reach 0.985–0.990. This sensitivity is relevant to the interpre-

tation of the benchmark. Many diffusion memorization detectors operate by identifying unusually strong or anisotropic conditional responses: such responses can characterize prompts associated with memorization, but do not by themselves establish that a particular generated image copies a training example. Indeed, the released prompt collections are themselves highly distinguishable from text alone: a character 3–5-gram classifier reaches 0.986 AUROC and uppercase-character count reaches 0.968.

The conditioning comparison provides a concrete example of how this prompt structure can enter model-based scores. Masked and no-mask CLIP representations agree through the first EOS token, while the post-EOS tail remains available to U-Net cross-attention when it is not masked. Removing this tail at cross-attention reduces fixed-state epsilon-difference discrimination from 0.970 to 0.774, close to its masked value of 0.783. We therefore view the masked results as the more conservative evidence that JER transfers to a large pretrained diffusion model. The substantially higher no-mask results also expose a limitation of prompt-list evaluation: success can depend on prompt-source and conditioning structure whose relation to generation-level copying may not transfer to less curated prompt distributions.

F.8 GeoSteer-DM Across Training Checkpoints

Table 9 provides the complete full-Jacobian GeoSteer-DM evaluation on CelebA across training checkpoints. Beyond reducing aggregate memorization rates, steering systematically shifts the full gap-ratio distribution away from training references: at 300k steps, the median gap ratio r increases from 0.31 to 0.90 for DDIM and from 0.26 to 0.85 for rectified flow.

While steering increases FID across all checkpoints, this reflects a known artifact of the metric: because FID is computed using 500 generations against a fixed 1,024-image reference set, exact training-set reproduction artificially decreases FID scores (Appendix F.3). We therefore treat FID as a secondary metric in this controlled setting. Qualitative examples at 160k steps (Figure 14) confirm that GeoSteer-DM redirects memorized trajectories initialized from identical noise toward distinct, plausible CelebA samples with altered facial attributes.

Table 9: **GeoSteer on CelebA across training checkpoints.** Each row reports unmodified sampling $\rightarrow$ GeoSteer on 500 paired generations (five seeds, 100 samples each) sharing initial noise. FID is computed from 500 generations per arm against a fixed 1,024-image CelebA reference set (Appendix E.4).

Sampler	Ckpt	Copy (%) $\downarrow$	Mean r $\uparrow$	Median r $\uparrow$	FID $\downarrow$
DDIM	30k	1.4 $\rightarrow$ 1.0	0.85 $\rightarrow$ 0.89	0.91 $\rightarrow$ 0.94	64.9 $\rightarrow$ 74.0
	60k	8.4 $\rightarrow$ 3.2	0.76 $\rightarrow$ 0.85	0.84 $\rightarrow$ 0.92	60.9 $\rightarrow$ 66.6
	100k	24.6 $\rightarrow$ 6.0	0.63 $\rightarrow$ 0.83	0.69 $\rightarrow$ 0.90	56.5 $\rightarrow$ 64.6
	160k	36.4 $\rightarrow$ 6.8	0.55 $\rightarrow$ 0.83	0.54 $\rightarrow$ 0.91	52.1 $\rightarrow$ 65.5
	200k	42.0 $\rightarrow$ 4.4	0.52 $\rightarrow$ 0.84	0.47 $\rightarrow$ 0.91	52.1 $\rightarrow$ 67.0
	300k	51.6 $\rightarrow$ 7.6	0.45 $\rightarrow$ 0.82	0.31 $\rightarrow$ 0.90	47.2 $\rightarrow$ 61.3
RF	30k	7.8 $\rightarrow$ 0.4	0.76 $\rightarrow$ 0.89	0.85 $\rightarrow$ 0.93	39.3 $\rightarrow$ 45.4
	60k	26.8 $\rightarrow$ 1.6	0.62 $\rightarrow$ 0.85	0.69 $\rightarrow$ 0.91	38.0 $\rightarrow$ 47.5
	100k	36.6 $\rightarrow$ 3.4	0.52 $\rightarrow$ 0.81	0.45 $\rightarrow$ 0.89	34.8 $\rightarrow$ 45.3
	160k	54.6 $\rightarrow$ 6.8	0.43 $\rightarrow$ 0.79	0.30 $\rightarrow$ 0.87	34.0 $\rightarrow$ 41.8
	200k	52.4 $\rightarrow$ 7.4	0.42 $\rightarrow$ 0.77	0.30 $\rightarrow$ 0.85	33.6 $\rightarrow$ 41.7
	300k	60.2 $\rightarrow$ 8.0	0.40 $\rightarrow$ 0.78	0.26 $\rightarrow$ 0.85	33.5 $\rightarrow$ 42.5

