

Determinism of Randomness: Prompt-Residual Seed Shaping for Diffusion Generation

Song Yan¹ Wei Zhai^{1*} Chenfeng Wang² Xinliang Bi³ Jian Yang⁴ Yancheng Cai⁵

Yusen Zhang³ Yunwei Lan¹ Tao Zhang⁶ Guanye Xiong³ Min Li^{3*}

Zheng-Jun Zha¹

¹USTC ²Li Auto Inc. ³Xi'an High-tech Research Institute
⁴Wechat Vision ⁵Cambridge University ⁶HUST
 gary_144@mail.ustc.edu.cn

Abstract

Diffusion models start generation from an isotropic Gaussian latent, yet changing only the random seed can lead to large differences in prompt faithfulness, composition, and visual quality. We study this seed sensitivity through the semantic map from initial noise to generated meaning. Although the sampling flow is locally invertible, the subsequent semantic projection is many-to-one, inducing a degenerate pullback semi-metric on the latent space: most local directions are nearly semantic-invariant, while semantic-sensitive variation is concentrated in a much smaller horizontal subspace. This provides an explanatory geometric view of the seed lottery. Motivated by this view, we introduce a training-free prompt-residual seed-shaping procedure. Rather than claiming to recover the exact horizontal space, the method uses a single high-noise cold-start prompt residual as a model-coupled proxy, injects only its tangential component, and retracts the seed to the original Gaussian radius shell. This keeps the initialization prior-compatible while adding only one conditional/unconditional probe before standard sampling. Across multiple generation benchmarks, the method improves alignment and quality metrics over standard sampling, supporting both the practical value of the proxy and the explanatory relevance of semantic anisotropy.

1 Introduction

Diffusion models are increasingly used in settings where generation is expected to be controllable, prompt-faithful, and stable under nuisance factors. A persistent source of instability is the initial random seed. With the prompt and model fixed, two seeds of comparable Euclidean norm can yield sharply different semantic correctness, compositional structure, typography, and artifact rates. This phenomenon, often described as a *Seed Lottery*, leads to repeated sampling, cherry-picking, and inflated evaluation variance.

A paradox: isotropic priors, anisotropic outcomes. Standard pipelines sample the initial latent from an isotropic Gaussian prior $z_T \sim \mathcal{N}(0, I)$, which assigns no privileged Euclidean direction. If this prior geometry were the geometry relevant to semantic behavior, seed perturbations of comparable size would have comparable effects. In practice, they do not: as illustrated in Figure 1, some directions induce large semantic changes or failures,

*Corresponding author.

whereas many others produce only minor variation. The issue is therefore not randomness per se, but the mismatch between an isotropic input law and an anisotropic semantic response.

Core viewpoint: semantics induces a degenerate latent geometry. The sampling flow itself cannot explain this collapse of degrees of freedom. Under standard local uniqueness conditions for the continuous probability-flow ODE, the map $\Psi_{T \rightarrow 0}$ is locally invertible on regular intervals. This continuous flow alone does not explain semantic many-to-one collapse; semantic identification, however, is many-to-one. We model the effective semantic outcome by

$$\Phi = \pi \circ \Psi_{T \rightarrow 0} : \mathcal{Z} \cong \mathbb{R}^d \rightarrow \mathcal{M}, \quad (1)$$

where π is an operational semantic projection, such as CLIP/DINO features followed by a local chart projection; Appendix B studies the sensitivity to this choice. On regular regions with $\text{rank}(D\Phi_z) = k \ll d$, the level sets $\Phi^{-1}(x)$ form high-dimensional fibers of semantic-invariant variation. Pulling back the metric on $\mathcal{M}$ gives

$$g_{\text{lat}} = \Phi^* g_{\mathcal{M}}, \quad (2)$$

which vanishes along fiber directions and is informative only on a low-dimensional horizontal complement. Thus Euclidean seed distance is not the right measure of semantic stability; local behavior is governed by the spectrum of $D\Phi_z$ on semantic-sensitive directions.

From explanatory geometry to an algorithmic approximation. The geometric framework identifies the type of direction that can affect semantic outcomes, but it does not make $D\Phi_z$ or the exact horizontal space available to the sampler. We therefore separate the theoretical object from the implemented update: the algorithm uses the pretrained model’s prompt residual as a *model-coupled proxy* for a prompt-sensitive horizontal component, rather than claiming to recover the true horizontal direction. The estimator uses a single high-noise cold-start probe: near the start of sampling, the scheduler-scaled seed is close to the native trajectory, making this proxy both cheap and stable.

Concretely, our *Prompt-Residual Seed Shaping* procedure has two parts. First, it estimates a **cold-start prompt-residual proxy direction** by querying the pretrained model at a high-noise timestep using scheduler-consistent scaling of the initial Gaussian seed. Second, it applies **tangential injection and spherical retraction**: the radial component is removed, a small prompt-dependent step is applied, and the result is projected back to the original radius shell. Our contributions are:

1. We systematically derive a degenerate pullback geometry of the semantic map $\Phi = \pi \circ \Psi$, showing that semantic projection creates large semantic-invariant fibers and low-dimensional sensitive directions. This provides, to the best of our knowledge, the first geometric explanation of seed lottery under isotropic Gaussian initialization.
2. Guided by this view, we propose prompt-residual seed shaping, a training-free initialization method that uses a single high-noise model probe as a prompt-sensitive proxy direction. The method performs tangential injection followed by spherical retraction, preserving the sampled Gaussian radius while suggesting a broader route toward geometry-aware optimization throughout the generation process.
3. We validate both the theory and the algorithm through degeneracy diagnostics, proxy-direction analyses, and targeted ablations. Experiments show that the proposed seed shaping improves generation quality and alignment across heterogeneous backbones.

2 Related Works

2.1 Seed Sensitivity, Initial-Condition Interventions, and Latent Geometry

Diffusion models exhibit a persistent *Seed Lottery* [1, 2]. This sensitivity is often attributed to early denoising and structure commitment, where the initial latent acts as a coarse scaffold for the final generation [3, 4]. Consequently, several works intervene directly on the initial condition, including attention-based initial-noise refinement [5] and supervised noise adjustment [6].

This seed dependence is also related to broader geometric views of generative latent spaces. Prior work shows that decoder-induced pullback metrics can make latent spaces curved and anisotropic [7],

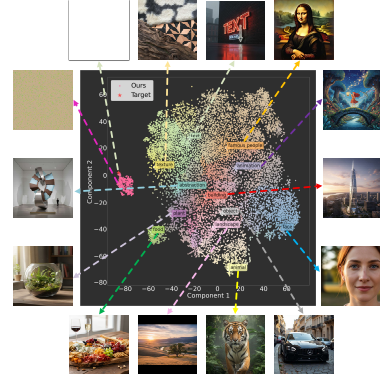


Figure 1: Semantic distribution of 20,000 random noises in FLUX.1 Dev. Details are in Appendix H.

and Riemannian formulations have been explored for diffusion and flow models on non-Euclidean domains [8, 9]. At the representation level, manifold/fiber perspectives separate semantic factors from nuisance variation [10], while Jacobian and dynamical analyses connect local spectra to stability and anisotropy [11, 12].

3 Explanatory Geometry: Latent Space Degeneracy

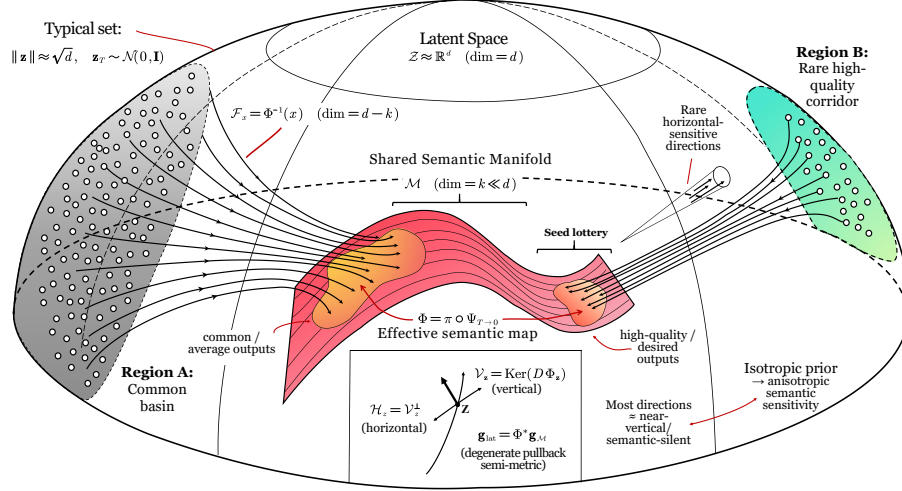


Figure 2: Geometric overview of latent-space degeneracy. Although the sampling flow is locally invertible, the semantic projection is many-to-one. Large latent fibers therefore map to similar semantic outcomes: directions tangent to a fiber are semantic-invariant, while transverse directions control semantic change. This mismatch between an isotropic Gaussian prior and an anisotropic semantic response underlies seed lottery.

3.1 Semantic map and fiber structure

Let z_t follow the probability-flow ODE

$$\frac{dz_t}{dt} = v_\theta(z_t, t; y), \quad z_0 = \Psi_{T \rightarrow 0}(z_T). \quad (3)$$

Under standard local Lipschitz regularity, $\Psi_{T \rightarrow 0}$ is locally invertible on regular intervals. Hence the collapse of degrees of freedom relevant to semantic outcomes does not come from the ODE flow alone, but from the semantic identification applied after generation. We model the effective semantic outcome by

$$\Phi := \pi \circ \Psi_{T \rightarrow 0} : \mathcal{Z} \simeq \mathbb{R}^d \rightarrow \mathcal{M}, \quad (4)$$

where π is an operational differentiable semantic projection, such as CLIP/DINO features followed by a local chart projection.

Assumption 3.1 (Regular semantic map). Locally in a semantic chart, Φ is C^2 on a dense regular set $\mathcal{Z}_{\text{reg}} = \{z : \text{rank}(D\Phi_z) = k\}$ with $k \ll d$, and the singular set has negligible Gaussian mass. When applying coarea, the chart is taken on the effective k -dimensional image manifold.

At a regular point, write $J_z := D\Phi_z$ and define the ambient vertical and horizontal spaces

$$\mathcal{V}_z := \text{Ker}(J_z), \quad \mathcal{H}_z := \mathcal{V}_z^\perp. \quad (5)$$

The fiber $\mathcal{F}_x = \Phi^{-1}(x)$ consists of seeds with the same semantic outcome, and $\mathcal{V}_z$ is its tangent space at regular points. Directions in $\mathcal{V}_z$ are semantic-invariant to first order, while directions in $\mathcal{H}_z$ are the semantic-sensitive transverse directions depicted in Figure 2.

3.2 Degenerate pullback geometry

The semantic map induces the pullback semi-metric

$$g_{\text{lat}}(z)(u, v) := \langle J_z u, J_z v \rangle_{g_{\mathcal{M}}}. \quad (6)$$

In local coordinates, with semantic metric matrix $G_{\mathcal{M}}(\Phi(z))$,

$$G_{\text{lat}}(z) = J_z^\top G_{\mathcal{M}}(\Phi(z)) J_z. \quad (7)$$

Theorem 3.2 (Degenerate geometry). *For $z \in \mathcal{Z}_{\text{reg}}$, g_{lat} vanishes exactly on $\mathcal{V}_z$ and is positive definite on $\mathcal{H}_z$. Equivalently, $G_{\text{lat}}(z)$ has rank k and nullspace $\mathcal{V}_z$.*

Proof. Appendix C.1. □

Let $\sigma_1(z) \geq \dots \geq \sigma_k(z) > 0$ be the nonzero singular values of J_z and define

$$\kappa_{\text{eff}}(z) := \frac{\sigma_1(z)^2}{\sigma_k(z)^2}. \quad (8)$$

The local semantic effect of a small perturbation is controlled by its horizontal component.

Lemma 3.3 (Local semantic sensitivity). *If Φ is C^2 near a regular point z and J_z is locally Lipschitz, then for sufficiently small Δz ,*

$$d_{\mathcal{M}}(\Phi(z + \Delta z), \Phi(z)) \leq c_2 \sigma_1(z) \|\text{Proj}_{\mathcal{H}_z} \Delta z\| + C \|\Delta z\|^2, \quad (9)$$

$$d_{\mathcal{M}}(\Phi(z + \Delta z), \Phi(z)) \geq c_1 \sigma_k(z) \|\text{Proj}_{\mathcal{H}_z} \Delta z\| - C \|\Delta z\|^2. \quad (10)$$

If $\|\text{Proj}_{\mathcal{H}_z} \Delta z\| \geq \beta \|\Delta z\|$, then, for small enough $\|\Delta z\|$,

$$d_{\mathcal{M}}(\Phi(z + \Delta z), \Phi(z)) \geq \frac{c_1}{2} \sigma_k(z) \|\text{Proj}_{\mathcal{H}_z} \Delta z\|. \quad (11)$$

Proof. Appendix C.2. □

3.3 Coarea view and interpretation

The horizontal–vertical split explains local anisotropy; the coarea formula explains why seed mass is unevenly distributed across semantic fibers. Define

$$J_k(\Phi, z) := \prod_{i=1}^k \sigma_i(z) = \sqrt{\det(J_z J_z^\top)}. \quad (12)$$

Theorem 3.4 (Coarea density transport). *If Φ is locally Lipschitz on $\mathcal{Z}_{\text{reg}}$ and $Z \sim p_Z$ has a density, then for regular values $x \in \mathcal{M}$,*

$$p_{\mathcal{M}}(x) = \int_{\mathcal{F}_x} \frac{p_Z(z)}{J_k(\Phi, z)} d\mathcal{H}^{d-k}(z). \quad (13)$$

Proof. Appendix C.3. □

Corollary 3.5 (Local compression energy). *For the Gaussian prior, the local coarea weight motivates*

$$E_{\text{eff}}(z) := \frac{1}{2} \|z\|^2 + \frac{1}{2} \log \text{pdet}(J_z J_z^\top), \quad (14)$$

up to an additive constant.

Proof. Appendix C.4. □

Mechanism. Figure 2 summarizes the mechanism. Although the Gaussian prior samples seeds at comparable norm, the semantic map collapses high-dimensional fibers to the same or nearby meanings. Fiber-tangent motion mainly changes nuisance details, whereas transverse motion can enter different semantic basins. Seed lottery therefore reflects variation in local fiber geometry and horizontal conditioning: most directions are nearly semantic-silent, while a few corridors have strong semantic response. Section 4 turns this view into a fixed-radius intervention that preserves the Gaussian norm while seeking prompt-sensitive transverse motion.

4 Algorithm: Prompt-Residual Seed Shaping

4.1 From ambient degeneracy to a radius-shell prompt update

We implement the geometric picture on the Gaussian radius shell. For $R = \|z_T\|$, keeping z_T on $\mathbb{S}_R^{d-1}$ preserves the sampled Gaussian radius and isolates directional effects. Define

$$\begin{aligned}\hat{z} &:= z/\|z\|, & Q_z &:= I - \hat{z}\hat{z}^\top, \\ A_z &:= J_z Q_z, & J_z &:= D\Phi_z.\end{aligned}\tag{15}$$

The ambient fiber split induces the shell split

$$\mathcal{V}_z^{\text{sh}} := \{u \in T_z \mathbb{S}_R^{d-1} : J_z u = 0\}, \quad \mathcal{H}_z^{\text{sh}} := (\mathcal{V}_z^{\text{sh}})^\perp \cap T_z \mathbb{S}_R^{d-1} = \text{Im}(Q_z J_z^\top).\tag{17}$$

Equivalently, the shell pullback metric is

$$G_{\text{sh}}(z) = Q_z J_z^\top G_{\mathcal{M}}(\Phi(z)) J_z Q_z,\tag{18}$$

whose null directions are shell-vertical and whose active directions lie in $\mathcal{H}_z^{\text{sh}}$. For a local prompt-semantic potential

$$F_y(z) := \ell_y(\Phi(z)),\tag{19}$$

the shell gradient is horizontal:

Lemma 4.1 (Shell prompt-gradient). *In local coordinates, absorbing the semantic metric into $\nabla_\Phi \ell_y$,*

$$\nabla_{\mathbb{S}_R} F_y(z) = Q_z J_z^\top \nabla_\Phi \ell_y(\Phi(z)) \in \mathcal{H}_z^{\text{sh}}.\tag{20}$$

Proof. For $u \in T_z \mathbb{S}_R^{d-1}$, $DF_y(z)[u] = \langle J_z^\top \nabla_\Phi \ell_y, u \rangle$ and $u = Q_z u$, so the tangent representative is $Q_z J_z^\top \nabla_\Phi \ell_y \in \text{Im}(Q_z J_z^\top)$. $\square$

Since J_z and ℓ_y are unavailable at inference time, we approximate this shell-horizontal prompt direction with a model residual.

4.2 Setup and cold-start proxy

Given prompt y , standard sampling draws $z_T \sim \mathcal{N}(0, I_d)$. We use a single high-noise cold-start probe at t^* :

$$g_\theta(z_t, t, y) := \epsilon_\theta(z_t, t, y) - \epsilon_\theta(z_t, t, \emptyset),\tag{21}$$

with ϵ_θ replaced by velocity for flow backbones, and

$$z_{t^*}^{\text{cs}} := \rho(t^*) z_T, \quad v_{\text{proxy}} := g_\theta(z_{t^*}^{\text{cs}}, t^*, y),\tag{22}$$

where $\rho(t)$ is the native scheduler scaling. Algorithm 1 then injects only the tangential component and retracts back to $\mathbb{S}_R^{d-1}$.

Residual as a noisy shell-gradient proxy. At the queried timestep, let Φ denote the downstream semantic map from the latent state to the final semantic outcome. For score- or ϵ -parameterized diffusion,

$$g_\theta(z_t, t, y) = -\sigma_t \nabla_{z_t} \log p_\theta(y | z_t),\tag{23}$$

up to score approximation, scalar factors, and sign convention. For flow backbones we absorb parameterization mismatch into a residual term. If locally $\log p_\theta(y | z) = \ell_y(\Phi(z)) + r_y(z)$, then

$$g_\theta = c_t J_z^\top \nabla_\Phi \ell_y + (c_t \nabla_z r_y + \xi_t) =: J_z^\top a_t(y) + e_t, \quad J_z := D\Phi_z.\tag{24}$$

Projecting to the shell gives

$$Q_z g_\theta(z, t, y) = c_t \nabla_{\mathbb{S}_R} F_y(z) + Q_z e_t,\tag{25}$$

so the implemented direction is a noisy shell-gradient proxy. Appendix C.5 gives the derivation.

