RiCoTA: Riemannian Companding Token Autoencoder

Anonymous authors
Paper under double-blind review

Abstract

Fast generation on Riemannian manifolds must preserve the intrinsic structure of data distributions with low sampling cost. We introduce the Riemannian Companding Token Autoencoder (RiCoTA), which learns a joint distribution over discrete latent codes for continuous manifold generation. Its design addresses a geometric mismatch: equal-width latent intervals can induce very different geodesic displacements after decoding. RiCoTA therefore organizes latent coordinates by average decoder sensitivity and allocates finer resolution where local changes have larger intrinsic effects, keeping the number of intervals per coordinate fixed. Training and generation share a uniform within-cell reconstruction rule. At inference, an autoregressive prior selects cells, continuous values are sampled within them, and one decoder evaluation produces manifold samples without trajectory integration. For a fixed representation, our asymptotic analysis derives a coordinatewise bound on intrinsic distortion and identifies the sensitivity-based allocation that minimizes its leading term. A complementary population bound controls generation error in terms of quantized reconstruction error and joint code-distribution mismatch. Held-out diagnostics show that estimated sensitivity predicts geometric risk and adaptive allocation reduces coordinatewise intrinsic distortion at a fixed budget. Experiments across 14 manifold benchmarks demonstrate strong support fidelity at low end-to-end sampling cost. Code is available at `https://anonymous.4open.science/r/RiCoTA-5D6B/`.

1 Introduction

Generative models on Riemannian manifolds must capture how probability mass is organized under intrinsic distance, not merely produce points that satisfy manifold constraints. Geographic locations, molecular angles, and rotations make this requirement concrete (Mardia & Jupp, 2000; Lovell et al., 2003; Murray et al., 2003; Falorsi et al., 2019). Diffusion, score-based models, and continuous flows incorporate geometry through dynamics whose simulation requires repeated network evaluations and manifold operations (Mathieu & Nickel, 2020; De Bortoli et al., 2022; Huang et al., 2022; Chen & Lipman, 2024; Jo & Hwang, 2024). One- and few-step methods reduce this cost, while direct feed-forward generators avoid numerical integration (Cheng et al., 2025; Davis et al., 2026; Zhong et al., 2026; Sorrenson et al., 2024). We study a complementary design axis for efficient manifold generation: the resolution at which a latent distribution is modeled and how it should reflect decoded geometry.

Finite codes make modeling resolution an explicit choice (Bennett, 1948; Gray & Neuhoff, 1998; van den Oord et al., 2017; Razavi et al., 2019). Partitioning latent coordinates into intervals, or cells, gives a learned joint code model—the generative prior—a concrete target: which combinations of intervals occur. A fixed within-cell sampling rule supplies continuous variation, so each code identifies a region rather than a reconstruction prototype. The challenge is geometric: equal changes in a latent coordinate can induce very different geodesic displacements after decoding. Uniform intervals therefore need not resolve comparable geometric detail. Even after marginal probability normalization, the decoder can remain more sensitive in some regions than in others. Equal-probability cells can therefore cover very different intrinsic variations, spending the same resolution on nearly invariant regions as on regions where small latent changes have larger geometric effects. This motivates treating latent resolution as a geometric resource: with the number of intervals per coordinate fixed, the partition specifies which geometric distinctions the code model must resolve.

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

We introduce the Riemannian Companding Token Autoencoder (RiCoTA), which calibrates latent resolution at two complementary scales. Globally, a generalized eigenvalue construction normalizes latent covariance and orders directions by average intrinsic decoder sensitivity. Marginal cumulative distribution functions express active coordinates on a common probability scale. Locally, RiCoTA measures the intrinsic effect of varying one coordinate while keeping the others at observed training values. This data-supported response determines a monotone warp, or compander, that allocates finer resolution to sensitive regions. Geometry thus determines both the coordinate system and the placement of resolution within it, rather than entering only as a reconstruction loss.

RiCoTA aligns reconstruction with generation by training the encoder and decoder through hard quantization and uniform within-cell sampling. This exposes the decoder during training to the continuous within-cell variation used by the generative sampler. Once the representation and partition are fixed, an autoregressive prior learns the resulting joint code distribution. Generation draws codes from this prior, samples continuous values within the selected intervals, and decodes once without numerical trajectory integration.

Our analysis provides a geometric basis for this allocation. For a fixed representation, we derive a coordinatewise intrinsic-distortion bound and identify the sensitivity-based allocation that minimizes its leading term as the number of intervals grows. A complementary population bound controls generation error through quantized reconstruction error and mismatch between the data-induced and learned joint code distributions.

We evaluate RiCoTA on 14 benchmarks spanning spheres, product tori, $\mathrm{SO}(3)$, and hyperbolic space. On the 13 tasks shared by all methods, RiCoTA achieves a macro-averaged multi-scale support F1 of 0.872 versus 0.743 for M-FFF, the best baseline on this metric. Its macro-averaged end-to-end time to generate 100,000 samples is 14.67 ms, a $2.81\times$ speedup over the next-fastest baseline, M-FFF. Held-out diagnostics show that estimated sensitivity predicts intrinsic risk, while frozen-representation comparisons isolate the benefit of adaptive cell placement at a fixed budget. A case study adapts RiCoTA to conditional generation of 1,024-base DNA promoters (Woo et al., 2026), illustrating applicability to high-dimensional sequences.