F.9 Computational Cost of GeoSteer-DM

We evaluate how the dimensionality of the Jacobian sketch affects the steering signal by varying $k \in \{8, 16, 32, 64\}$ and comparing against the full-Jacobian intervention. Table 10 reports memorization rates pooled across the six controlled-model checkpoints, together with the median runtime of each sketched steering configuration.

Increasing geometric resolution generally improves steering effectiveness. For rectified flow, memorization decreases consistently from 32.7% at $k = 8$ to 14.0% at $k = 64$, compared with 6.3% using

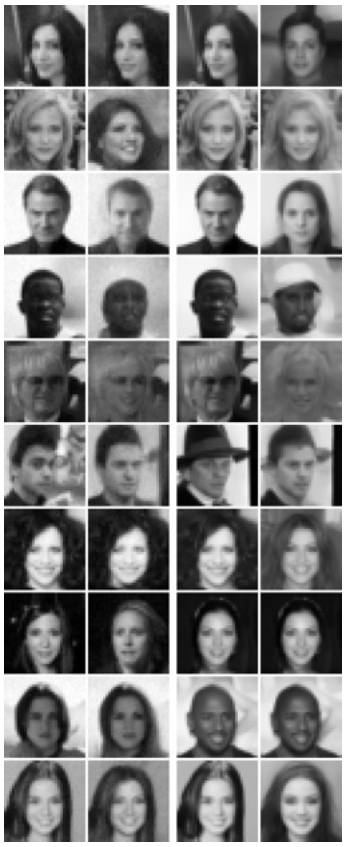


Figure 14: **Qualitative examples of geometric steering on CelebA.** Matched generations at checkpoint 160k. Columns, left to right: DDIM, DDIM + GeoSteer, rectified flow, rectified flow + GeoSteer. Each steered sample starts from the same initial noise as the unsteered sample to its left. All unsteered samples are memorized ($r < 1/3$), and all steered samples are not ($r > 1/3$). Steering redirects the trajectory to a distinct, plausible CelebA generation, which can change identity attributes.

Table 10: **Effect of geometric resolution on steering.** Memorization rates are pooled across the six controlled-model checkpoints. Runtime denotes the median wall-clock time per evaluated steering arm across checkpoints. Full-spectrum timing was not recorded in this benchmark.

Geometry	DDIM mem. ↓	RF mem. ↓	DDIM time (s) ↓	RF time (s) ↓
$k = 8$	18.0%	32.7%	3.50	5.96
$k = 16$	19.7%	24.7%	3.47	6.29
$k = 32$	19.0%	18.0%	3.52	6.21
$k = 64$	10.0%	14.0%	3.82	6.84
Full spectrum	**3.0%**	**6.3%**	–	–

the full Jacobian. DDIM exhibits the same overall dependence on resolution, although the smaller sketches are not monotonically ordered: memorization remains near 18–20% for $k \in \{8, 16, 32\}$ before decreasing to 10.0% at $k = 64$, compared with 3.0% under full-spectrum steering.