Lemma 4.2 (Shell cone condition for the prompt residual). *Let $q_z := Q_z g_\theta(z, t, y)$. If Eq. (24) holds, then*

$$\text{Proj}_{\mathcal{V}_z^{\text{sh}}} q_z = \text{Proj}_{\mathcal{V}_z^{\text{sh}}}(Q_z e_t), \quad \|\text{Proj}_{\mathcal{V}_z^{\text{sh}}} q_z\| \leq \|Q_z e_t\|. \quad (26)$$

If, moreover,

$$\|\text{Proj}_{\mathcal{V}_z^{\text{sh}}} q_z\| \leq \rho \|\text{Proj}_{\mathcal{H}_z^{\text{sh}}} q_z\|, \quad (27)$$

then for $u = q_z / \|q_z\|$,

$$\|\text{Proj}_{\mathcal{V}_z^{\text{sh}}} u\| \leq \frac{\rho}{\sqrt{1 + \rho^2}}, \quad \angle(u, \mathcal{H}_z^{\text{sh}}) \leq \arctan \rho. \quad (28)$$

Proof. Appendix C.6. □

Remark 4.3 (Empirical form of the cone condition). The exact spaces depend on the unobserved $J_z = D\Phi_z$. In Section 5.1, we estimate a finite-dimensional shell-horizontal section $\hat{\mathcal{H}}_{z,k}^{\text{sh}}$ by radius-preserving finite differences of $\pi \circ \Psi$ around the initial seed, and measure

$$\hat{\rho}_k(z, y) := \frac{\|\text{Proj}_{\hat{\mathcal{V}}_{z,k}^{\text{sh}}} q_z\|}{\|\text{Proj}_{\hat{\mathcal{H}}_{z,k}^{\text{sh}}} q_z\| + \varepsilon}, \quad \hat{\mathcal{V}}_{z,k}^{\text{sh}} := (\hat{\mathcal{H}}_{z,k}^{\text{sh}})^\perp \cap T_z \mathbb{S}_R^{d-1}. \quad (29)$$

Small $\hat{\rho}_k$ is an empirical test of Eq. (27).

4.3 Local bridge and finite-step update

The proxy is evaluated at $z_{t^*}^{\text{cs}} = \rho(t^*)z_T$. Under the Lipschitz and boundedness assumptions in Appendix C.11, with $h = T - t^*$,

$$\|z_{t^*}^{\text{traj}} - z_{t^*}^{\text{cs}}\| \leq (B_v + L_\rho \|z_T\|) \frac{e^{L_v h} - 1}{L_v} = \mathcal{O}(T - t^*), \quad (30)$$

where the fraction is interpreted as h if $L_v = 0$.

Proposition 4.4 (Cold-start state deviation). *Under the assumptions stated above, Eq. (30) holds.*

Proof. Appendix C.11. □

Proposition 4.5 (Cold-start shell-proxy transfer). *Let $\tilde{G}_{t^*}(z) := Q_z g_\theta(\rho(t^*)z, t^*, y)$. If $\tilde{G}_{t^*}$ and $\nabla_{\mathbb{S}_R} F_y$ are locally Lipschitz with constants L_G^{sh} and L_F^{sh} , and if $\tilde{G}_{t^*}(z^{\text{ref}}) = c_{t^*} \nabla_{\mathbb{S}_R} F_y(z^{\text{ref}}) + r_{\text{ref}}$, then*

$$\tilde{G}_{t^*}(z_T) = c_{t^*} \nabla_{\mathbb{S}_R} F_y(z_T) + r_T, \quad \|r_T\| \leq \|r_{\text{ref}}\| + (L_G^{\text{sh}} + c_{t^*} L_F^{\text{sh}}) \|z_T - z^{\text{ref}}\|. \quad (31)$$

When Eq. (30) controls the reference discrepancy, the extra error is $\mathcal{O}(T - t^)$.*

Proof. Appendix C.7. □

Lemma 4.2 controls vertical leakage; the following stronger bridge interprets the residual as a noisy ascent direction for F_y . Our diagnostics directly test the former and provide operational evidence for the latter.

Theorem 4.6 (Local algorithmic bridge). *Let $R = \|z_T\|$, $F_y(z) = \ell_y(\Phi(z))$, and*

$$\bar{g}_T := Q_{z_T} g_\theta(z_{t^*}^{\text{cs}}, t^*, y), \quad u_0 := \frac{\bar{g}_T}{\|\bar{g}_T\|} \quad (\bar{g}_T \neq 0). \quad (32)$$

Assume $\nabla_{\mathbb{S}_R} F_y(z_T) \neq 0$ and, for $c_{t^} > 0$ and $0 \leq \rho_c < 1$,*

$$\bar{g}_T = c_{t^*} \nabla_{\mathbb{S}_R} F_y(z_T) + r_T, \quad \|r_T\| \leq \rho_c c_{t^*} \|\nabla_{\mathbb{S}_R} F_y(z_T)\|. \quad (33)$$

For the ε -stabilized implementation, also assume $\|\bar{g}_T\| \geq m_g > 0$. For

$$z_T^* = R \frac{z_T + \delta u_0}{\|z_T + \delta u_0\|}, \quad (34)$$

if F_y is C^2 near the shell segment and δ is small, then

$$F_y(z_T^*) - F_y(z_T) \geq \delta \frac{1 - \rho_c}{1 + \rho_c} \|\nabla_{\mathbb{S}_R} F_y(z_T)\| - C_F \delta^2 - \mathcal{O}(\delta \varepsilon / m_g). \quad (35)$$

With exact normalization the last term is absent; Proposition 4.5 adds an $\mathcal{O}(\delta(T - t^))$ cold-start term.*

Proof. Appendix C.8. $\square$

The theorem is explanatory rather than a runtime certificate because J_z , e_t , and F_y are unobserved. We therefore test the proxy using radius-preserving probes

$$z_T(\theta; u) = \|z_T\| \left(\cos \theta \frac{z_T}{\|z_T\|} + \sin \theta u \right), \quad u \perp z_T, \quad (36)$$

and semantic displacement

$$D_\pi(z_T, u; \theta) = d_\pi(\pi(\Psi(z_T(\theta; u))), \pi(\Psi(z_T))). \quad (37)$$

4.4 Tangential injection and spherical retraction

Let $\hat{z} = z_T / \|z_T\|$ and $\hat{v} = v_{\text{proxy}} / (\|v_{\text{proxy}}\| + \varepsilon)$. The implementation is

$$\tilde{v}_\perp = \hat{v} - (\hat{v}^\top \hat{z}) \hat{z}, \quad \hat{v}_\perp = \frac{\tilde{v}_\perp}{\|\tilde{v}_\perp\| + \varepsilon}, \quad z_T^* = \|z_T\| \frac{z_T + \delta \hat{v}_\perp}{\|z_T + \delta \hat{v}_\perp\| + \varepsilon}. \quad (38)$$

Proposition 4.7 (Retraction as projection onto the radius shell). *Ignoring the numerical stabilizer, $\|z_T\|(z_T + \delta \hat{v}_\perp) / \|z_T + \delta \hat{v}_\perp\|$ is the Euclidean projection of $z_T + \delta \hat{v}_\perp$ onto $\{z : \|z\| = \|z_T\|\}$; the implemented $+\varepsilon$ denominator adds only an $O(\varepsilon)$ radial perturbation.*

Proof. Appendix C.9. $\square$

Lemma 4.8 (First-order shell retraction). *Let $R = \|z_T\|$ and $u \perp z_T$, $\|u\| = 1$. For $z_T^+ = R(z_T + \delta u) / \|z_T + \delta u\|$,*

$$z_T^+ = z_T + \delta u - \frac{\delta^2}{2R^2} z_T + O\left(\frac{\delta^3}{R^2}\right), \quad (39)$$

and, for any C^2 prompt-semantic objective,

$$F_y(z_T^+) - F_y(z_T) = \delta \langle \nabla_{\mathbb{S}_R} F_y(z_T), u \rangle + O(\delta^2). \quad (40)$$

Proof. Appendix C.10. $\square$

Algorithm 1 Prior-Compatible Prompt-Residual Seed Shaping

Require: Prompt $\mathbf{y}$, model M , probe timestep t^* , rescale $\rho(\cdot)$, strength δ

Ensure: Output x_0

- 1: Sample $\mathbf{z}_T \sim \mathcal{N}(0, I)$.
 - 2: Set $\mathbf{z}_{t^*}^{\text{cs}} \leftarrow \rho(t^*) \mathbf{z}_T$ and compute $\mathbf{r} \leftarrow M(\mathbf{z}_{t^*}^{\text{cs}}, t^*, \mathbf{y}) - M(\mathbf{z}_{t^*}^{\text{cs}}, t^*, \emptyset)$.
 - 3: Normalize, remove the radial component, and re-normalize: $\hat{\mathbf{v}} \leftarrow \mathbf{r} / (\|\mathbf{r}\| + \varepsilon)$, $\mathbf{v}_\perp \leftarrow \hat{\mathbf{v}} - \frac{\langle \hat{\mathbf{v}}, \mathbf{z}_T \rangle}{\|\mathbf{z}_T\|^2 + \varepsilon} \mathbf{z}_T$,
 $\hat{\mathbf{v}}_\perp \leftarrow \mathbf{v}_\perp / (\|\mathbf{v}_\perp\| + \varepsilon)$.
 - 4: Retract to the original shell: $\tilde{\mathbf{z}}_T \leftarrow \mathbf{z}_T + \delta \hat{\mathbf{v}}_\perp$, $\mathbf{z}_T^* \leftarrow \|\mathbf{z}_T\| \tilde{\mathbf{z}}_T / (\|\tilde{\mathbf{z}}_T\| + \varepsilon)$.
 - 5: $x_0 \leftarrow \text{Sampler}(M, \mathbf{z}_T^*, \mathbf{y})$; **return** x_0 .
-

5 Experiment and Analysis

5.1 Prompt-residual proxy diagnostics

We test the empirical cone condition in Remark 4.3. For each prompt-seed pair, let $q_z := Q_{zg\theta}(z, t^*, y)$. We form a radius-preserving local section from the normalized residual direction and m random tangent controls. For tangent u , we estimate $s(u) = \frac{\pi(\Psi(\text{Ret}_z(\epsilon u))) - \pi(\Psi(z))}{\epsilon}$, $\text{Ret}_z(\xi) = R \frac{z + \xi}{\|z + \xi\|}$. The top- k right singular directions define $\hat{\mathcal{H}}_{z,k}^{\text{sh}}$. We report $\hat{\rho}_k$ from Eq. (29), horizontal

energy $E_{H,k} = \frac{\|\text{Proj}_{\hat{\mathcal{H}}_{z,k}^{\text{sh}}} q_z\|^2}{\|q_z\|^2 + \varepsilon}$, and semantic gain over random tangent controls.

Table 1 reports the main $k = 16$ setting. For FLUX, SD3.5M, and Z-Image, $\hat{\rho}_{16} < 0.2$ and horizontal energy exceeds 96%; semantic gain is also above random controls with percentile 1.0. Thus the residual is not only tangent, but concentrated in the estimated semantic-horizontal shell, empirically instantiating Lemma 4.2.

Since Φ and J_z are unavailable at runtime, these diagnostics are not certificates; they validate the proxy’s geometric premise. Additional percentiles, residual concentration, and high-noise consistency are in Appendix D.6.

Table 1: Finite-difference sectional cone diagnostics for the prompt residual. All models use official inference settings. Detailed $k = 10$ results and protocols are in Appendix A.

Model	$\hat{\rho}_{16} \downarrow$	Horizontal energy $\uparrow$	Semantic gain $\uparrow$	Percentile $\uparrow$	Model	$\hat{\rho}_{16} \downarrow$	Horizontal energy $\uparrow$	Semantic gain $\uparrow$	Percentile $\uparrow$
FLUX [13]	0.110	98.8%	2.124	1.000	Z-Image [14]	0.159	97.5%	2.420	1.000
SDXL [2]	0.189	97.2%	1.976	1.000	WAN [15]	0.114	98.2%	2.560	1.000
SD3.5M [?]]	0.198	96.2%	2.467	1.000	TRELLIS [16]	0.158	97.4%	2.133	1.000

Table 2: Main generation results across image, video, and 3D generation. See Appendix 16a for PickScore and VQAScore $\Delta[95\%CI]$ and Win Rate.

(a) Image generation.						(b) Video generation.					
Model	Dataset	Method	Pick $\uparrow$	ImgR $\uparrow$	CLIP $\uparrow$	Method	AQ $\uparrow$	AS $\uparrow$	BC $\uparrow$	DD $\uparrow$	IQ $\uparrow$
SDXL	Pick-a-Pic	Standard	17.049	-1.971	16.223	Standard	0.572	0.220	0.963	0.644	0.644
SDXL	Pick-a-Pic	Initno [†] [5] (Training-free)	17.050	-1.969	16.237	Ours	0.651	0.341	0.985	0.701	0.664
SDXL	Pick-a-Pic	NPNet [6] (Training-based)	17.051	-1.968	16.250						
SDXL	Pick-a-Pic	Ours (Training-free)	17.371	-1.852	16.642						
SDXL	DrawBench	Standard	17.416	-2.084	16.618	Method	MS $\uparrow$	OC $\uparrow$	SC $\uparrow$	TF $\uparrow$	VQA $\uparrow$
SDXL	DrawBench	Initno [†] [5] (Training-free)	17.419	-2.085	16.632	Standard	0.972	0.244	0.937	0.984	0.534
SDXL	DrawBench	NPNet [6] (Training-based)	17.420	-2.084	16.640	Ours	0.997	0.303	0.994	0.997	0.600
SDXL	DrawBench	Ours (Training-free)	17.619	-1.984	16.957						
SDXL	HPD	Standard	16.748	-1.955	14.986	(c) 3D generation.					
SDXL	HPD	Initno [†] [5] (Training-free)	16.765	-1.952	15.010	Method	FD _I $\downarrow$	KD _I $\downarrow$	FD _D $\downarrow$	KD _D $\downarrow$	VQA $\uparrow$
SDXL	HPD	NPNet [6] (Training-based)	16.780	-1.948	15.040	Standard	29.544	0.011	340.530	0.802	0.876
SDXL	HPD	Ours (Training-free)	17.041	-1.875	15.851	Ours	29.403	0.008	335.425	0.733	0.911
FLUX	Pick-a-Pic	Standard	17.134	-1.942	16.295						
FLUX	Pick-a-Pic	NPNet [6] (Training-based)	17.140	-1.936	16.305						
FLUX	Pick-a-Pic	Ours (Training-free)	17.732	-1.905	16.914						
FLUX	DrawBench	Standard	17.275	-2.051	16.642						
FLUX	DrawBench	NPNet [6] (Training-based)	17.282	-2.044	16.655						
FLUX	DrawBench	Ours (Training-free)	17.626	-1.968	17.013						
FLUX	HPD	Standard	16.730	-1.951	15.021						
FLUX	HPD	NPNet [6]	16.750	-1.944	15.900						
FLUX	HPD	Ours (Training-free)	17.239	-1.874	16.731						

I: Inception; *D*: DINOv2.

5.2 Latent degeneracy visualization

Our theory predicts anisotropic semantic sensitivity on the Gaussian radius shell. For $R = \|z_T\|$ and $\hat{z} = z_T/R$, we probe tangent directions by $\hat{u} = \frac{u - \langle u, \hat{z} \rangle \hat{z}}{\|u - \langle u, \hat{z} \rangle \hat{z}\|}$ and $z_T(\theta) = R(\cos \theta \hat{z} + \sin \theta \hat{u})$, so semantic changes are directional rather than radial. The diagnostics show heavy-tailed CLIP displacement and stronger movement along the prompt-residual proxy than random tangent controls, supporting the horizontal-vertical view of seed lottery. Full protocols and plots are in Appendix D, with detailed proxy diagnostics in Appendix D.6.

5.3 Main results across image, video, and 3D generation

Tables 2a, 16b, and 16c summarize the main image, video, and 3D results. The method improves SDXL and FLUX metrics over standard sampling and available initial-condition baselines, and transfers to WAN video and TRELLIS 3D generation without retraining. Human evaluation results are reported in Appendix I.

5.4 Seed robustness and proxy diagnostics

Matched prompt-seed HPSv3 results in Table 3a improve both mean and Q20, indicating fewer brittle low-quality seeds. The $t^* = 900$ cold-start residual already outperforms standard sampling, so the method acts as a single-probe initialization proxy rather than cross-timestep optimization.

5.5 Ablation: component choice, probe stage, and strength

Figure 3 ablates t^* and δ . Moderate δ works best: $\delta = 0$ has no effect, while large updates can over-perturb structure; with $\delta = 6$, the high-noise probe around $t^* = 900$ is most stable. This supports the method as a bounded initialization-side correction. Figure 4 compares Base, Random Tangent, Raw CFG, Wrong Prompt, and the full method on matched FLUX cases. Random Tangent removes prompt information, Raw CFG removes tangential projection and shell retraction, and Wrong Prompt uses an unrelated residual. Raw CFG often destabilizes structure, while Random Tangent and Wrong Prompt fail to reliably recover target layout, relations, or readable text. The full method more consistently adds prompt-specified elements.

Table 3: Robustness, diagnostic behavior, and operating cost on FLUX.

(a) **Seed robustness and probe diagnostic.** HPSv3 is the reported score; Q20 is the 20th percentile over matched prompt–seed cases. p denotes prompts, s/p denotes seeds per prompt, and “diag.” denotes diagnostic.

Block	Setting	HPSv3 $\uparrow$	Gain / note
Seed robust.	48p, 32s/p	10.668	+0.146, Q20: 8.697 $\rightarrow$ 8.784
	48p, 8s/p	10.667	+0.201, Q20: 8.613 $\rightarrow$ 8.897
Probe diag.	Standard	10.466	—
	$t^* = 900$	10.667	+0.201 (default)

(b) **Operating boundary and overhead.** Mod., high, and ext. denote moderate, high, and extreme guidance stress. Guidance rows report HPSv3 ranges; cost rows report runtime / peak memory.

Block	Setting	Metric / reading
Guidance	Mod. CFG	HPSv3 11.027–11.129 (stable gain)
	High CFG	HPSv3 6.00–6.31 (stress-test degradation)
	Ext. CFG	HPSv3 –1.44––1.92 (unstable stress test)
Cost	Standard	9.50–9.53s / 34.65 GB
	Ours	9.62–9.66s / 34.65 GB



Figure 3: Probe stage and injection strength.

5.6 Operating boundary and cost

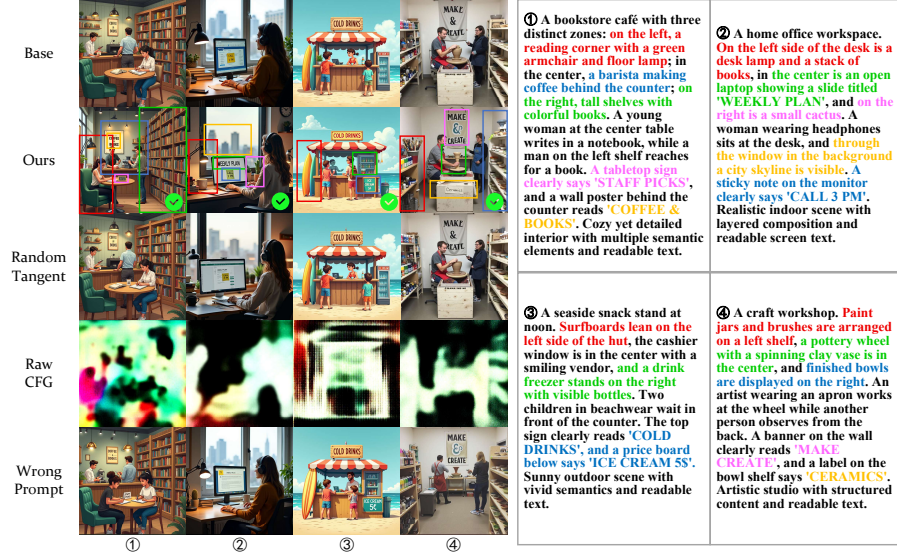


Figure 4: Qualitative component ablation of seed shaping on FLUX.

Table 3b summarizes the operating boundary and cost. The method remains stable under moderate CFG, while high/extreme CFG act as out-of-range stress tests and may overshoot. This supports viewing the method as a bounded local seed correction rather than a global steering mechanism. Its only extra cost is one conditional/unconditional cold-start probe, with modest runtime overhead and nearly unchanged peak memory. Diversity is largely preserved (Appendix E).

6 Conclusion

We presented a geometric view of seed sensitivity in diffusion and flow generation. Although the initialization prior is isotropic, the induced semantic map is not: many latent directions are nearly semantic-invariant, while meaningful variation concentrates in a low-dimensional sensitive subspace. This degenerate pullback geometry offers an explanatory account of the seed lottery. Motivated by this view, we proposed a training-free prompt-residual seed-shaping method. Rather than recovering the exact horizontal structure, the method uses a single high-noise prompt residual as a model-coupled proxy, updates the seed tangentially, and retracts it to the original Gaussian radius shell. The finite-difference sectional diagnostics further show that the shell-projected prompt residual has low vertical leakage relative to the empirical semantic-horizontal subspace. This empirically bridges the cone condition used in our local theory with the behavior of official-pipeline image generation backbones. The resulting intervention is prior-compatible, requires no retraining or cross-timestep aggregation, and improves alignment metrics across image, video, and 3D generation. Limitations see J.

References

- [1] Z. Qi, L. Bai, H. Xiong, and Z. Xie, “Not all noises are created equally: diffusion noise selection and optimization,” *ArXiv*, vol. abs/2407.14041, 2024.
- [2] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, “Sdxl: Improving latent diffusion models for high-resolution image synthesis,” in *The Twelfth International Conference on Learning Representations*, 2023.
- [3] M. F. Anjum, “Advancing diffusion models: Alias-free resampling and enhanced rotational equivariance,” 2024.
- [4] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, V. Jampani, and R. Rombach, “Stable video diffusion: Scaling latent video diffusion models to large datasets,” *arXiv preprint arXiv:2311.15127*, 2023.
- [5] X. Guo, J. Liu, M. Cui, J. Li, H. Yang, and D. Huang, “Initno: Boosting text-to-image diffusion models via initial noise optimization,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 2024, pp. 9380–9389.
- [6] Z. Zhou, S. Shao, L. Bai, S. Zhang, Z. Xu, B. Han, and Z. Xie, “Golden noise for diffusion models: A learning framework,” in *International Conference on Computer Vision*, 2025.
- [7] G. Arvanitidis, L. K. Hansen, and S. Hauberg, “Latent space oddity: On the curvature of deep generative models,” in *International Conference on Learning Representations (ICLR)*, 2018.
- [8] A. Lou, M. Xu, and S. Ermon, “Scaling riemannian diffusion models,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.
- [9] R. T. Q. Chen and Y. Lipman, “Flow matching on general geometries,” in *International Conference on Representation Learning*, B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun, Eds., vol. 2024, 2024, pp. 47 922–47 945.
- [10] X. Li, “Visual language hypothesis,” *arXiv preprint arXiv:2512.23335*, 2025, arXiv version may appear as early 2026 depending on revision.
- [11] G. Biroli, T. Bonnaire, V. de Bortoli, and M. Mézard, “Dynamical regimes of diffusion models,” *arXiv preprint arXiv:2402.18491*, 2024.
- [12] E. Ventura, B. Achilli, G. Silvestri, C. Lucibello, and L. Ambrogioni, “Manifolds, random matrices and spectral gaps: The geometric phases of generative diffusion,” in *International Conference on Representation Learning*, Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, Eds., vol. 2025, 2025, pp. 88 317–88 338.
- [13] B. F. Labs, “Flux,” <https://github.com/black-forest-labs/flux>, 2024.
- [14] Z.-I. Team, “Z-image: An efficient image generation foundation model with single-stream diffusion transformer,” *arXiv preprint arXiv:2511.22699*, 2025.
- [15] A. Wang, B. Ai, B. Wen, C. Mao, C. Xie, D. Chen, F. Yu, H. Zhao, J. Yang, J. Zeng, J. Wang, J. Zhang, J. Zhou, J. Wang, J. Chen, K. Zhu, K. Zhao, K. Yan, L. Huang, X. Meng, N. Zhang, P. Li, P. Wu, R. Chu, R. Feng, S. Zhang, S. Sun, T. Fang, T. Wang, T. Gui, T. Weng, T. Shen, W. Lin, W. Wang, W. Wang, W. Zhou, W. Wang, W. Shen, W. Yu, X. Shi, X. Huang, X. Xu, Y. Kou, Y. Lv, Y. Li, Y. Liu, Y. Wang, Y. Zhang, Y. Huang, Y. Li, Y. Wu, Y. Liu, Y. Pan, Y. Zheng, Y. Hong, Y. Shi, Y. Feng, Z. Jiang, Z. Han, Z. Wu, and Z. Liu, “Wan: Open and advanced large-scale video generative models,” *CoRR*, vol. abs/2503.20314, 2025.
- [16] J. Xiang, Z. Lv, S. Xu, Y. Deng, R. Wang, B. Zhang, D. Chen, X. Tong, and J. Yang, “Structured 3d latents for scalable and versatile 3d generation,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*. Computer Vision Foundation / IEEE, 2025, pp. 21 469–21 480.

- [17] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, and R. Rombach, “Scaling rectified flow transformers for high-resolution image synthesis,” in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.
- [18] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *J. Mach. Learn. Res.*, vol. 21, pp. 140:1–140:67, 2020.
- [19] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, and O. Levy, “Pick-a-pic: an open dataset of user preferences for text-to-image generation,” in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS ’23. Red Hook, NY, USA: Curran Associates Inc., 2023.
- [20] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, “Photorealistic text-to-image diffusion models with deep language understanding,” in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 36 479–36 494.
- [21] X. Wu, Y. Hao, K. Sun, Y. Chen, F. Zhu, R. Zhao, and H. Li, “Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis,” *arXiv preprint arXiv:2306.09341*, 2023.
- [22] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit, Y. Wang, X. Chen, L. Wang, D. Lin, Y. Qiao, and Z. Liu, “VBench: Comprehensive benchmark suite for video generative models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [23] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong, “Imagereward: Learning and evaluating human preferences for text-to-image generation,” in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023.
- [24] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, “Clipscore: A reference-free evaluation metric for image captioning,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, M. Moens, X. Huang, L. Specia, and S. W. Yih, Eds. Association for Computational Linguistics, 2021, pp. 7514–7528.
- [25] Z. Lin, D. Pathak, B. Li, J. Li, X. Xia, G. Neubig, P. Zhang, and D. Ramanan, “Evaluating text-to-visual generation with image-to-text generation,” *arXiv preprint arXiv:2404.01291*, 2024.
- [26] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, “Qwen2.5-vl technical report,” 2025.
- [27] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 6626–6637.
- [28] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 2818–2826.
- [29] J. M. Phillips and S. Venkatasubramanian, “A gentle introduction to the kernel distance,” *CoRR*, vol. abs/1103.1625, 2011.

- [30] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P. Huang, S. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jégou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, “Dinov2: Learning robust visual features without supervision,” *Trans. Mach. Learn. Res.*, vol. 2024, 2024.

A Finite-Difference Sectional Cone Diagnostics

This appendix provides the detailed protocol behind Table 1. For each prompt–seed pair, we evaluate the official pipeline at the base seed and at radius-preserving perturbed seeds. All perturbations stay on the Gaussian radius shell, so the measured differences are directional rather than radial. We use $m = 32$ random tangent controls. The empirical horizontal space $\hat{\mathcal{H}}_{z,k}^{\text{sh}}$ is obtained from the top- k right singular directions of the finite-difference semantic response matrix.

We report four quantities. First, the empirical cone ratio

$$\hat{\rho}_k = \frac{\|\text{Proj}_{\hat{\mathcal{V}}_{z,k}^{\text{sh}}} q_z\|}{\|\text{Proj}_{\hat{\mathcal{H}}_{z,k}^{\text{sh}}} q_z\| + \varepsilon}. \quad (41)$$

Second, the horizontal energy fraction

$$E_{H,k} = \frac{\|\text{Proj}_{\hat{\mathcal{H}}_{z,k}^{\text{sh}}} q_z\|^2}{\|q_z\|^2 + \varepsilon}. \quad (42)$$

Third, semantic gain compares the semantic displacement induced by the prompt residual against the mean displacement of random tangent controls. Fourth, percentile measures the fraction of random tangent controls whose semantic displacement is below the prompt-residual displacement.

Table 4: Sectional cone diagnostics at $k = 10$ and $k = 16$. This table expands Table 1; $k = 5$ is omitted because the main theory concerns sufficiently wide empirical horizontal sections.

Model	k	$\hat{\rho}_k \downarrow$	Horizontal energy $\uparrow$	Semantic gain $\uparrow$	Percentile $\uparrow$
FLUX	10	0.176	97.0%	2.124	1.000
FLUX	16	0.110	98.8%	2.124	1.000
SDXL	10	0.251	96.1%	1.976	1.000
SDXL	16	0.189	97.2%	1.976	1.000
SD3.5M	10	0.379	87.4%	2.467	1.000
SD3.5M	16	0.198	96.2%	2.467	1.000
Z-Image	10	0.288	92.3%	2.420	1.000
Z-Image	16	0.159	97.5%	2.420	1.000
WAN	10	0.126	96.7%	2.560	1.000
WAN	16	0.114	98.2%	2.560	1.000
TRELLIS	10	0.201	90.1%	2.133	1.000
TRELLIS	16	0.158	97.4%	2.133	1.000

The trend is consistent across the six backbones: increasing k from 10 to 16 reduces the estimated vertical-to-horizontal ratio and increases the horizontal energy. At $k = 16$, all five models satisfy the strong empirical cone pattern reported in the main text.

B Sensitivity of the geometric diagnostics to the choice of semantic map π

This appendix studies the robustness of our **geometry-only** diagnostics to the choice of semantic map π in $\Phi = \pi \circ \Psi$. We focus on the diagnostic procedure itself—prompt stratification via local Jacobian-spectrum statistics—and evaluate whether its qualitative behavior remains consistent across a set of frozen, widely used semantic representations.

Notation. Throughout this appendix, Ψ (a.k.a. G in the main text) denotes the generator mapping base noise z to an output image $x = \Psi(z)$. We instantiate Ψ with the Z-Image generator [14] and study the composed map $\Phi = \pi \circ \Psi$.

Semantic maps π considered. We evaluate six frozen choices of π spanning (i) CLIP², (ii) OpenCLIP/VLM-style embeddings³, and (iii) DINOv2 image embeddings fused with CLIP-text⁴. For

²<https://github.com/openai/CLIP>

³https://github.com/mlfoundations/open_clip

⁴<https://github.com/facebookresearch/dinov2>

each family, we use two fusion rules: elementwise image–text product (`itprod`) and concatenation (`concat`).

CLIP: `clip_ViT-L-14_openai_itprod`, `clip_ViT-L-14_openai_concat`.

VLM: `vlm_ViT-H-14_laion2b_s32b_b79k_itprod`, `vlm_ViT-H-14_laion2b_s32b_b79k_concat`.

DINOv2 + CLIP-text: `dinov2_vitb14_cliptext_itprod`, `dinov2_vitb14_cliptext_concat`.

Jacobian estimator and spectrum statistics. For each prompt, we estimate a randomized local sketch of the Jacobian of $\Phi = \pi \circ \Psi$ at a fixed base noise using symmetric finite differences along random directions. We probe $r = 6$ random directions, so the maximum observable rank, effective rank, and participation ratio are upper-bounded by 6. Accordingly, the measured spectrum should be interpreted as a finite local section of semantic sensitivity, not as a full-rank estimate of $D\Phi$. We report four statistics from the probed singular values: (i) smallest singular value $s_{\min}$, (ii) effective rank, (iii) participation ratio, and (iv) condition number κ reported as `cond`.

How to read the plots. The evaluator produces three plot types:

- **Violin plots** (`fig_spectrum_stats_violin_*`): prompt-level mean distributions, with seeds aggregated per prompt.
- **Heatmaps** (`fig_pi_spearman_heatmap_*`): pairwise Spearman correlations of prompt-level means across different π .
- **Seed-stability boxplots** (`fig_seed_stability_box_*`): within-prompt seed variability, e.g., per-prompt standard deviation.

The conclusions below correspond to these plots.

Main observation. Across all six choices of π , the diagnostics show a consistent finite-section spectral pattern. Although the numerical scale of some quantities, especially $s_{\min}$ and `cond`, naturally depends on the representation, the rank-type statistics are stable across semantic maps. Effective rank remains close to the probe ceiling $r = 6$, and participation ratio stays in a narrow range below that ceiling. This indicates that the probed local semantic response is not dominated by an idiosyncratic single direction and that the diagnostic procedure remains well-behaved across different choices of π . Thus, these diagnostics can be used for prompt-level stratification without committing to one particular semantic encoder.

Scale and structural consistency. Table 5 reports prompt-averaged metrics with 95% bootstrap CIs over 211 HPD prompts. The results show two complementary trends:

- **Representation-dependent scale.** The prompt-averaged $s_{\min}$ varies across representations, roughly from ≈ 0.040 to ≈ 0.066 , reflecting the different sensitivity units induced by each π .
- **Stable finite-section spectral structure.** Across π , effective rank remains around ≈ 5.37 – 5.44 out of the maximum observable value 6, while participation ratio remains around ≈ 4.82 – 4.98 . This shows that the local six-direction Jacobian sketch has a stable spectral shape across semantic maps. The result should not be read as an estimate of the full semantic rank of $D\Phi$; rather, it indicates that the same finite-section diagnostic behaves consistently under different semantic representations.

Thus, while different semantic maps induce different measurement scales, the relative spectral structure used by our diagnostics remains consistent.

Seed stability. Table 5 also reports the mean within-prompt standard deviation across seeds for each metric and π . The rank-type metrics show similar seed-stability behavior across semantic maps, while scale-dependent quantities such as $s_{\min}$ and `cond` reflect the expected representation-specific units. Overall, the diagnostics remain stable enough to support prompt-level comparisons.

Geometry– Δ correlations. We additionally compute prompt-level Spearman correlations between geometry metrics and Δ under a fixed evaluator. These results provide a complementary view of

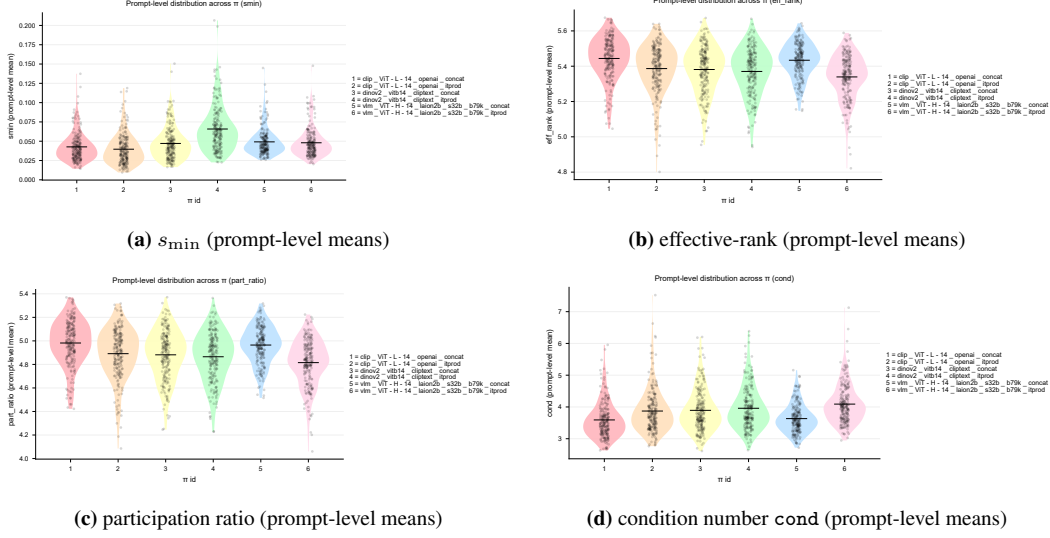


Figure 5: Prompt-level distributions of Jacobian-spectrum statistics across π . Each panel shows the distribution across prompts of the prompt-level mean, with seeds aggregated per prompt, for a given metric and semantic map. These plots diagnose how the *scale* and *spread* of each local finite-section statistic change with π .

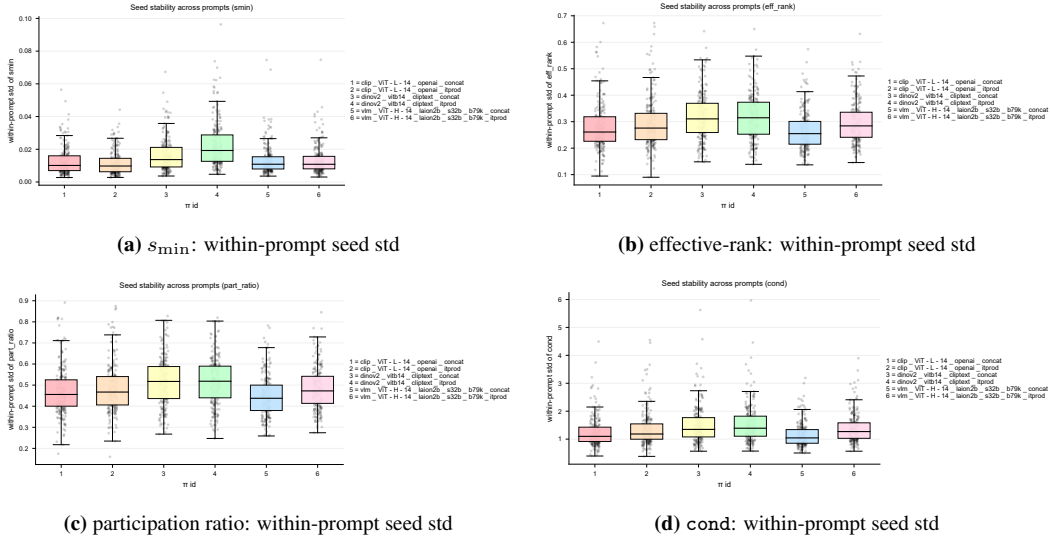


Figure 6: Seed stability of geometry diagnostics across π . Each panel shows the distribution across prompts of the within-prompt standard deviation over seeds for the indicated metric. This quantifies how much seed variation affects the diagnostic at the prompt level, and how that sensitivity varies with the semantic map.

how geometric diagnostics interact with downstream scoring. Since Δ depends on the evaluator, its relationship with a particular semantic representation can vary across π . This is consistent with our interpretation: the geometry-only diagnostics capture local latent sensitivity, while downstream score prediction also depends on evaluator alignment.

Scope of interpretation. The diagnostics are randomized local sketches, not full measurements of the global Jacobian spectrum. They are most informative when the semantic map provides a meaningful representation of the prompt-relevant attributes. Prompts involving fine-grained counting, exact attribute conjunctions, or text-heavy content can exhibit higher seed sensitivity, which is reflected in larger within-prompt variability. Similarly, agreement across π is strongest for well-aligned, high-capacity semantic maps. These observations clarify the intended scope of the diagnostics and support their use as practical tools for analyzing latent-space anisotropy, while showing that the qualitative behavior is not tied to a single semantic representation in this Z-Image study.