The computational cost varies comparatively little across the tested sketch dimensions. Median runtime ranges from 3.47–3.82 s for DDIM and 5.96–6.84 s for rectified flow. Thus, increasing k substantially improves the geometric resolution available to the intervention without a corresponding increase of the same magnitude in runtime. Full-Jacobian timings were not recorded in this benchmark and are therefore not included in the runtime comparison.

Table 11: **Direction-specificity of geometric steering.** Matched comparison at the `300k` checkpoint across 500 trajectories per sampler. The random-plane control matches the instantaneous displacement norm and rotation angle of GEOSTEER-DM relative to the unmodified update while randomizing the orientation plane.

Sampler	Method	Copy rate ↓	Mean r ↑	Median r ↑
DDIM	Baseline	51.6%	0.454	0.314
	Matched random rotation	50.8%	0.460	0.326
	GEOSTEER-DM	**7.6%**	**0.821**	**0.897**
RF	Baseline	60.2%	0.397	0.257
	Matched random rotation	58.6%	0.401	0.262
	GEOSTEER-DM	**8.0%**	**0.778**	**0.851**

F.10 DIRECTION-SPECIFICITY OF GEOSTEER-DM

A key control question is whether the anti-memorization effect of GEOSTEER-DM arises specifically from the identified spectral direction, or whether perturbing trajectories by a comparable magnitude produces a similar reduction. We test this hypothesis using a matched random-direction control that preserves the instantaneous magnitude and rotation angle of GEOSTEER-DM while randomizing its orientation.

At each active sampler step, let $\delta = x_{\text{base}} - x_{\text{anchor}}$ denote the unsteered model displacement, and let δ_{geo} denote the steered displacement. We construct a random control matching both $\|\delta_{\text{geo}}\|_2$ and the deflection angle $\theta = \angle(\delta, \delta_{\text{geo}})$. Sampling a random Gaussian direction q, projecting it onto the orthogonal complement of δ, and normalizing yields:

$$\delta_{\text{rand}} = \|\delta_{\text{geo}}\|_2 \left[\cos\theta\, \widehat{\delta} + \sin\theta\, q\right], \qquad q^\top \widehat{\delta} = 0. \tag{39}$$

By construction, $\|\delta_{\text{rand}}\|_2 = \|\delta_{\text{geo}}\|_2$ and $\angle(\delta, \delta_{\text{rand}}) = \angle(\delta, \delta_{\text{geo}})$, matching the Euclidean edit magnitude $\|\delta_{\text{rand}} - \delta\|_2$ to numerical precision. All intervention windows, initial noise vectors, and sampler settings remain identical, with local geometry recomputed dynamically along each edited trajectory.

We evaluate this control on 500 paired trajectories at the `300k` checkpoint. For DDIM, GEOSTEER-DM reflects across the leading direction of the full noise-prediction Jacobian over steps 60–98.

As detailed in Table 11, matching intervention magnitude alone is insufficient to escape memorization basins. For DDIM, baseline memorization (51.6%) remains essentially unchanged under matched random rotation (50.8%), whereas GEOSTEER-DM reduces it to 7.6%. Similarly, for rectified flow, memorization shifts from 60.2% to 58.6% under random control, compared to 8.0% under GEOSTEER-DM. Paired bootstrap analysis confirms this gap: relative to the random control, GEOSTEER-DM reduces copying by 43.16 percentage points on DDIM (95% CI $[-47.8, -38.4]$) and 50.58 percentage points on rectified flow (95% CI $[-55.4, -46.0]$). In contrast, the random rotation induces negligible change relative to baseline (DDIM: -0.81%, 95% CI $[-2.6, 1.0]$; RF: -1.59%, 95% CI $[-3.2, 0.0]$).

Crucially, these interventions are locally small: the median deflection angle is only $4.00°$ for DDIM and $4.06°$ for rectified flow. Applying these minor rotations along the leading spectral direction systematically redirects trajectories away from memorization basins, whereas equally strong rotations in random planes produce almost no effect. This confirms that the anti-memorization property of GEOSTEER-DM is strictly direction-specific.