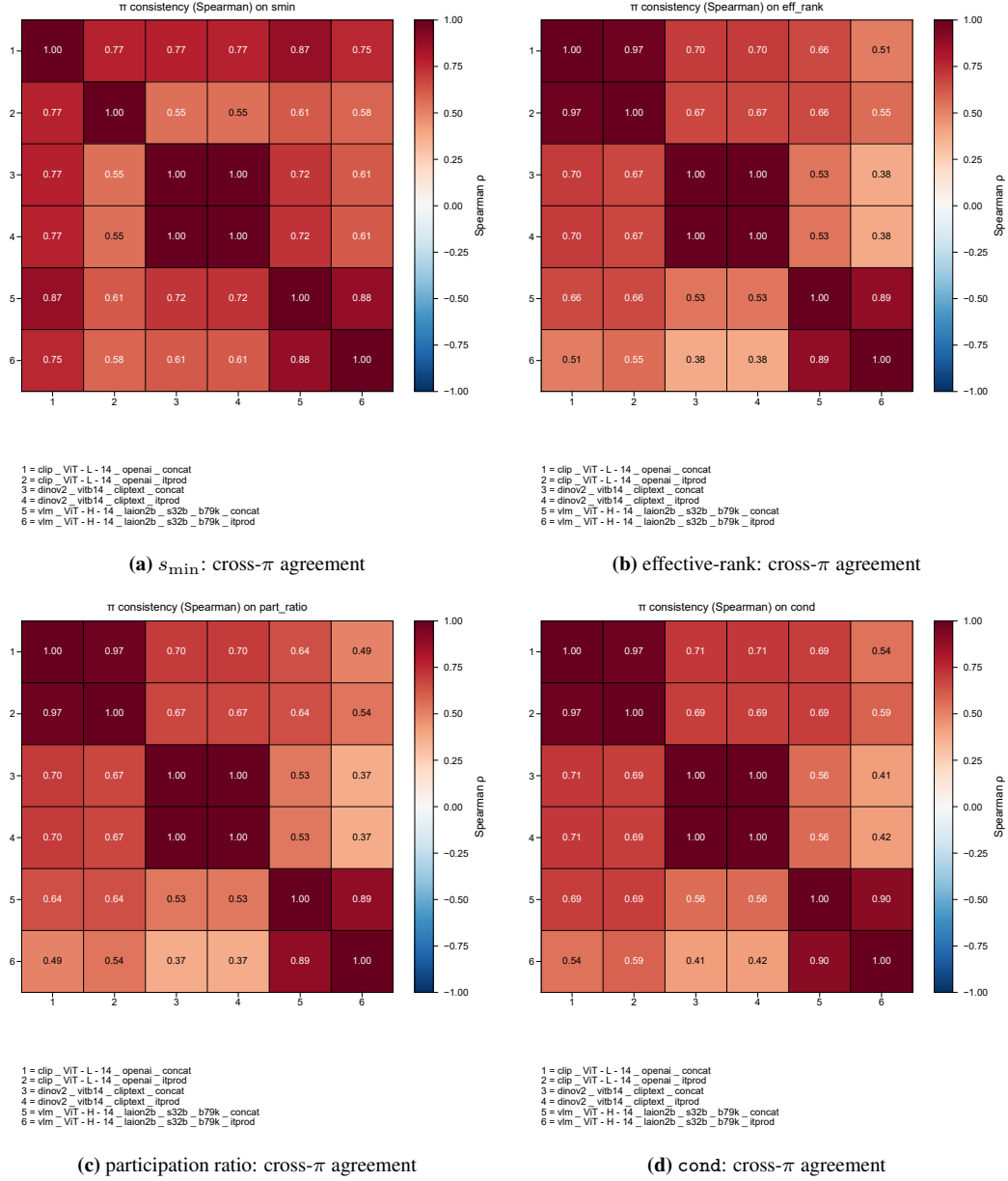


Figure 7: Pairwise Spearman agreement across π using prompt-level means. Each heatmap reports promptwise rank correlation between two choices of π for the same diagnostic statistic. Positive agreement supports the use of the diagnostics for prompt stratification without relying on a single semantic representation.

Takeaway. Across six strong and diverse semantic maps π , the geometry-only diagnostics remain well-defined and practically usable. Within the probed $r = 6$ tangent section, the local Jacobian-spectrum statistics show stable finite-section structure across semantic representations, enabling prompt stratification without committing to any single embedding choice. Downstream geometry- Δ correlations vary with π as expected because Δ is evaluator-dependent, but this does not affect the core geometry-only conclusion.

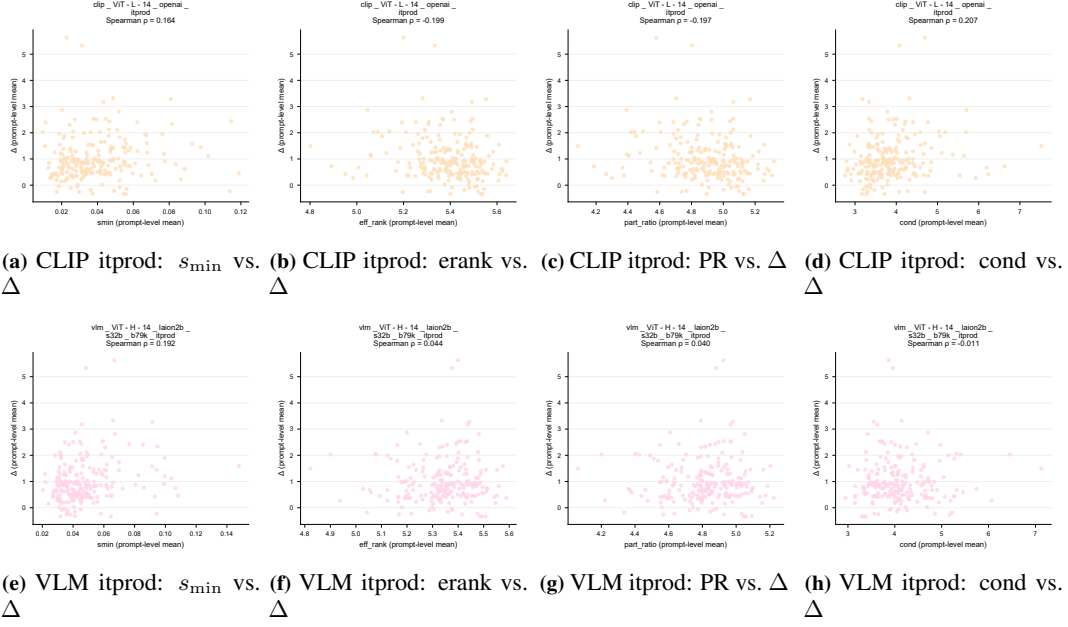
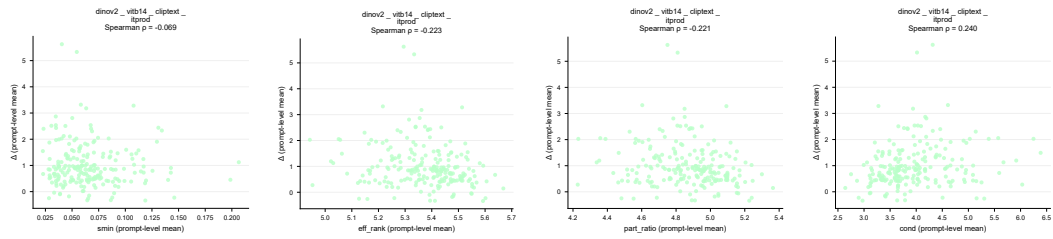


Figure 8: Geometry- Δ correlations depend on evaluator alignment. Each scatter uses prompt-level means and reports Spearman correlation in the title. Variability across π is expected because Δ is evaluator-dependent; the geometry-only conclusions rely instead on prompt stratification and finite-section spectrum shape.

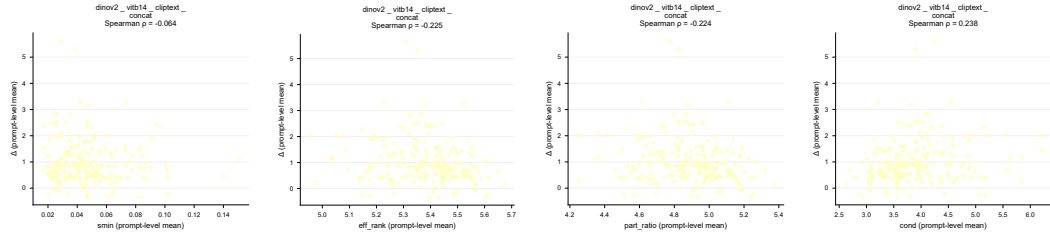
Table 5: Prompt-averaged geometry diagnostics across π (this run). Mean and 95% bootstrap CI are computed across prompts using prompt-level means, with seeds aggregated per prompt. Seed stability is summarized by the mean within-prompt standard deviation across seeds. All spectrum statistics are computed from the $r = 6$ finite-difference Jacobian sketch.

π (name)	$s_{\min}$ mean [95% CI]	$erank$ mean [95% CI]	PR mean [95% CI]	$cond$ mean [95% CI]
clip_ViT-L-14_itprod	0.03965 [0.03703, 0.04239]	5.3866 [5.3674, 5.4051]	4.8914 [4.8619, 4.9202]	3.8715 [3.7811, 3.9691]
clip_ViT-L-14_concat	0.04267 [0.04020, 0.04530]	5.4439 [5.4269, 5.4600]	4.9820 [4.9550, 5.0080]	3.5917 [3.5167, 3.6722]
vlm_ViT-H-14_itprod	0.04799 [0.04547, 0.05066]	5.3393 [5.3217, 5.3559]	4.8153 [4.7881, 4.8411]	4.0897 [4.0105, 4.1768]
vlm_ViT-H-14_concat	0.04917 [0.04682, 0.05160]	5.4338 [5.4206, 5.4462]	4.9647 [4.9433, 4.9846]	3.6332 [3.5763, 3.6945]
dinov2_vitb14_itprod	0.06581 [0.06177, 0.06997]	5.3703 [5.3519, 5.3889]	4.8652 [4.8362, 4.8941]	3.9588 [3.8689, 4.0507]
dinov2_vitb14_concat	0.04712 [0.04416, 0.05012]	5.3813 [5.3628, 5.3998]	4.8816 [4.8520, 4.9107]	3.8933 [3.8036, 3.9844]

Seed stability (mean within-prompt std across seeds). $s_{\min}$: {0.0116, 0.0128, 0.0134, 0.0135, 0.0163, 0.0228}; $erank$: {0.2818–0.3195}; PR: {0.4465–0.5166}; $cond$: {1.1624–1.5354}.



(a) DINO itprod: $s_{\min}$ vs. Δ . (b) DINO itprod: $erank$ vs. Δ . (c) DINO itprod: PR vs. Δ . (d) DINO itprod: cond vs. Δ .



(e) DINO concat: $s_{\min}$ vs. Δ . (f) DINO concat: $erank$ vs. Δ . (g) DINO concat: PR vs. Δ . (h) DINO concat: cond vs. Δ .

Figure 9: Geometry- Δ correlations for DINO-based π . Same as Fig. 8, but for DINOv2 image features fused with CLIP-text. These plots are included for completeness; variability across π is expected because Δ is evaluator-dependent.

C Additional Proofs for Preliminaries and Method

C.1 Proof of Theorem 3.2: Degenerate geometry

Proof. Let $J := D\Phi_z$ and let $G_{\mathcal{M}}(\Phi(z))$ be positive definite in a local chart. For any $w \in \mathcal{V}_z = \text{Ker}(J)$,

$$\|w\|_{g_{\text{lat}}}^2 = \langle Jw, Jw \rangle_{g_{\mathcal{M}}} = 0.$$

Conversely, if $\|u\|_{g_{\text{lat}}} = 0$, positive definiteness of $g_{\mathcal{M}}$ implies $Ju = 0$, hence $u \in \mathcal{V}_z$. Thus the nullspace of g_{lat} is exactly $\mathcal{V}_z$. Since $\mathcal{H}_z = \mathcal{V}_z^\perp$, no nonzero vector in $\mathcal{H}_z$ lies in this nullspace; hence g_{lat} is positive definite on $\mathcal{H}_z$. In matrix form, $G_{\text{lat}}(z) = J^\top G_{\mathcal{M}} J$. Because multiplication by the invertible matrix $G_{\mathcal{M}}^{1/2}$ does not change rank or kernel after composition with J , we have

$$\text{Ker}(G_{\text{lat}}(z)) = \text{Ker}(J) = \mathcal{V}_z, \quad \text{rank}(G_{\text{lat}}(z)) = \text{rank}(J) = k$$

on the regular set. $\square$

C.2 Proof of Lemma 3.3: Local semantic sensitivity

Proof. Work in a local semantic chart around $\Phi(z)$ and write $J := D\Phi_z$. Since Φ is C^2 and J is locally Lipschitz, Taylor's theorem gives, for sufficiently small Δz ,

$$\Phi(z + \Delta z) - \Phi(z) = J\Delta z + R(\Delta z), \quad \|R(\Delta z)\| \leq C\|\Delta z\|^2.$$

Decompose $\Delta z = \Delta z_H + \Delta z_V$ with $\Delta z_H = \text{Proj}_{\mathcal{H}_z} \Delta z$ and $\Delta z_V = \text{Proj}_{\mathcal{V}_z} \Delta z$. Since $J\Delta z_V = 0$, the linear term is $J\Delta z_H$. On $\mathcal{H}_z$, the nonzero singular values of J satisfy

$$\sigma_k(z)\|\Delta z_H\| \leq \|J\Delta z_H\| \leq \sigma_1(z)\|\Delta z_H\|.$$

Local equivalence between the chart norm and the manifold distance $d_{\mathcal{M}}$ gives constants $c_1, c_2 > 0$, yielding the stated upper and lower bounds after absorbing Taylor constants into C . If $\|\Delta z_H\| \geq \beta\|\Delta z\|$, then the quadratic term is dominated by the linear term for sufficiently small $\|\Delta z\|$, which gives Eq. (11). $\square$

C.3 Proof of Theorem 3.4: Coarea density transport

Proof. We derive the stated density formula directly from the coarea formula.

Step 1: Coarea formula. Let $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}^k$ be locally Lipschitz with $k \leq d$. Denote by $J_k(\Phi, z)$ the k -dimensional Jacobian (the product of the top k singular values of $D\Phi_z$ on the regular set). The coarea formula states that for any nonnegative measurable function h ,

$$\int_{\mathbb{R}^d} h(z) J_k(\Phi, z) dz = \int_{\mathbb{R}^k} \left(\int_{\Phi^{-1}(x)} h(z) d\mathcal{H}^{d-k}(z) \right) dx, \quad (43)$$

where $\mathcal{H}^{d-k}$ is $(d-k)$ -dimensional Hausdorff measure on the level set $\Phi^{-1}(x)$.

In our setting $\Phi : \mathcal{Z} \rightarrow \mathcal{M}$ maps into a manifold, but locally in a chart φ we can apply (43) to $\tilde{\Phi} = \varphi \circ \Phi$; this yields the same expression up to chart Jacobian factors, which are absorbed into the density on $\mathcal{M}$. The theorem statement is the intrinsic version of this local expression.

Step 2: Express the pushforward probability using an indicator. Let $Z \sim p_{\mathcal{Z}}$ with density w.r.t. Lebesgue measure on $\mathcal{Z} \subset \mathbb{R}^d$. Let $X = \Phi(Z)$. For any measurable set $A \subset \mathcal{M}$,

$$\mathbb{P}(X \in A) = \mathbb{P}(\Phi(Z) \in A) = \int_{\mathcal{Z}} \mathbf{1}_A(\Phi(z)) p_{\mathcal{Z}}(z) dz. \quad (44)$$

Step 3: Insert $J_k(\Phi, z)$ and apply coarea. Rewrite the integrand as

$$\mathbf{1}_A(\Phi(z)) p_{\mathcal{Z}}(z) = \left(\mathbf{1}_A(\Phi(z)) \frac{p_{\mathcal{Z}}(z)}{J_k(\Phi, z)} \right) J_k(\Phi, z),$$

on the regular set where $J_k(\Phi, z) > 0$ (the singular set is negligible by assumpt submersion). Define

$$h(z) := \mathbf{1}_A(\Phi(z)) \frac{p_{\mathcal{Z}}(z)}{J_k(\Phi, z)}.$$

Then (44) becomes

$$\mathbb{P}(X \in A) = \int_{\mathcal{Z}} h(z) J_k(\Phi, z) dz.$$

Applying the coarea formula (43) (in local coordinates) gives

$$\begin{aligned} \mathbb{P}(X \in A) &= \int_{\mathcal{M}} \left(\int_{\Phi^{-1}(x)} h(z) d\mathcal{H}^{d-k}(z) \right) d\mathcal{H}^k(x) \\ &= \int_{\mathcal{M}} \mathbf{1}_A(x) \left(\int_{\Phi^{-1}(x)} \frac{p_{\mathcal{Z}}(z)}{J_k(\Phi, z)} d\mathcal{H}^{d-k}(z) \right) d\mathcal{H}^k(x). \end{aligned} \quad (45)$$

Step 4: Read off the density. Since (45) holds for all measurable A , the term in parentheses must be the density $p_{\mathcal{M}}(x)$ (w.r.t. k -dimensional Hausdorff/Riemannian volume measure on $\mathcal{M}$):

$$p_{\mathcal{M}}(x) = \int_{\Phi^{-1}(x)} \frac{p_{\mathcal{Z}}(z)}{J_k(\Phi, z)} d\mathcal{H}^{d-k}(z),$$

which is exactly (13). $\square$

C.4 Proof of Corollary 3.5: Local “compression energy” surrogate

Proof. Start from Theorem 3.4, which implies that along a fiber point z contributes to $p_{\mathcal{M}}(x)$ with weight

$$\frac{p_{\mathcal{Z}}(z)}{J_k(\Phi, z)}.$$

Taking negative log yields a local “energy”:

$$-\log \left(\frac{p_{\mathcal{Z}}(z)}{J_k(\Phi, z)} \right) = -\log p_{\mathcal{Z}}(z) + \log J_k(\Phi, z).$$

Gaussian prior term. For $p_{\mathcal{Z}}(z) = \mathcal{N}(0, I_d)$,

$$p_{\mathcal{Z}}(z) = (2\pi)^{-d/2} \exp \left(-\frac{1}{2} \|z\|^2 \right),$$

so

$$-\log p_{\mathcal{Z}}(z) = \frac{1}{2} \|z\|^2 + \text{const.}$$

Jacobian term and pseudodeterminant. On the regular set $\text{rank}(J_z) = k$, the k -Jacobian satisfies

$$J_k(\Phi, z) = \prod_{i=1}^k \sigma_i(z),$$

where $\{\sigma_i(z)\}_{i=1}^k$ are the nonzero singular values of J_z in local coordinates. Equivalently,

$$J_k(\Phi, z)^2 = \det(J_z J_z^\top).$$

Since $J_z J_z^\top \in \mathbb{R}^{k \times k}$ is positive definite at regular points, its determinant equals the product of its (nonzero) eigenvalues; in invariant notation this is the pseudodeterminant $\text{pdet}(J_z J_z^\top)$, and here $\text{pdet} = \det$. Hence

$$\log J_k(\Phi, z) = \frac{1}{2} \log(J_k(\Phi, z)^2) = \frac{1}{2} \log \det(J_z J_z^\top) = \frac{1}{2} \log \text{pdet}(J_z J_z^\top).$$

Putting the two terms together yields

$$-\log \left(\frac{p_{\mathcal{Z}}(z)}{J_k(\Phi, z)} \right) = \frac{1}{2} \|z\|^2 + \frac{1}{2} \log \text{pdet}(J_z J_z^\top) + \text{const.},$$

which is exactly (14). $\square$

C.5 Derivation of the score-factorized bridge

This subsection makes explicit how the prompt residual used by the algorithm relates to the horizontal-vertical geometry. The goal is not to prove that every pretrained generator exactly materializes the true horizontal space, but to identify a sufficient score-factorization condition under which the residual has a horizontal component, and to state precisely what the remaining terms represent.

From conditional scores to prompt residuals. For an ϵ -parameterized diffusion model with noisy latent $z_t = \alpha_t z_0 + \sigma_t \epsilon$, the learned noise predictor determines the conditional score through

$$s_\theta(z_t, t, y) := \nabla_{z_t} \log p_\theta(z_t | y) \approx -\frac{1}{\sigma_t} \epsilon_\theta(z_t, t, y), \quad (46)$$

up to the usual approximation error of the learned score model. By Bayes' rule,

$$\nabla_{z_t} \log p_\theta(y | z_t) = \nabla_{z_t} \log p_\theta(z_t | y) - \nabla_{z_t} \log p_\theta(z_t) \quad (47)$$

$$= s_\theta(z_t, t, y) - s_\theta(z_t, t, \emptyset), \quad (48)$$

where the unconditional branch is used as the model prior score. Combining this identity with Eq. (46) gives

$$\epsilon_\theta(z_t, t, y) - \epsilon_\theta(z_t, t, \emptyset) \approx -\sigma_t \nabla_{z_t} \log p_\theta(y | z_t). \quad (49)$$

Thus, for ϵ -prediction, the prompt residual is a time-scaled prompt-likelihood gradient. The scalar factor is immaterial after normalization; the sign convention is fixed by defining the local potential ℓ_y so that the semantic component of the implemented residual has positive coefficient $c_t > 0$.

Semantic factorization implies horizontal gradients. Assume first that the prompt likelihood depends on the latent only through the semantic map,

$$\log p_\theta(y | z) = \ell_y(\Phi(z)). \quad (50)$$

By the chain rule, with $J_z := D\Phi_z$,

$$\nabla_z \log p_\theta(y | z) = J_z^\top \nabla_\Phi \ell_y(\Phi(z)) \in \text{Im}(J_z^\top). \quad (51)$$

At a regular point, $\text{Im}(J_z^\top) = \text{Ker}(J_z)^\perp = \mathcal{H}_z$, so a purely semantic prompt-likelihood gradient is horizontal. This is the clean mathematical case behind the proxy construction.

Relaxed factorization and Eq. (24). In practice, the likelihood need not factor exactly through the chosen semantic map. Write

$$\log p_\theta(y | z) = \ell_y(\Phi(z)) + r_y(z), \quad (52)$$

where r_y captures prompt-related dependence not represented by Φ . Then

$$\nabla_z \log p_\theta(y | z) = J_z^\top \nabla_\Phi \ell_y + \nabla_z r_y. \quad (53)$$

For ϵ -prediction, substituting Eq. (53) into the score-residual identity yields

$$g_\theta = -\sigma_t J_z^\top \nabla_\Phi \ell_y - \sigma_t \nabla_z r_y = J_z^\top a_t(y) + e_t, \quad (54)$$

with $a_t(y) := -\sigma_t \nabla_\Phi \ell_y$ and $e_t := -\sigma_t \nabla_z r_y$. For velocity- or flow-parameterized backbones, there is no scheduler-independent exact scalar relating the velocity residual to the conditional score. We therefore write

$$g_\theta^{\text{flow}} = c_t \nabla_z \log p_\theta(y | z) + \xi_t, \quad (55)$$

where c_t is the effective time-dependent calibration induced by the chosen parameterization and scheduler, and ξ_t is the remaining mismatch. Combining Eq. (55) with Eq. (53) gives

$$g_\theta^{\text{flow}} = c_t J_z^\top \nabla_\Phi \ell_y + (c_t \nabla_z r_y + \xi_t) = J_z^\top a_t(y) + e_t. \quad (56)$$

This is Eq. (24); it should be read as a sufficient structural approximation, not as an unconditional property of the pretrained model. A convenient normalized measure of the flow-specific mismatch is

$$\eta_{\text{flow}} := \frac{\|\xi_t\|}{\|c_t \nabla_z \log p_\theta(y | z_t)\| + \varepsilon}. \quad (57)$$

When either $\nabla_z r_y$ or ξ_t is large, the vertical or non-semantic component of the proxy can also be large; the method therefore relies on operational validation rather than exact recovery of $\mathcal{H}_z$.

What the residual term r_y represents. The term r_y is not an arbitrary error symbol: it collects prompt-likelihood dependence that is not encoded by the semantic projection Φ . Examples include rendering artifacts, typography failures, fine-grained counting, spatial relations missed by the chosen embedding, and finite-capacity biases of the backbone. If $\nabla_z r_y$ dominated the semantic term $J_z^\top \nabla_\Phi \ell_y$, the normalized prompt residual would behave more like a non-semantic or random tangent direction under the radius-preserving probe. Conversely, for a small spherical perturbation $z(\theta; u)$,

$$D_\pi(z, u; \theta) = \theta \|J_z u\| + O(\theta^2), \quad (58)$$

so vertical directions contribute only at second order. The semantic-displacement percentiles in Table 6 therefore provide an indirect constraint: they do not prove $\|\nabla_z r_y\|$ is small or that the vertical component vanishes, but they show that the normalized prompt residual has a semantic-active component far stronger than random tangent controls. This operational evidence is sufficient for the seed-shaping objective, which only requires a useful tangent direction rather than a certified horizontal projection.

C.6 Proof of Lemma 4.2: Shell-projected proxy compatibility

Proof. Let $J := D\Phi_z$ and $Q := Q_z$. Projecting Eq. (24) onto the shell tangent space gives

$$Qg_\theta = c_t QJ^\top \nabla_\Phi \ell_y + Qe_t.$$

By Lemma 4.1, $QJ^\top \nabla_\Phi \ell_y \in \text{Img}(QJ^\top) = \mathcal{H}_z^{\text{sh}}$. Therefore its projection onto $\mathcal{V}_z^{\text{sh}}$ is zero, and

$$\text{Proj}_{\mathcal{V}_z^{\text{sh}}}(Qg_\theta) = \text{Proj}_{\mathcal{V}_z^{\text{sh}}}(Qe_t),$$

which gives Eq. (26). For the cone bound, write

$$Qg_\theta = s + r, \quad s := \text{Proj}_{\mathcal{H}_z^{\text{sh}}}(Qg_\theta), \quad r := \text{Proj}_{\mathcal{V}_z^{\text{sh}}}(Qg_\theta).$$

The two components are orthogonal. If $\|r\| \leq \rho \|s\|$ and $u = (s + r)/\|s + r\|$, then

$$\|\text{Proj}_{\mathcal{V}_z^{\text{sh}}} u\| = \frac{\|r\|}{\sqrt{\|s\|^2 + \|r\|^2}} \leq \frac{\rho}{\sqrt{1 + \rho^2}},$$

and the corresponding angle to $\mathcal{H}_z^{\text{sh}}$ is at most $\arctan \rho$. $\square$

C.7 Proof of Proposition 4.5: Cold-start shell-proxy transfer

Proof. By adding and subtracting $\tilde{G}_{t^*}(z^{\text{ref}})$ and $c_{t^*} \nabla_{\mathbb{S}_R} F_y(z^{\text{ref}})$, we obtain

$$\begin{aligned} & \tilde{G}_{t^*}(z_T) - c_{t^*} \nabla_{\mathbb{S}_R} F_y(z_T) \\ &= (\tilde{G}_{t^*}(z^{\text{ref}}) - c_{t^*} \nabla_{\mathbb{S}_R} F_y(z^{\text{ref}})) + (\tilde{G}_{t^*}(z_T) - \tilde{G}_{t^*}(z^{\text{ref}})) - c_{t^*} (\nabla_{\mathbb{S}_R} F_y(z_T) - \nabla_{\mathbb{S}_R} F_y(z^{\text{ref}})). \end{aligned}$$

Taking norms and using the Lipschitz assumptions gives

$$\|r_T\| \leq \|r_{\text{ref}}\| + (L_G^{\text{sh}} + c_{t^*} L_F^{\text{sh}}) \|z_T - z^{\text{ref}}\|.$$

If the reference discrepancy is controlled by Proposition 4.4, the last term is $\mathcal{O}(T - t^*)$. $\square$

C.8 Proof of Theorem 4.6: Local algorithmic bridge

Proof. Let $f := \nabla_{\mathbb{S}_R} F_y(z_T)$, $c := c_{t^*}$, and $r := r_T$. With exact normalization, the update direction is

$$u_0 = \frac{cf + r}{\|cf + r\|}.$$

The cone assumption gives $\|r\| \leq \rho_c c \|f\|$. Hence

$$\langle f, cf + r \rangle \geq c \|f\|^2 - \|f\| \|r\| \geq c(1 - \rho_c) \|f\|^2,$$

and

$$\|cf + r\| \leq c(1 + \rho_c) \|f\|.$$

Therefore

$$\langle f, u_0 \rangle \geq \frac{1 - \rho_c}{1 + \rho_c} \|f\|. \quad (59)$$

If the ε -stabilized normalization is used and $\|\bar{g}_T\| \geq m_g > 0$ locally, then the direction changes by $O(\varepsilon/m_g)$, producing the $O(\delta\varepsilon/m_g)$ term. With exact normalization this term is absent.

Because $u_0 \perp z_T$ and $\|u_0\| = 1$, the spherical retraction satisfies

$$z_T^* = z_T + \delta u_0 - \frac{\delta^2}{2\|z_T\|^2} z_T + O\left(\frac{\delta^3}{\|z_T\|^2}\right).$$

Taylor expanding F_y around z_T gives

$$F_y(z_T^*) - F_y(z_T) = \delta \langle \nabla_{\mathbb{S}_R} F_y(z_T), u_0 \rangle + O(\delta^2).$$

Combining this expansion with Eq. (59) proves Eq. (35). The statement about cold-start error follows by substituting the residual bound of Proposition 4.5 into the cone residual term. $\square$

C.9 Proof of Proposition 4.7: Retraction as projection onto the sphere

Proof. Let $R := \|z_T\|$ and $u := z_T + \delta \hat{v}$. We need to solve

$$\min_{z \in \mathbb{R}^d} \|z - u\|^2 \quad \text{s.t.} \quad \|z\| = R.$$

Method 1: Geometry (maximize inner product). Expand:

$$\|z - u\|^2 = \|z\|^2 + \|u\|^2 - 2\langle z, u \rangle = R^2 + \|u\|^2 - 2\langle z, u \rangle.$$

Since $R^2 + \|u\|^2$ is constant under the constraint, minimizing $\|z - u\|^2$ is equivalent to maximizing $\langle z, u \rangle$ subject to $\|z\| = R$. By Cauchy-Schwarz, $\langle z, u \rangle \leq \|z\|\|u\| = R\|u\|$, with equality iff z is a positive scalar multiple of u . Thus the maximizer is

$$z^* = R \frac{u}{\|u\|} = \|z_T\| \cdot \frac{z_T + \delta \hat{v}}{\|z_T + \delta \hat{v}\|},$$

Method 2: Lagrange multipliers. Consider $\mathcal{L}(z, \lambda) = \|z - u\|^2 + \lambda(\|z\|^2 - R^2)$. Stationarity in z gives $2(z - u) + 2\lambda z = 0$, i.e. $(1 + \lambda)z = u$. Thus $z = \frac{1}{1+\lambda}u$. Enforce $\|z\| = R$ to get $|1 + \lambda| = \|u\|/R$ and choose the sign giving the minimizer (same direction as u), yielding $z^* = Ru/\|u\|$. $\square$

C.10 Proof of Lemma 4.8: First-order shell retraction

Proof. Let $R = \|z_T\|$, $\hat{z} = z_T/R$, and let $u \perp z_T$ with $\|u\| = 1$. Then

$$\|z_T + \delta u\| = \sqrt{R^2 + \delta^2} = R\sqrt{1 + \delta^2/R^2}.$$

Therefore

$$z_T^+ = R \frac{z_T + \delta u}{\sqrt{R^2 + \delta^2}} = (z_T + \delta u) \left(1 + \frac{\delta^2}{R^2}\right)^{-1/2}.$$

Using $(1 + s)^{-1/2} = 1 - s/2 + O(s^2)$ gives

$$z_T^+ = z_T + \delta u - \frac{\delta^2}{2R^2} z_T + O\left(\frac{\delta^3}{R^2}\right),$$

where the remainder includes the product of δu with $O(\delta^2/R^2)$. Exact radius preservation follows directly from the definition. Finally, for any C^2 function F_y , Taylor's theorem gives

$$F_y(z_T^+) - F_y(z_T) = \langle \nabla F_y(z_T), z_T^+ - z_T \rangle + O(\|z_T^+ - z_T\|^2).$$

The first-order displacement is the tangent vector δu , and $u \in T_{z_T} \mathbb{S}_R$. Thus $\langle \nabla F_y(z_T), u \rangle = \langle \nabla_{\mathbb{S}_R} F_y(z_T), u \rangle$, while all remaining terms are $O(\delta^2)$. $\square$

C.11 Proof of Proposition 4.4: Cold-start state deviation

Proof. We use a reverse-time reparameterization and Grönwall's inequality in integral form to avoid differentiating $\|e_\tau\|$ at $\tau = 0$.

Step 1: Reverse-time reparameterization. Set $\tau := T - t$, and define

$$\tilde{z}_\tau := z_{T-\tau}^{\text{traj}}, \quad \tilde{z}_\tau^{\text{cs}} := \rho(T - \tau)z_T, \quad \tau \in [0, h].$$

Then $\tilde{z}_0 = \tilde{z}_0^{\text{cs}} = z_T$, and the chain rule gives

$$\frac{d\tilde{z}_\tau}{d\tau} = -v_\theta(\tilde{z}_\tau, T - \tau; y), \quad \frac{d\tilde{z}_\tau^{\text{cs}}}{d\tau} = -\rho'(T - \tau)z_T.$$

Step 2: Integral form of the deviation. Let $e_\tau := \tilde{z}_\tau - \tilde{z}_\tau^{\text{cs}}$. Since $e_0 = 0$, the fundamental theorem of calculus yields

$$e_\tau = \int_0^\tau \frac{de_s}{ds} ds = \int_0^\tau [-v_\theta(\tilde{z}_s, T - s; y) + \rho'(T - s)z_T] ds.$$

Adding and subtracting $v_\theta(\tilde{z}_s^{\text{cs}}, T - s; y)$ inside the integrand and applying the triangle inequality:

$$\begin{aligned} \|e_\tau\| &\leq \int_0^\tau \|v_\theta(\tilde{z}_s, T - s; y) - v_\theta(\tilde{z}_s^{\text{cs}}, T - s; y)\| ds \\ &\quad + \int_0^\tau \|v_\theta(\tilde{z}_s^{\text{cs}}, T - s; y) - \rho'(T - s)z_T\| ds. \end{aligned}$$

Step 3: Apply the Lipschitz and boundedness assumptions. By the L_v -Lipschitz assumption on $v_\theta(\cdot, t; y)$,

$$\|v_\theta(\tilde{z}_s, T - s; y) - v_\theta(\tilde{z}_s^{\text{cs}}, T - s; y)\| \leq L_v \|e_s\|.$$

By the boundedness assumptions $\|v_\theta\| \leq B_v$ and $|\rho'| \leq L_\rho$,

$$\|v_\theta(\tilde{z}_s^{\text{cs}}, T - s; y) - \rho'(T - s)z_T\| \leq B_v + L_\rho \|z_T\|.$$

Substituting:

$$\|e_\tau\| \leq L_v \int_0^\tau \|e_s\| ds + (B_v + L_\rho \|z_T\|) \tau.$$

Step 4: Grönwall's inequality (integral form). The inequality $\|e_\tau\| \leq \alpha(\tau) + L_v \int_0^\tau \|e_s\| ds$, with $\alpha(\tau) = (B_v + L_\rho \|z_T\|)\tau$, gives by the Grönwall–Bellman inequality

$$\|e_\tau\| \leq \alpha(\tau) + L_v \int_0^\tau \alpha(s) e^{L_v(\tau-s)} ds.$$

Since

$$L_v \int_0^\tau s e^{L_v(\tau-s)} ds = \frac{e^{L_v\tau} - 1}{L_v} - \tau,$$

we obtain

$$\|e_\tau\| \leq (B_v + L_\rho \|z_T\|) \left[\tau + \frac{e^{L_v\tau} - 1}{L_v} - \tau \right] = (B_v + L_\rho \|z_T\|) \frac{e^{L_v\tau} - 1}{L_v}.$$

Step 5: Conclude. Setting $\tau = h$ gives Eq. (30); the case $L_v = 0$ follows from Taylor expansion of $e^{L_v h}$. If $\rho \equiv 1$ on $[t^*, T]$, the direct bound $\|z_{t^*}^{\text{traj}} - z_T\| \leq \int_{t^*}^T \|v_\theta(z_t, t; y)\| dt \leq B_v(T - t^*)$ also applies. $\square$

D Additional Experiments: Visualizing Degenerate Latent Geometry

This appendix details the empirical diagnostics used to visualize the pullback geometry induced by the semantic map $\Phi = \pi \circ \Psi$. Our theory predicts that semantic sensitivity is anisotropic on the Gaussian typical-radius shell: most tangent directions induce small semantic changes, while a smaller set of prompt-relevant directions produces stronger semantic movement. We verify this through three complementary diagnostics: direction-sensitivity distributions, local spectrum estimation, and prompt-residual proxy comparisons.

D.1 Experimental Setup and Notation

Backbone and sampler. We use FLUX.1-Dev with the official inference configuration in Appendix G. All probes modify only the initial seed while keeping the prompt and sampler fixed.

Prompts and seeds. We evaluate complex prompts involving multi-object composition, typography, and spatial relations, with multiple random seeds per prompt. Reported curves and distributions aggregate over prompts and seeds, while strip visualizations show representative examples.

Semantic features and metrics. Let $x(z_T)$ denote the generated output from initial noise z_T under prompt y . We use CLIP image/text embeddings $e_I(x)$ and $e_T(y)$, normalized as $\hat{e}_I$ and $\hat{e}_T$.

$$\text{CLIPSim}(x, y) = \langle \hat{e}_I(x), \hat{e}_T(y) \rangle. \quad (60)$$

For a baseline seed z_T and perturbed seed z'_T , we measure semantic displacement by

$$\|\Delta\text{CLIP}\| = \|\hat{e}_I(x(z'_T)) - \hat{e}_I(x(z_T))\|_2. \quad (61)$$

D.2 Typical-Radius Shell Probing

To probe sensitivity while staying on the Gaussian typical-radius shell, we fix $R = \|z_T\|$. Let $\hat{z} = z_T/\|z_T\|$. A random tangent direction is obtained by

$$\tilde{u} := u - \langle u, \hat{z} \rangle \hat{z}, \quad \hat{u} := \tilde{u} / \|\tilde{u}\|. \quad (62)$$

We then use the radius-preserving spherical geodesic

$$z_T(\theta) = R(\cos \theta \hat{z} + \sin \theta \hat{u}), \quad (63)$$

which satisfies $\|z_T(\theta)\| = R$ for all θ . For small δ/R , this is equivalent to a tangent step followed by spherical retraction,

$$z'_T = R \cdot \frac{z_T + \delta \hat{u}}{\|z_T + \delta \hat{u}\|}, \quad (64)$$

matching the retraction used in our method.