F.11 GUIDANCE-RESIDUAL VERSUS FEATURE-JACOBIAN STEERING

At Stable Diffusion scale, multiple local Jacobian operators can provide directions for geometric steering. We compare the guidance-residual Jacobian J_g used in our primary method against a feature-space alternative,

$$J_F = \frac{\partial F_c(z_t, t)}{\partial z_t},$$

Table 12: **Stable Diffusion steering with alternative geometric operators.** Results on 500 memorized prompts with four paired generations per prompt. J_F denotes the conditional mid-block feature Jacobian, while J_g denotes the Jacobian of the classifier-free guidance residual. Lower SSCD indicates lower similarity to memorized training references.

Method	Mean SSCD ↓	SSCD ≥ 0.50 ↓	CLIP ↑	FID ↓
No mitigation	0.362	30.30%	0.279	119.42
J_F GeoSteer	0.259	14.15%	0.270	109.15
J_g GeoSteer	**0.229**	**9.60%**	0.262	**102.74**

where F_c denotes the conditional U-Net mid-block representation. For J_F, we obtain the steering direction from the leading right singular vector of a fixed rank-32 left sketch WJ_F. For J_g, we use the matrix-free power iteration described in Appendix C.5. All other components of the steering rule are held fixed: both methods act on the classifier-free guidance residual via

$$g_{\text{geo}} = 0.7(I - 2v_1v_1^\top)g$$

over denoising-order indices 0–17.

The two operators capture distinct notions of local sensitivity. J_F identifies latent directions that most strongly affect an internal conditional representation, whereas J_g directly differentiates the conditional–unconditional residual modified by GeoSteer. Consistent with this distinction, a separate pilot comparison finds a squared subspace overlap of only approximately 0.286 between their leading directions.

Table 12 reports a matched full-scale comparison on the memorized-prompt cohort. Replacing J_F with J_g produces a stronger reduction in training-reference similarity: mean SSCD decreases from 0.259 to 0.229, while the fraction of generations with SSCD ≥ 0.50 decreases from 14.15% to 9.60%, compared with 30.30% for unmodified generation. This stronger intervention is accompanied by a larger reduction in CLIP prompt alignment.

The same comparison on non-memorized prompts shows that neither intervention is confined to memorized generations. Mean SSCD changes from 0.191 for the unmodified sampler to 0.172 under J_F and 0.167 under J_g, while CLIP changes from 0.268 to 0.257 and 0.251, respectively. Thus, directly differentiating the guidance residual yields a stronger intervention on the memorized cohort, but also produces measurable changes on non-memorized prompts.

F.12 Control Study on Non-Memorized Generations

Table 13 shows that J_g-based GeoSteer substantially reduces training-reference similarity on the memorized-prompt cohort, both independently and when composed with existing mitigation methods. We next test whether these reductions could instead reflect a generic tendency to perturb all generations away from their reference images.

We construct a separate control cohort of 100 prompts that are not memorized by the unmodified Stable Diffusion model. For these prompts, lower SSCD is not itself an objective; rather, the measurements quantify how strongly the interventions change reference similarity when memorization is absent. The unmodified control cohort has a mean SSCD of 0.191, a 90th percentile of 0.262, and only 1% of generations above the SSCD ≥ 0.50 threshold.

Under J_g-based GeoSteer, mean SSCD decreases from 0.191 to 0.167 and CLIP alignment from 0.268 to 0.251 (Table 14). These changes are substantially smaller than on the memorized cohort, where mean SSCD decreases from 0.362 to 0.229 and the fraction above the SSCD ≥ 0.50 threshold decreases from 30.30% to 9.60%. The same qualitative asymmetry is observed when GeoSteer is composed with the existing mitigation methods. GeoSteer-DM is not inert on non-memorized prompts, but its impact on reference similarity is substantially stronger on memorized ones. This contrast suggests that the reductions on memorized samples are not driven by generic trajectory noise.