D.3 Direction Sensitivity Distribution

For each prompt–seed pair, we sample M random tangent directions, perturb each seed by a small angle θ_0 , and compute $\|\Delta\text{CLIP}\|$. We aggregate the resulting semantic displacements and visualize both the histogram and CCDF.

For small θ_0 , the semantic change follows the local linear approximation

$$\Phi(z_T(\theta_0)) - \Phi(z_T) \approx D\Phi_{z_T}(R\theta_0 \hat{u}). \quad (65)$$

Thus, the distribution of $\|\Delta\text{CLIP}\|$ across random tangent directions reveals the anisotropy of semantic sensitivity.

D.4 Local Spectrum Estimation

We further estimate the effective local spectrum of semantic sensitivity. Let $\phi(z_T) := \hat{e}_I(x(z_T))$ be the normalized CLIP image embedding. For each tangent direction $\hat{u}_j$, we compute the symmetric geodesic finite difference

$$g_j = \frac{\phi(z_T(\theta_0; \hat{u}_j)) - \phi(z_T(-\theta_0; \hat{u}_j))}{2\theta_0} \in \mathbb{R}^D. \quad (66)$$

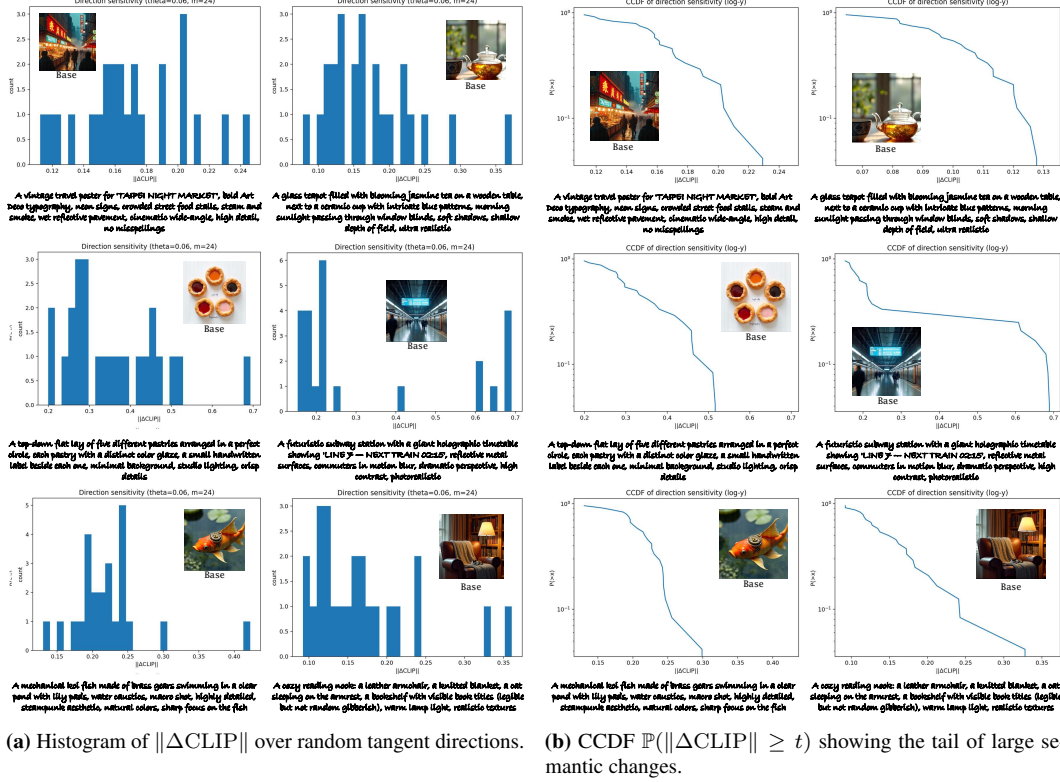


Figure 10: Anisotropic semantic sensitivity on the Gaussian radius shell. Most tangent directions induce small semantic displacement, while a subset produces strong semantic changes, consistent with the pullback-geometry view.

Stacking g_j as columns of $G = [g_1, \dots, g_M]$, the spectrum of $G^\top G$ provides a randomized estimate of the dominant local sensitivity modes.

A concentrated spectrum indicates that semantic changes are governed by a small set of dominant directions. Across prompts, the spectral shape also reflects conditioning-dependent structure, especially for typography, layout, and compositional prompts.

D.5 Prompt-Residual Proxy vs. Control Directions

Since $D\Phi$ is not explicitly available during sampling, we use the model-derived prompt residual as an operational proxy direction. At probing timestep t , we compute

$$r := \epsilon_\theta(z_t, t, y) - \epsilon_\theta(z_t, t, \emptyset), \quad (67)$$

normalize it, and project it to the tangent space:

$$\hat{u}_H = \frac{\hat{r} - \langle \hat{r}, \hat{z} \rangle \hat{z}}{\|\hat{r} - \langle \hat{r}, \hat{z} \rangle \hat{z}\|}. \quad (68)$$

We compare this prompt-residual proxy with an independently sampled random tangent control $\hat{u}_C$ under matched radius-preserving perturbations.

D.6 Operational Diagnostics for the Prompt-Residual Proxy

We evaluate whether the prompt-residual proxy behaves as a structured tangent direction on FLUX.1-Dev.

Semantic-displacement percentile. For each matched prompt–seed pair, we compare the proxy direction against random tangent controls. Table 6 shows that the proxy lies in the upper tail of random controls for CLIP displacement metrics.

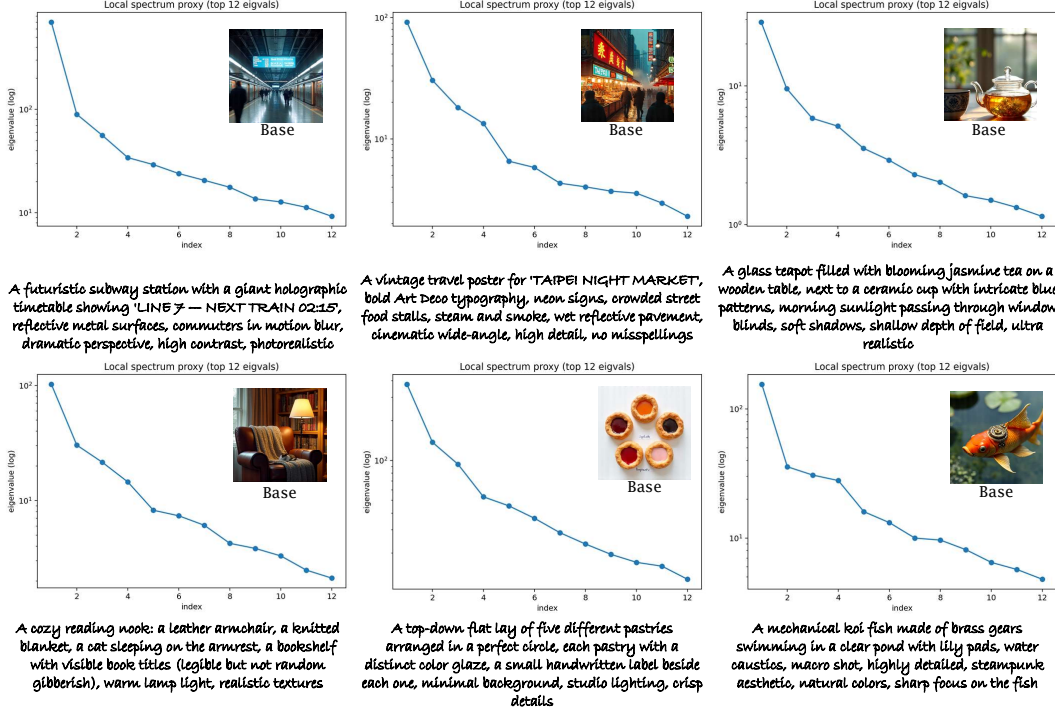


Figure 11: Local spectrum proxy. The spectrum summarizes how semantic sensitivity concentrates across dominant tangent directions.

Table 6: Semantic-displacement percentile of the prompt-residual proxy on FLUX. Percentiles are computed over 48 matched prompt-seed pairs against random tangent controls.

Metric	Mean percentile $\uparrow$	Median percentile $\uparrow$	Top-2 rate $\uparrow$
CLIP L2 displacement	0.995	1.000	1.000
CLIP cosine displacement	0.997	1.000	1.000
Alignment absolute change	0.922	1.000	0.854
Alignment signed change	0.529	0.578	0.438

Residual concentration. For local prompt variants at $t^* = 900$, Table 7 shows that prompt-residual directions form a concentrated family compared with random tangent controls. The tangential projection retains nearly all residual norm, with mean radial ratio 0.066 and retained tangent norm ratio 0.997.

High-noise direction consistency. We compare prompt-residual directions at different high-noise timesteps with the default $t^* = 900$ direction. Table 8 shows consistent alignment above the random tangent baseline around the default probe. Across these timesteps, the retained tangent norm ratio remains high, ranging from 0.998 to 1.000.

Together, these diagnostics show that the prompt-residual proxy is structured, concentrated, and semantically active on the Gaussian radius shell.

D.7 Geodesic Curves and Strip Visualizations

For each (y, z_T) , we evaluate a grid of angles θ along both the prompt-residual proxy direction $\hat{u}_H$ and the random control direction $\hat{u}_C$. We report semantic displacement $\|\Delta_{\text{CLIP}}\|(\theta)$ and alignment $\text{CLIPSim}(\theta)$.

The main signature is an early-slope gap: semantic displacement increases faster along $\hat{u}_H$ than along $\hat{u}_C$. Alignment curves further show that moderate prompt-residual movement can improve prompt agreement, supporting the controlled injection strength used in our method.

Table 7: Concentration of prompt-residual directions at $t^* = 900$. Local prompt variants induce a concentrated residual family compared with random tangent controls.

Direction family	Pairwise cosine $\uparrow$	Effective rank $\downarrow$	Top-5 energy $\uparrow$
Local prompt variants	0.723	9.373	0.766
Random tangent controls	0.000	28.999	0.175

Table 8: High-noise consistency of prompt-residual directions on FLUX. Cosine similarity is computed with the $t = 900$ direction as reference.

Timestep	Cosine with $v_{900} \uparrow$	95% CI	Random baseline
1000	0.150	[0.132, 0.166]	≈ 0
950	0.461	[0.421, 0.501]	≈ 0
900	1.000	[1.000, 1.000]	≈ 0
850	0.215	[0.187, 0.243]	≈ 0
800	0.095	[0.077, 0.113]	≈ 0

We also visualize generation strips along increasing θ . Each column corresponds to one radius-preserving geodesic step, with $\theta = 0$ as the baseline.

D.8 Diagnostic Evaluation of Geometric Assumptions

We further quantify the geometry diagnostics over $N = 200$ prompts. For each prompt, we generate one baseline image and one image with our method under matched settings. We report prompt-level CLIPScore improvement

$$\Delta = \text{CLIPScore(ours)} - \text{CLIPScore(baseline)}. \quad (69)$$

All confidence intervals are bootstrap 95% intervals over prompts using 2,000 resamples.

Local Jacobian spectrum. We approximate the Jacobian of $\pi \circ G$ around the baseline initial noise z_0 using symmetric finite differences:

$$J_i \approx \frac{\pi(G(z_0 + \varepsilon u_i)) - \pi(G(z_0 - \varepsilon u_i))}{2\varepsilon}. \quad (70)$$

We use the minimum singular value $s_{\min}$ as a local geometry proxy.

Overall gains. Across 200 prompts, our method improves CLIPScore with:

- Mean improvement: $\Delta = +0.312$.
- Bootstrap 95% CI: [0.153, 0.466].
- Prompt-level win rate: 65.0%.
- Effect size: Cohen’s $d = 0.266$.

Regime analysis by $s_{\min}$. We stratify prompts into three quantile buckets by $s_{\min}$. Table 9 shows that gains remain positive across all buckets, with the largest improvement in the low- $s_{\min}$ regime.

D.9 Summary

Together, the direction-sensitivity distribution, local spectrum proxy, prompt-residual diagnostics, geodesic curves, strip visualizations, and prompt-level gain analysis provide a consistent empirical picture: under an isotropic Gaussian prior, semantic sensitivity is anisotropic and concentrated in structured directions. This supports the pullback-geometry viewpoint and motivates the prompt-residual seed shaping method in Sec. 4.

E Diversity Analysis: Alignment Gains Without Diversity Sacrifice

A potential concern for inference-time alignment improvements is that they might arise from reduced output diversity (e.g., implicitly collapsing samples to fewer modes). To rule out this explanation, we

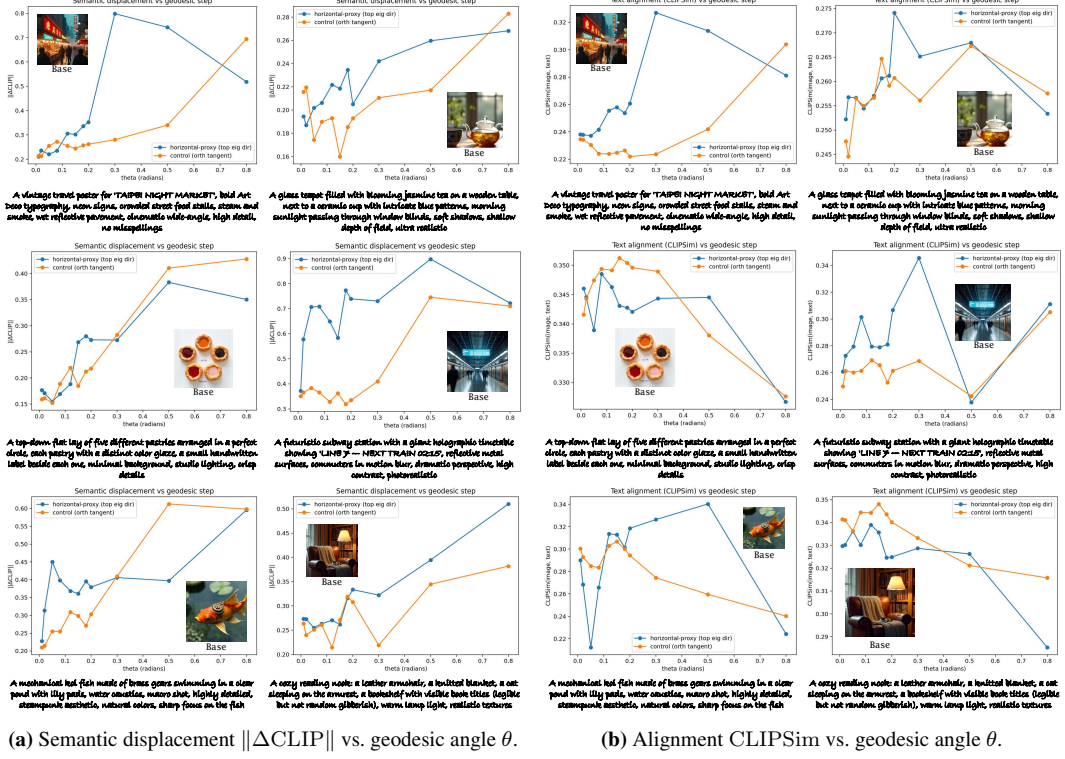


Figure 12: Prompt-residual proxy vs. control perturbations. Prompt-residual proxy directions typically yield faster semantic displacement and improved alignment at moderate angles.

Table 9: Prompt-level gains stratified by the geometry proxy $s_{\min}$. Δ denotes $\text{CLIPScore}(\text{ours}) - \text{CLIPScore}(\text{baseline})$. Confidence intervals are bootstrap 95% over prompts within each bucket.

Bucket by $s_{\min}$	# Prompts	$\mathbb{E}[\Delta]$	95% CI	Win rate
Low	67	0.424	[0.150, 0.699]	0.716
Mid	66	0.236	[-0.016, 0.484]	0.606
High	67	0.274	[-0.059, 0.573]	0.627

compare seed-to-seed diversity between the **Base** sampler (standard generation) and the **Full** method (our inference-time procedure). As shown in Fig. 15, we observe no meaningful reduction in diversity under our method, suggesting that alignment gains are *not* obtained by sacrificing diversity.

E.1 Evaluation Protocol

For each prompt, we generate n samples using n distinct random seeds under identical generation settings (model, resolution, guidance, sampler, and number of steps). For a diversity metric $d(\cdot, \cdot)$, we compute the average pairwise distance across the n samples:

$$D(\mathcal{S}) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} d(x_i, x_j), \quad (71)$$

where $\mathcal{S} = \{x_1, \dots, x_n\}$ denotes the set of samples for a fixed prompt. We then report the mean of $D(\mathcal{S})$ over all prompts.

E.2 Diversity Metrics

We use complementary metrics capturing semantic and perceptual diversity:

- **Embedding cosine distance (cosdist).** The average cosine distance in a semantic embedding space (e.g., CLIP), capturing *semantic diversity* such as changes in composition and scene semantics.



A vintage travel poster for 'TAIPEI NIGHT MARKET', bold Art Deco typography, neon signs, crowded street food stalls, steam and smoke, wet reflective pavement, cinematic wide-angle, high detail, no misspellings



A glass teapot filled with blooming jasmine tea on a wooden table, next to a ceramic cup with intricate blue patterns, morning sunlight passing through window blinds, soft shadows, shallow depth of field, ultra realistic



A top-down flat lay of five different pastries arranged in a perfect circle, each pastry with a distinct color glaze, a small handwritten label beside each one, minimal background, studio lighting, crisp details



A futuristic subway station with a giant holographic timetable showing 'LINE 3 - NEXT TRAIN 02:15', reflective metal surfaces, commuters in motion blur, dramatic perspective, high contrast, photorealistic



A cyclist in a bright yellow rain jacket riding through heavy rain at night, streetlights creating bokeh, water droplets frozen mid-air, dynamic motion, cinematic realism



A product photo of a matte black perfume bottle labeled 'NOCTURN' in silver foil, on a dark marble surface with soft reflections, controlled studio lighting, minimalist luxury aesthetic

(a) Control strip along $\hat{u}_C$.



A vintage travel poster for 'TAIPEI NIGHT MARKET', bold Art Deco typography, neon signs, crowded street food stalls, steam and smoke, wet reflective pavement, cinematic wide-angle, high detail, no misspellings



A glass teapot filled with blooming jasmine tea on a wooden table, next to a ceramic cup with intricate blue patterns, morning sunlight passing through window blinds, soft shadows, shallow depth of field, ultra realistic



A top-down flat lay of five different pastries arranged in a perfect circle, each pastry with a distinct color glaze, a small handwritten label beside each one, minimal background, studio lighting, crisp details



A futuristic subway station with a giant holographic timetable showing 'LINE 3 - NEXT TRAIN 02:15', reflective metal surfaces, commuters in motion blur, dramatic perspective, high contrast, photorealistic



A cyclist in a bright yellow rain jacket riding through heavy rain at night, streetlights creating bokeh, water droplets frozen mid-air, dynamic motion, cinematic realism



A product photo of a matte black perfume bottle labeled 'NOCTURN' in silver foil, on a dark marble surface with soft reflections, controlled studio lighting, minimalist luxury aesthetic

(b) Prompt-residual proxy strip along $\hat{u}_H$.

Figure 13: Strip visualization along spherical geodesics. Prompt-residual proxy movement induces coherent, prompt-relevant semantic changes earlier than random control directions.

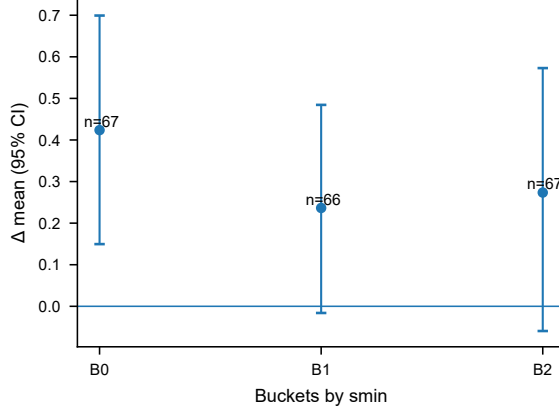


Figure 14: Bucketed analysis: mean Δ with bootstrap 95% CI across three $s_{\min}$ quantile buckets.

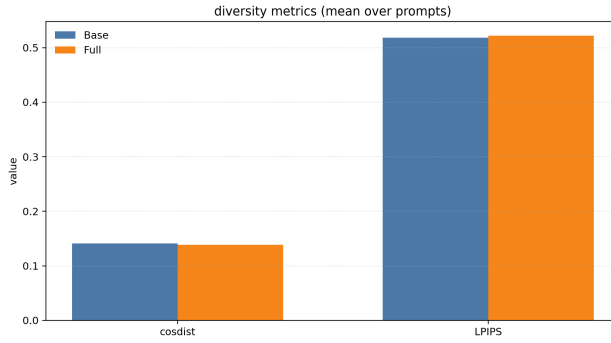


Figure 15: Diversity metrics (mean over prompts). Full improves alignment without reducing diversity: semantic diversity measured by embedding cosine distance remains essentially unchanged, while perceptual diversity measured by LPIPS is comparable or slightly higher.

- **LPIPS.** A perceptual similarity metric; higher values indicate larger perceptual differences and thus greater *appearance-level diversity*.

E.3 Results and Discussion

Figure 15 shows that diversity is preserved when switching from Base to Full:

- **Semantic diversity is unchanged.** The mean embedding cosine distance is nearly identical for Base and Full (cosdist_base vs. cosdist_full), indicating that our method does not compress variation in semantic space.
- **Perceptual diversity is maintained (or slightly increased).** LPIPS under Full is comparable to, and marginally higher than, Base (lpips_full vs. lpips_base), suggesting that outputs do not become more similar in appearance across seeds.

Overall, these results support that the alignment improvements of Full are not explained by reduced randomness or mode collapse, but rather by structured inference-time manipulation that improves alignment while preserving seed-to-seed diversity.

F Additional Diagnostics for the Single-Probe Method

This appendix reports two diagnostics for Algorithm 1. Our method uses a *single* scheduler-consistent cold-start probe at $t^* = 900$, followed by tangential injection and spherical retraction. All experiments below keep this default single-probe configuration unchanged.

F.1 Single-Probe Cold-Start Diagnostic

We directly evaluate the default configuration on FLUX.1-Dev: one cold-start residual at $t^* = 900$, one tangential update, and one radius-shell retraction. All settings use the same sampler, prompt set, resolution, and matched random seeds. The diagnostic set contains 48 prompts with 8 seeds per prompt, giving 384 matched prompt–seed cases. We report HPSv3 and Q20, the 20th percentile over matched cases, to measure both average quality and lower-tail robustness.

Table 10: Single-probe cold-start diagnostic on FLUX. The method uses the default single high-noise probe at $t^* = 900$.

Setting	Probe configuration	HPSv3 $\uparrow$	Q20 $\uparrow$
Standard	none	10.466	8.613
Ours	single cold-start probe, $t^* = 900$	10.667	8.897

The single-probe configuration improves both the mean score and Q20, supporting the effectiveness of the cold-start residual as a model-coupled initialization-side correction.

F.2 Complementarity with Feature-Side Inference Control

We also test whether seed-side shaping composes with training-free feature-side inference control. Our method changes only the initial seed direction on the Gaussian radius shell, while feature-side controls modify denoising behavior during sampling. We use FreeU as a representative feature-side method and keep our single-probe procedure unchanged in the combined setting.

Table 11: Complementarity between single-probe seed shaping and feature-side inference control on FLUX. FreeU modifies denoising-side feature behavior, while our method modifies initialization geometry.

Setting	HPSv3 $\uparrow$	Gain vs. standard	Q20 $\uparrow$
Standard	10.466	–	8.613
FreeU only	10.502	+0.036	–
Ours only, single probe $t^* = 900$	10.667	+0.201	8.897
Ours + FreeU	11.167	+0.700	9.202

The combined setting achieves the strongest mean score and Q20, indicating that the proposed seed-side intervention is compatible with denoising-side improvements.

G Implementation Details

G.1 Model Details

FLUX.1 dev [13] is an open-weight rectified-flow transformer for high-resolution text-to-image synthesis. The network stacks 19 MMDiT [17] blocks that jointly attend to 4 096 T5-XXL [18] text embeddings and a $1\,024 \times 1\,024$ pixel latent grid.

SDXL [2] is a cascaded latent diffusion model that generates $1\,024 \times 1\,024$ images through a base+refiner ensemble.

SD3.5 Medium is a text-to-image diffusion transformer used in our sectional cone diagnostics under its official inference setting.

Z-Image [14] is a text-to-image generation backbone used in our finite-difference cone diagnostics under its official base-model setting.

Wan 2.1 1.3B [15] is a bilingual diffusion transformer optimized for Chinese–English text-to-video generation.

Trellis text Xlarge [16] is a sparse-voxel rectified-flow transformer that maps natural-language prompts directly to 3-D assets without 2-D distillation.

Table 12: Applicability and comparability of initial-noise optimization methods. “Training-free” indicates no additional optimization of model weights and no extra training dataset. “Arch.-agnostic” means not tied to UNet-only design choices and can be instantiated across different backbones/formulations given the standard inference interface.

Method	Training-free	Requires training	Custom dataset	Optimizes init noise	UNet diffusion	DiT Flow	Multimodal
ours	✓	×	×	✓	✓	✓	✓
InitNO	✓	×	×	✓	✓	×	×
NPNet	×	✓	✓	✓/–	✓/–	×	×

G.2 Baseline

InitNO is a *training-free* initial-noise optimization approach, but it is *designed specifically for UNet-based diffusion pipelines* and is *image-only*. It serves as the closest training-free baseline when comparing within UNet diffusion settings.

NPNet requires *additional training* with a *separately customized dataset* and is *image-only*. Thus, NPNet is *not a strictly comparable training-free baseline*. We include it as a representative reference point for the line of work that improves generation via *learned, data-dependent optimization modules*, and we report/interpret it separately from training-free comparisons.

G.3 Benchmarks

Pick-a-Pic [19] benchmark is a large-scale, open dataset designed to evaluate text-to-image generation models based on real user preferences.

DrawBench [20] benchmark is a comprehensive evaluation suite designed to assess the performance of text-to-image generation models in rendering complex spatial relationships and compositional scenes.

HPD [21] benchmark is a large-scale, human-annotated dataset designed to evaluate the alignment of text-to-image generative models with human aesthetic preferences.

Vbench [22] is a comprehensive evaluation suite designed to assess the performance of video generative models across a multitude of dimensions.

Toy4K [22] benchmark is a synthetic 3D object dataset designed to facilitate the evaluation of 3D object recognition and generation models.

G.4 Metrics

Pickscore [19] is a human preference prediction model designed to evaluate the alignment of images generated from textual prompts with human aesthetic judgments.

ImageReward [23] is a human preference reward model designed to evaluate and improve text-to-image generation models by aligning their outputs with human aesthetic judgments.

CLIPScore [24] is a reference-free evaluation tool designed to assess the alignment between generated captions and their corresponding images.

VQAScore with Qwen-2.5-VL [25, 26] To assess the semantic alignment between video content and textual descriptions, we employ VQAScore. This facilitates the computation of similarity between video frames and their corresponding textual prompts.

Vbench [22] is a comprehensive and versatile benchmark suite designed to address the critical challenge of evaluating video generative models by decomposing “video generation quality” into 16 specific, hierarchical, and disentangled dimensions, including subject consistency, motion smoothness, spatial relationship, and aesthetic quality.

VQAScore with Qwen-2.5-VL [25, 26] computes the VQAScore of rendered images and input prompts. It is averaged over multiple views to ensure consistency. Higher scores indicate better alignment, reflecting the model’s ability to adhere to prompts.

Fréchet Distance [27] (FD) with Inception-v3 [28] measures the similarity between feature distributions of generated and real images using Inception-v3. Lower FD values indicate that generated assets are closer to the real data distribution in feature space, reflecting higher visual quality.

Kernel Distance [29] (KD) with Inception-v3 [28] uses kernel methods (e.g., MMD) to compare feature distributions from Inception-v3. It assesses distribution alignment, with lower scores suggesting that generated outputs are realistic and diverse.

Fréchet Distance [27] (FD) with DINOv2 [30] This variant employs DINOv2 features, which capture rich semantic information. Lower FD values indicate better semantic consistency with real-world objects, leveraging the model’s strong representation power.

Kernel Distance [29] (KD) with DINOv2 [30] By applying kernel distance to DINOv2 features, this metric emphasizes structural and semantic similarities. Lower scores denote that generated assets align well with expected visual characteristics.

G.5 Model Setting

For the standard method, we adopted the hyperparameters presented in Table 13, we increased the number of inference steps for the standard pipeline so that its total inference time matches ours, which incurs additional computational cost due to our design.

Table 13: Inference Settings for SDXL, FLUX.1 Dev, Wan2.1 T2V 1.3B, and TRELLIS Text XLarge

Parameter	SDXL	FLUX.1 Dev
Number of Inference Steps	51 (Before: 50)	29 (Before: 28)
Guidance Scale	7.5	3.5
Eta (for DDIM Scheduler)	0.0	-
Output Image Size	1024 × 1024	1024 × 1024
True CFG Scale	-	1
Parameter	Wan2.1 T2V 1.3B	TRELLIS Text XLarge
Resolution	832 × 480 (480p)	-
Sample Guide Scale	6	-
Sample Shift	10	-
Sparse Sampler Steps	-	26 (Before: 25)
Sparse Sampler CFG Strength	-	7.5
SLAT Sampler Steps	-	26 (Before: 25)
SLAT Sampler CFG Strength	-	7.5

Table 14: Recommended hyperparameters for Algorithm 1.

Hyperparam	Rec.	Range	Notes / tuning tips
t^*	900 ~ 1000	high-noise timesteps	Single cold-start probe used for the prompt residual.
δ	5 ~ 6	0 ~ 15	Stronger injection can improve alignment but may overshoot.
$\rho(t)$	model-native	model-dep.	Use native schedule for time-consistent probing.
Orthogonalize	on	on/off	Decouple from $\mathbf{z}_T$; turn off if effect weakens.
Normalize	on	on/off	Stabilizes the single-step residual direction.

G.5.1 Hyperparameter

This section analyzes the two key hyperparameters of our method: the single cold-start probe timestep t^* and the injection strength δ .

Single cold-start probe timestep t^* The main method uses one high-noise cold-start probe and does not aggregate residuals across timesteps. All main experiments extract the prompt residual at a fixed high-noise timestep $t^* = 900$. This single-step probe provides a stable semantic signal across

Table 15: User study results and adult participant demographics.

(a) Preference results across generation modalities.			(b) Adult participant demographics.			
Category	Ours (%)	Standard (%)	Age Group	Male	Female	Percentage
T2I Basic Objects & Colors	92.5	7.5	18–25	64	42	24.54%
T2I Counterfactual Concepts	85.0	15.0	26–30	62	46	25.00%
T2I Spatial & Textual Relations	55.0	45.0	31–40	32	22	12.50%
T2I Stylistic & Artistic Rendering	81.4	18.6	41–50	30	24	12.50%
T2I Complex & Specific Concepts	72.9	27.1	51–60	38	32	16.20%
T2V Dynamic Scenes	66.4	33.6	Over 60	26	12	8.80%
T2V Static Scenes	53.8	46.2	Total	254	178	100.00%
Text to 3D Detailed Objects	58.6	41.4				
Text to 3D Specialized Objects	51.9	48.1				
Overall (T2I)	74.1	25.9				
Overall (T2V)	64.5	35.5				
Overall (Text to 3D)	55.3	44.7				

both simple and complex prompts while requiring only one additional conditional/unconditional probing pair.

Injection Strength δ The injection strength δ controls the magnitude of proxy-guided tangential injection before retraction, and thus directly trades off semantic enhancement and visual quality. We sweep δ over a wide range and evaluate performance on both simple prompts and complex prompts.

Overall, we observe that small δ leads to limited semantic gain, while overly large δ may introduce artifacts and degrade aesthetic quality. Across our tests, $\delta = 6$ **consistently provides strong improvements for both simple and complex prompts while maintaining aesthetic metrics without degradation**. We therefore adopt $\delta = 6$ as the default setting.

H Visualization Method in Figure 1

The overall visualization method in Figure 1 is summarized in Fig. 16.

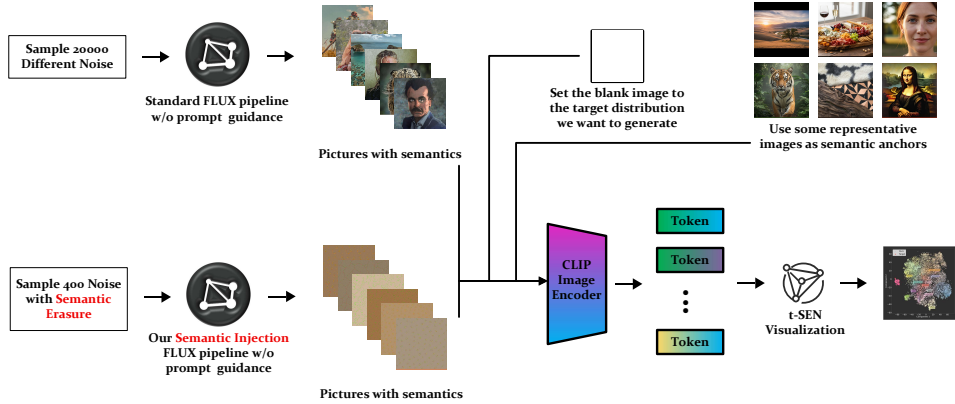


Figure 16: Visualization Method in Fig. 1.

I Details about User Study

I.1 Participants and Demographics

To evaluate generation quality and prompt alignment from a human-preference perspective, we conducted a user study across image, video, and 3D generation tasks. We report only responses from adult participants who self-reported being at least 18 years old. Responses from participants who reported being under 18 were excluded before analysis. The final analyzed pool contains 432 adult participants. The study was voluntary and anonymous, and participants were asked to choose which result better matched the given textual description.

I.2 Survey Methodology

As shown in Fig. 17, participants were presented with paired outputs generated by our method and the standard baseline, together with the corresponding textual description. For each pair, participants were asked to select the image, video, or 3D asset that better matched the description. The order of the two methods was randomized to reduce position bias. The prompts were designed to evaluate several aspects of alignment, including object presence, color fidelity, spatial relationships, text rendering, motion consistency, and object-level detail.

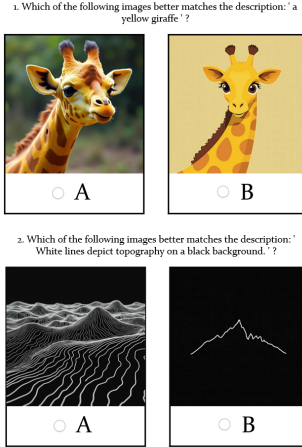


Figure 17: Schematic diagram of the questionnaire design.

I.3 Results

I.3.1 T2I

For text-to-image generation, our method is preferred over the standard baseline across all evaluated categories. Preference rates range from 55.0% on spatial and textual relations to 92.5% on basic objects and colors, with an overall T2I preference rate of 74.1%. The gains are strongest on prompts involving object presence, color fidelity, counterfactual concepts, and stylistic rendering. Spatial and textual relation prompts remain more challenging, indicating that fine-grained layout and text rendering are still difficult failure modes.

I.3.2 T2V

For text-to-video generation, our method is also preferred over the standard baseline. Preference rates are 66.4% for dynamic scenes and 53.8% for static scenes, with an overall T2V preference rate of 64.5%. The stronger gain on dynamic prompts suggests improved text-video alignment for motion-centric scenes, while static or less visually distinctive scenes show smaller margins.

I.3.3 Text to 3D

For text-to-3D generation, our method obtains a moderate but consistent preference advantage. Preference rates are 58.6% for detailed objects and 51.9% for specialized objects, with an overall Text-to-3D preference rate of 55.3%. The results suggest that seed shaping can improve prompt adherence in 3D generation, especially for detailed object descriptions, while highly specialized objects remain challenging.

J Limitations

Our framework separates three objects that are easy to conflate: (i) the ideal geometric object induced by the semantic map, (ii) the practical prompt-residual update used by the sampler, and (iii) the empirical diagnostics used to test whether the practical update behaves like the ideal object. This separation is important because our method is not intended to recover the exact semantic-horizontal

subspace of the diffusion model. Rather, it provides a training-free, inference-time approximation whose validity is supported empirically under the tested models, prompts, and schedulers.

The ideal horizontal space is not available at inference time. The geometric analysis in Section 3 characterizes the effective semantic map $\Phi = \pi \circ \Psi_{T \rightarrow 0}$ in Eq. (4). At a regular point, the Jacobian $J_z = D\Phi_z$ induces the vertical–horizontal decomposition in Eq. (5), where vertical directions are locally semantic-invariant and horizontal directions are locally semantic-sensitive. On the Gaussian radius shell, this induces the shell split in Eq. (17), with shell-horizontal directions lying in $\text{Img}(Q_z J_z^\top)$, where the shell projector Q_z is defined in Eq. (15). If this shell-horizontal space were directly accessible, one could in principle design seed interventions that move preferentially along semantic-sensitive directions while preserving the Gaussian radius shell.

This ideal intervention can be written explicitly. For the prompt-semantic potential $F_y(z) = \ell_y(\Phi(z))$ defined in Eq. (19), the geometry-matched shell direction is the shell gradient in Eq. (20):

$$g_{\text{oracle}}(z) := \nabla_{\mathbb{S}_R} F_y(z) = Q_z J_z^\top \nabla_\Phi \ell_y(\Phi(z)).$$

Equivalently, if the semantic metric is kept explicit rather than absorbed into $\nabla_\Phi \ell_y$, this direction can be written as

$$g_{\text{oracle}}(z) = Q_z D(\pi \circ \Psi_{T \rightarrow 0})_z^\top G_{\mathcal{M}}(\Phi(z)) \nabla_\Phi \ell_y(\Phi(z)).$$

This direction is exactly shell-horizontal by Lemma 4.1. The corresponding radius-preserving oracle update would use the same spherical retraction structure as Eq. (38):

$$z^+ = \|z\| \frac{z + \delta g_{\text{oracle}}(z) / (\|g_{\text{oracle}}(z)\| + \varepsilon)}{\|z + \delta g_{\text{oracle}}(z) / (\|g_{\text{oracle}}(z)\| + \varepsilon)\| + \varepsilon}.$$

By the first-order shell expansion in Eqs. (39)–(40), this oracle update performs first-order ascent on F_y for sufficiently small δ , while remaining on the original Gaussian radius shell.