Table 13: **Guidance-residual GeoSteer complements existing Stable Diffusion mitigation methods.** Results on the memorized-prompt cohort. The first block compares unmodified generation with J_g-based GeoSteer. In the composition block, each entry reports the corresponding mitigation method alone $\rightarrow$ the same method combined with GeoSteer. Lower mean SSCD and a smaller fraction of generations with SSCD ≥ 0.50 indicate lower training-reference similarity.

Method	Mean SSCD ↓	SSCD ≥ 0.50 ↓	CLIP ↑	FID ↓
No mitigation	0.362	30.30%	0.279	119.42
J_g GeoSteer	0.229	9.60%	0.262	102.74
Composition with existing mitigation methods				
Ren et al. Ren et al. (2024)	$0.298 \rightarrow 0.217$	$13.80\% \rightarrow 3.80\%$	$0.295 \rightarrow 0.281$	$93.75 \rightarrow 79.67$
RTA Somepalli et al. (2023b)	$0.334 \rightarrow 0.213$	$23.95\% \rightarrow 5.50\%$	$0.279 \rightarrow 0.263$	$95.59 \rightarrow 82.12$
Wen et al. Wen et al. (2024)	$0.338 \rightarrow 0.210$	$26.80\% \rightarrow 6.60\%$	$0.278 \rightarrow 0.260$	$117.03 \rightarrow 102.75$
Han et al. Han et al. (2026)	$0.283 \rightarrow 0.184$	$19.05\% \rightarrow 2.50\%$	$0.271 \rightarrow 0.257$	$116.29 \rightarrow 105.97$

Table 14: **Control experiment on non-memorized prompts.** Results on 100 prompts that are not memorized by the unmodified Stable Diffusion model.

Method	Mean SSCD	SSCD $q_{0.90}$	CLIP
No mitigation	0.191	0.262	0.268
J_g GeoSteer	0.167	0.214	0.251
RTA	0.204	0.298	0.281
RTA + GeoSteer	0.183	0.238	0.271
Wen et al.	0.184	0.271	0.268
Wen et al. + GeoSteer	0.178	0.228	0.264
Han et al.	0.176	0.226	0.255
Han et al. + GeoSteer	0.174	0.225	0.247

F.13 Separating Geometric Reflection from Guidance Attenuation

The Stable Diffusion setup in our main experiments applies both a geometric reflection and a global gain reduction to the classifier-free guidance displacement. To isolate these mechanisms, we perform a paired four-arm ablation comparing:

$$I, \qquad 0.7I, \qquad I - 2P, \qquad 0.7(I - 2P),$$

where P denotes the rank-one projector onto the leading singular direction of the guidance-residual Jacobian. For every prompt and seed, all four arms are evaluated from identical initial noise states.

We evaluate on a collection of memorized and non-memorized prompt families used in the other experiments. On memorized prompts, adding reflection at a fixed gain of 0.7 reduces mean SSCD by 0.222 relative to attenuation alone (95% paired-bootstrap CI $[-0.377, -0.081]$) and lowers the copying rate by 26.7 percentage points (95% CI $[-53.3, -6.7]$), with a corresponding CLIP score shift of -0.016 (95% CI $[-0.042, 0.004]$).

Geometric reflection is also effective without attenuation: compared to the unmodified sampler (I), norm-preserving reflection ($I - 2P$) reduces mean SSCD by 0.254 (95% CI $[-0.440, -0.076]$) and the copying rate by 33.3 percentage points (95% CI $[-60.0, -13.3]$).

In the main experiments, we use $a_n = 0.7$ because it slightly improves prompt alignment over un-attenuated reflection while preserving most of the anti-memorization effect. Because attenuation alone fails to suppress memorization, this confirms that the reduction is driven primarily by geometric reflection.

On non-memorized control prompts, no sample crosses the copying threshold under any arm, and SSCD shifts remain noticeably smaller than on memorized prompts. These results confirm that the observed anti-copying effect is not driven by guidance attenuation alone; rather, geometric reflection provides substantial, independent suppression of training-reference similarity.

Table 15: **Reflection versus attenuation on Stable Diffusion.** Paired ablation on memorized prompt families.

Method	Mean SSCD ↓	Copy rate ↓	CLIP ↑
Baseline	0.847	93.3%	0.325
Attenuation only	0.870	93.3%	0.324
Reflection only	0.593	60.0%	0.302
Reflection + attenuation	0.649	66.7%	0.308

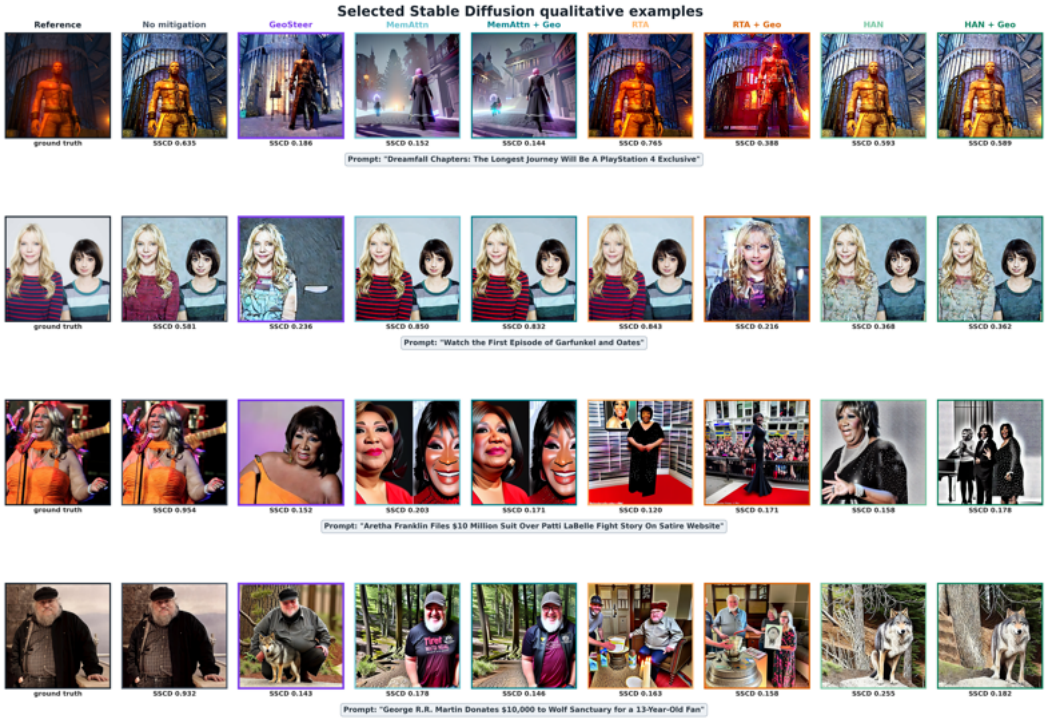


Figure 15: **Qualitative Stable Diffusion examples.** Representative generations from the memorized-prompt evaluation. For each example, the training reference, unmodified generation, and GeoSteer generation are shown for the same prompt and initial noise. GeoSteer reduces training-reference similarity while retaining the principal semantic content of the prompt.

F.14 Qualitative Stable Diffusion Examples

Figure 15 shows representative paired generations from the Stable Diffusion evaluation. For each example, the unmodified and GeoSteer generations use the same prompt and initial noise. GeoSteer moves outputs away from the memorized training reference while largely preserving the semantic content of the prompt, complementing the aggregate SSCD and CLIP results in Sec. 5.2.