In practice, however, this oracle direction is computationally inaccessible. Evaluating Φ in Eq. (4) at a single seed already requires a complete denoising trajectory followed by a semantic-encoder forward pass. Computing $D\Phi_z^\top$ would require differentiating through the entire sampler and semantic encoder, or approximating the Jacobian with many radius-preserving finite-difference generations. Both options are far more expensive than the intended use case: shaping one initial seed before generation. Moreover, $D\Phi_z$ is local to a particular seed, prompt, scheduler, and model state, so a Jacobian computed at one point does not provide a reusable horizontal space for other seeds. For this reason, our algorithm should not be interpreted as an exact estimator of $\mathcal{H}_z$ or $\mathcal{H}_z^{\text{sh}}$. Instead, it should be viewed as a low-cost approximation to the geometry-matched oracle direction above.

The prompt residual is a proxy, not an oracle. Our practical method replaces the unavailable oracle direction with a model-coupled prompt-residual proxy. Specifically, we use the conditional–unconditional residual in Eq. (21), evaluated at the high-noise cold-start state in Eq. (22). The motivation is that, under the local decomposition in Eq. (24), the prompt residual contains a component aligned with the pullback of a prompt-semantic gradient, plus an error term e_t . After shell projection, Eq. (25) gives

$$Q_z g_\theta(z, t, y) = c_t \nabla_{\mathbb{S}_R} F_y(z) + Q_z e_t,$$

so the implemented direction can be interpreted as a noisy approximation to the ideal shell gradient in Eq. (20). The actual update then removes the radial component and retracts the seed to its original Gaussian radius shell as in Eq. (38).

This argument has an important limitation: Eq. (24) is a modeling assumption, not a runtime guarantee. The residual may contain components unrelated to the semantic-horizontal space, and the residual-to-gradient relationship can vary across model families, parameterizations, schedulers, guidance scales, and prompt types. Accordingly, we do not claim that the prompt residual exactly recovers the true horizontal direction. Our claim is more limited: for the tested backbones and settings, the shell-projected residual exhibits the empirical signatures expected of a semantic-horizontal proxy.

The diagnostics provide evidence, not a formal certificate. Because $D\Phi_z$ is unavailable, the true vertical and horizontal spaces in Eq. (5) and Eq. (17) cannot be directly measured. We therefore use finite-difference semantic probes to construct an empirical local section of the radius shell and test whether the prompt-residual direction behaves differently from random tangent controls. The

cone diagnostics in Section 5.1 and Appendix D.6 measure whether the shell-projected residual has low estimated vertical leakage, high concentration in the empirical semantic-sensitive subspace, and larger semantic displacement than random tangent directions.

These diagnostics are useful because they test the empirical version of the cone condition in Eq. (27). In practice, we estimate the finite-dimensional cone ratio $\hat{\rho}_k$ using Eq. (29). However, this remains an empirical diagnostic rather than a formal certificate. The estimate depends on the chosen semantic projection π , the finite-difference scale, the number of tangent probes, and the evaluator used to measure semantic displacement. Thus, values such as $\hat{\rho}_k < 0.2$, horizontal energy above 96%, and semantic-displacement percentile 1.000 should be interpreted as strong empirical support for the proxy in the evaluated regimes, not as a mathematical certificate that the exact horizontal space has been recovered.

Failure modes and operating boundaries. The method can degrade when the prompt-residual proxy no longer approximates a semantic-horizontal direction. We identify several such regimes.

1. **Extreme classifier-free guidance.** Very large guidance scales can amplify the residual error term e_t in Eq. (24) and push the update outside the local regime where the shell approximation is meaningful. In this case, the method may overshoot, distort structure, or reduce quality. The guidance stress tests in Table 3b illustrate this operating boundary.
2. **Out-of-distribution backbones or schedulers.** The relationship between the prompt residual in Eq. (21) and the shell-gradient proxy in Eq. (25) can change across model architectures, parameterizations, timestep schedules, and noise scalings. A probe timestep and update strength that work for one backbone may therefore require retuning for another.
3. **Large update strengths.** The bridge in Theorem 4.6 is local. If the tangential injection strength δ in Eq. (38) is too large, the update may leave the neighborhood in which the residual is a reliable proxy. This is consistent with the first-order expansion in Eq. (40) and with the ablation in Figure 3, where moderate values of δ are beneficial while overly large values can degrade results.

In these regimes, the method should be treated as a bounded local seed correction rather than a general steering mechanism. When applying the method to a new model family or prompt distribution, we recommend re-running the cone diagnostics in Section 5.1 and strength ablations in Section 5.5 before drawing conclusions about effectiveness.

Scope of the contribution. These limitations do not undermine the central contribution, but they clarify its scope. The geometric analysis explains why an isotropic Gaussian prior can induce highly anisotropic semantic behavior after generation: the semantic map in Eq. (4) induces the degenerate pullback geometry in Eqs. (6)–(7), with semantic-sensitive directions concentrated in the horizontal component of Eq. (5). The geometry-matched oracle update above shows the exact seed intervention implied by this theory, while also making clear why it is impractical for standard inference. Our algorithm provides a cheap prompt-residual approximation to this unavailable shell-horizontal direction via Eqs. (21)–(25) and implements the final prior-compatible update through Eq. (38). The diagnostics test whether this approximation behaves as predicted in practice. Together, they support a practical and empirically validated seed-shaping procedure, while stopping short of claiming exact recovery of the latent semantic geometry or universal improvement across all generation regimes.

Table 16: Main generation results across image, video, and 3D generation. Δ and Win are computed against Standard using paired bootstrap over matched cases. The primary metric is PickScore for image generation and VQA for video/3D generation.

(a) Image generation.							
Model	Dataset	Method	Pick $\uparrow$	ImgR $\uparrow$	CLIP $\uparrow$	Δ Pick [95% CI]	Win
SDXL	Pick-a-Pic	Standard	17.049	-1.971	16.223	–	–
SDXL	Pick-a-Pic	Initno [†] [5]	17.050	-1.969	16.237	–	–
SDXL	Pick-a-Pic	NPNNet [6]	17.051	-1.968	16.250	–	–
SDXL	Pick-a-Pic	Ours	17.371	-1.852	16.642	+0.322 [0.241, 0.404]	68.4%
SDXL	DrawBench	Standard	17.416	-2.084	16.618	–	–
SDXL	DrawBench	Initno [†] [5]	17.419	-2.085	16.632	–	–
SDXL	DrawBench	NPNNet [6]	17.420	-2.084	16.640	–	–
SDXL	DrawBench	Ours	17.619	-1.984	16.957	+0.203 [0.089, 0.314]	63.1%
SDXL	HPD	Standard	16.748	-1.955	14.986	–	–
SDXL	HPD	Initno [†] [5]	16.765	-1.952	15.010	–	–
SDXL	HPD	NPNNet [6]	16.780	-1.948	15.040	–	–
SDXL	HPD	Ours	17.041	-1.875	15.851	+0.293 [0.171, 0.417]	66.2%
FLUX	Pick-a-Pic	Standard	17.134	-1.942	16.295	–	–
FLUX	Pick-a-Pic	NPNNet [6]	17.140	-1.936	16.305	–	–
FLUX	Pick-a-Pic	Ours	17.732	-1.905	16.914	+0.598 [0.481, 0.716]	74.5%
FLUX	DrawBench	Standard	17.275	-2.051	16.642	–	–
FLUX	DrawBench	NPNNet [6]	17.282	-2.044	16.655	–	–
FLUX	DrawBench	Ours	17.626	-1.968	17.013	+0.351 [0.236, 0.462]	69.8%
FLUX	HPD	Standard	16.730	-1.951	15.021	–	–
FLUX	HPD	NPNNet [6]	16.750	-1.944	15.900	–	–
FLUX	HPD	Ours	17.239	-1.874	16.731	+0.509 [0.374, 0.641]	72.1%

(b) Video generation.					
Method	AQ $\uparrow$	AS $\uparrow$	BC $\uparrow$	DD $\uparrow$	IQ $\uparrow$
Standard	0.572	0.220	0.963	0.644	0.644
Ours	0.651	0.341	0.985	0.701	0.664
Method	MS $\uparrow$	OC $\uparrow$	SC $\uparrow$	TF $\uparrow$	VQA $\uparrow$
Standard	0.972	0.244	0.937	0.984	0.534
Ours	0.997	0.303	0.994	0.997	0.600
Primary metric		Δ VQA [95% CI]		Win	
Ours vs. Standard		+0.066 [0.041, 0.091]		71.3%	

(c) 3D generation.					
Method	FD _I $\downarrow$	KD _I $\downarrow$	FD _D $\downarrow$	KD _D $\downarrow$	VQA $\uparrow$
Standard	29.544	0.011	340.530	0.802	0.876
Ours	29.403	0.008	335.425	0.733	0.911
Primary metric		Δ VQA [95% CI]		Win	
Ours vs. Standard		+0.035 [0.018, 0.053]		64.7%	

I: Inception; *D*: DINOv2.

K Additional Main Experiment Results

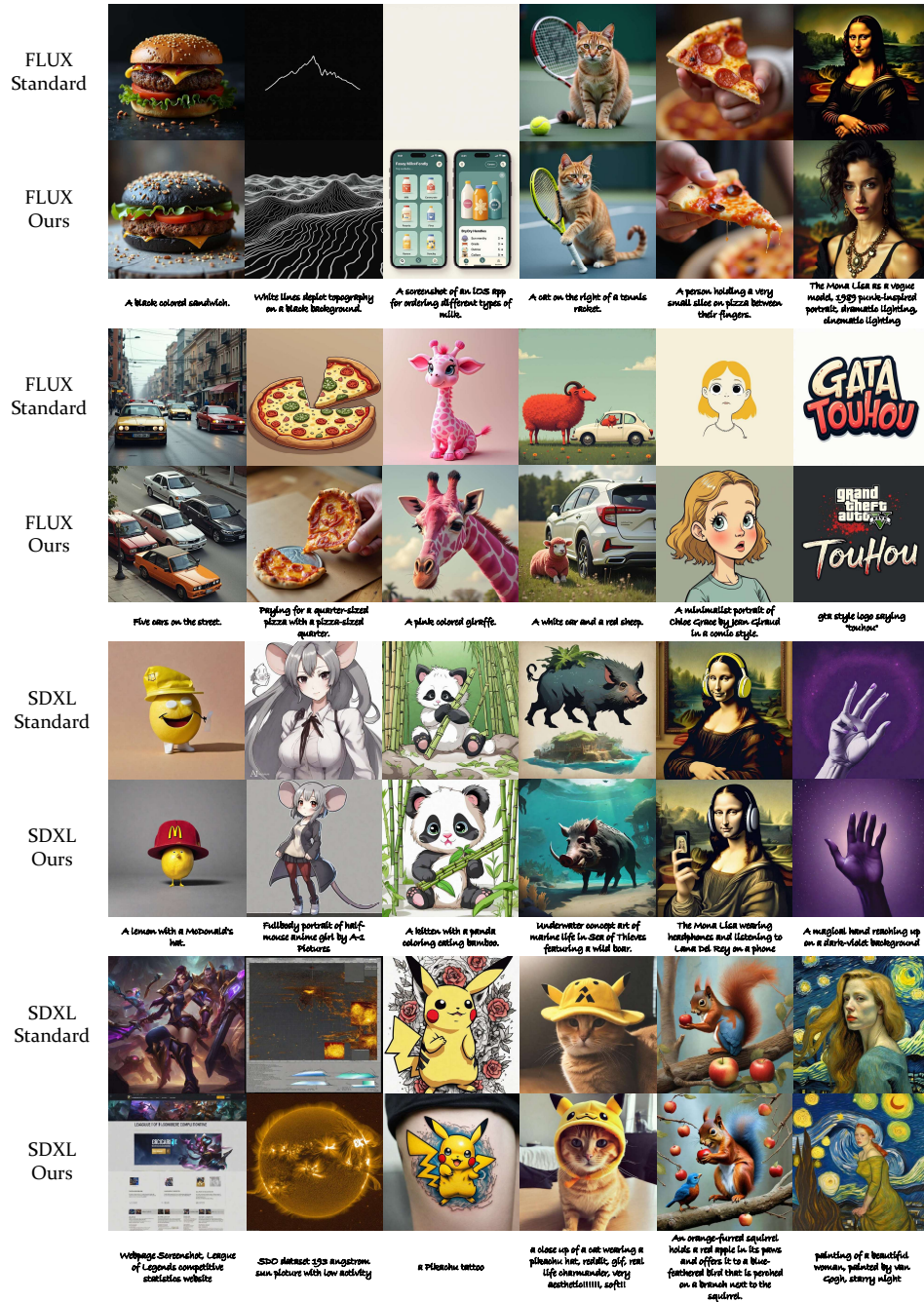


Figure 18: More Results of SDXL and FLUX

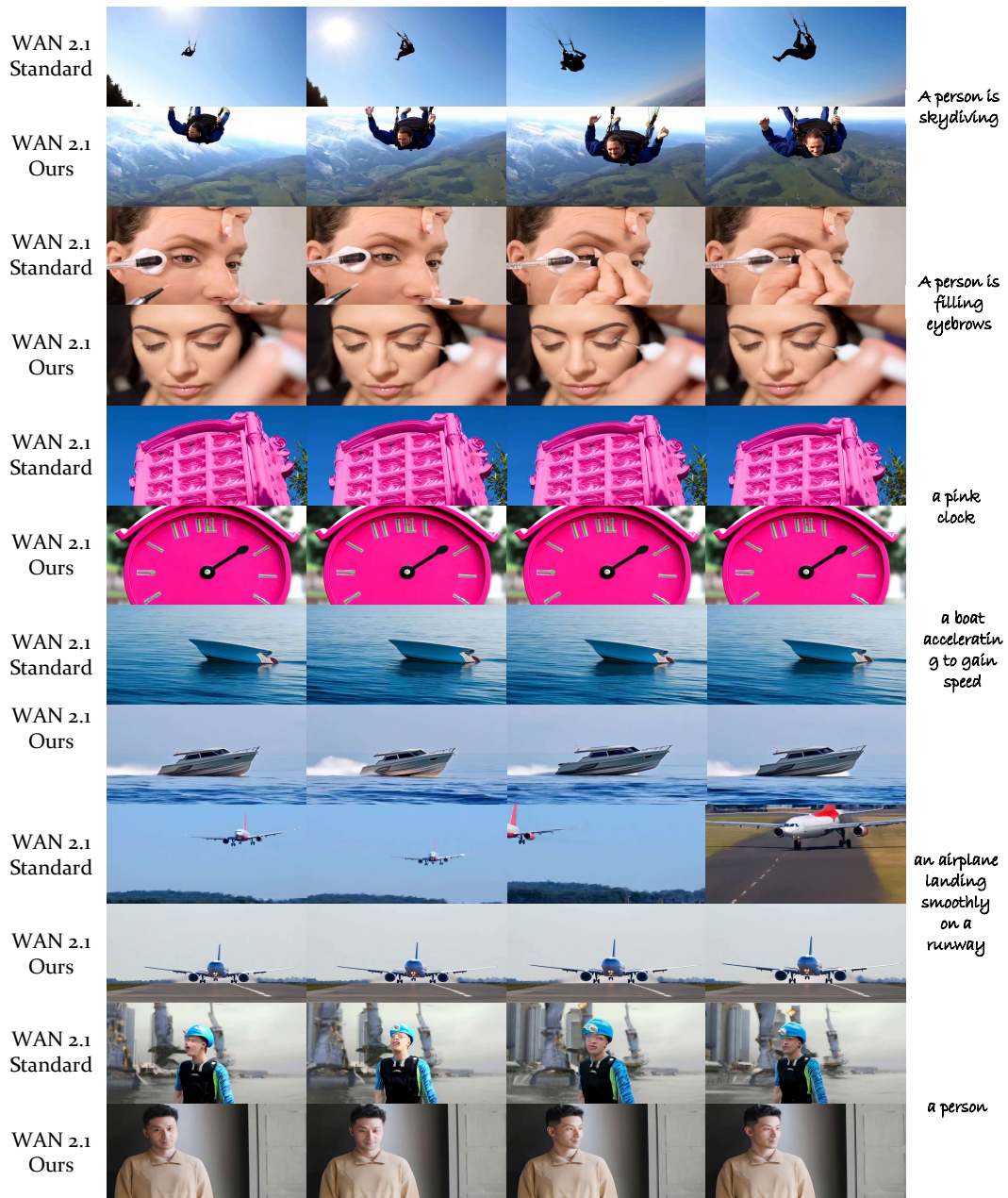
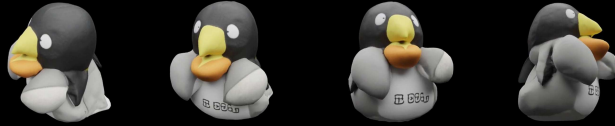


Figure 19: More Results of WAN 2.1 1.3B

TRELLIS
Standard



Stuffed penguin toy with a rounded, teardrop-shaped body; solid dark head and back, white front; oval head with white

TRELLIS
Ours



eyes, black centers; yellow triangular beak and feet; stitched fabric; wings flat and narrow; label 'TUX' on back.

TRELLIS
Standard

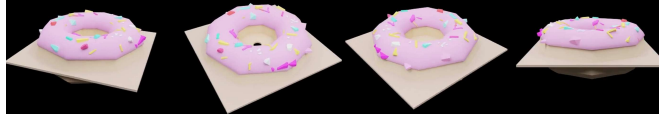


A vintage telephone with a wooden cylindrical handset, black rounded earpiece and mouthpiece with white ends, wooden supports with black bases, and a black multi-tiered rectangular base featuring a black rotary dial with white markings.

TRELLIS
Ours



TRELLIS
Standard



Low-poly torus-shaped donut with beige base and light pink icing, scattered rectangular sprinkles in blue, red, yellow, and white. The surface is faceted and polygonal, with the donut oriented horizontally and a visible central hole.

TRELLIS
Ours

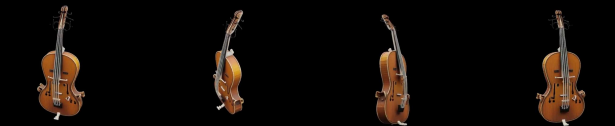


TRELLIS
Standard

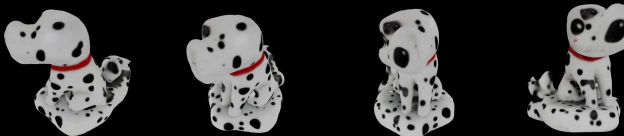


A 3D model of a violin with an amber body and darker front, white neck and scroll, black fingerboard, pegs, and tailpiece, metallic silver strings, a light beige bridge, and black f-holes.

TRELLIS
Ours



TRELLIS
Standard



A small figurine of a seated Dalmatian dog with a predominantly white body, black spots, detailed black facial features, large floppy black ears, a black spot above the left eye, a black tail laid flat, and a red collar.

TRELLIS
Ours

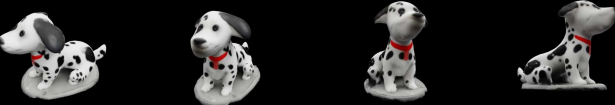


Figure 20: More Results of TRELLIS Text Xlarge