2 Problem Setup

Core notation. Given observations $X \sim \rho^*$ on a connected Riemannian manifold $(\mathcal{M}, g)$, we seek a generative model using finite codes and one decoder evaluation per sample. At $x \in \mathcal{M}$, $T_x\mathcal{M}$ is the tangent space and g_x its smoothly varying positive-definite inner product. We write $\|v\|_{g,x}^2 := g_x(v, v)$ and denote the induced geodesic distance by d_g. RiCoTA uses an encoder $E_\phi : \mathcal{M} \to \mathbb{R}^{m_0}$ and a manifold-valued decoder $D_\theta : \mathbb{R}^{m_0} \to \mathcal{M}$, with m_0 latent coordinates and K_0 symbols per active coordinate:

$$Z = E_\phi(X), \qquad \mathcal{C} = \{0, \dots, K_0 - 1\}^{m_0}. \tag{1}$$

The deterministic code map $Q_{\phi,\theta} : \mathcal{M} \to \mathcal{C}$ produces $C = Q_{\phi,\theta}(X)$, whose joint law $\Pi_{\phi,\theta}$ is approximated by the autoregressive prior P_ψ^C. The finite interface $\mathcal{I}$ stores calibrated coordinate transforms and cell geometry; dependence on $\mathcal{I}$ is left implicit where unambiguous. Given independent $\varepsilon \sim \mathrm{Unif}([0,1)^{m_0})$, the reconstruction map $\mathrm{Rec}_{\theta,\mathcal{I}} : \mathcal{C} \times [0,1)^{m_0} \to \mathcal{M}$ samples uniformly within the selected cells, maps the sampled values back to the decoder's latent coordinates, and evaluates D_θ once. Writing $f_\sharp\mu$ for the distribution obtained by applying f to samples from μ, the data-code and generated distributions are

$$\begin{aligned} \Pi_{\phi,\theta} &= (Q_{\phi,\theta})_\sharp \rho^*, \\ \rho_{\theta,\psi} &= (\mathrm{Rec}_{\theta,\mathcal{I}})_\sharp \left(P_\psi^C \otimes \mathrm{Unif}([0,1)^{m_0}) \right). \end{aligned} \tag{2}$$

Phase 1 learns the representation and calibrates the interface; Phase 2 freezes both and fits the code prior so that $\rho_{\theta,\psi}$ approximates ρ^*. Section 3 constructs the interface and one-decode sampler; Appendix B gives numerical conventions.

3 RIEMANNIAN COMPANDING TOKEN AUTOENCODER

Sections 3.1–3.3 construct the finite interface for a fixed encoder–decoder pair along $X \to Z \to \Xi \to U \to C$; Section 3.4 couples it to encoder–decoder training and fits the frozen joint code law. Figure 1 summarizes geometry-adapted cell construction and one-decode generation.

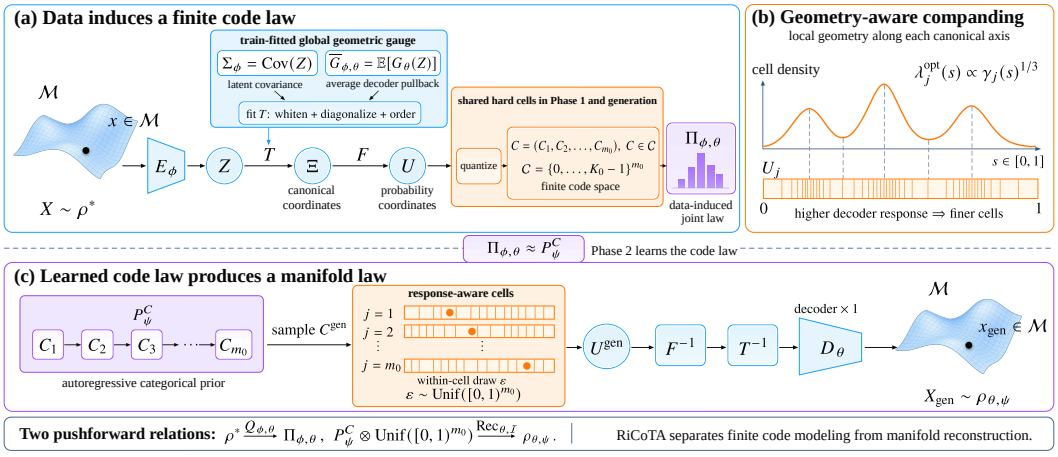


Figure 1: **Overview of RiCoTA. (a)** Global geometry canonicalizes latent coordinates. **(b)** Intrinsic response allocates finer cells to sensitive regions. **(c)** Phase 2 learns the joint code law; independent within-cell draws pass through frozen inverse transforms and one decoder call. Phase 1 and generation share cells and the sampling rule. Cells are schematic.

3.1 CANONICAL PROBABILITY COORDINATES

For $Z = E_\phi(X)$, let $\mu_\phi = \mathbb{E}[Z]$ and $\Sigma_\phi = \mathrm{Cov}(Z)$ be the population mean and covariance. Assume D_θ is differentiable on the active latent region. Its differential $dD_{\theta,z} : T_z\mathbb{R}^{m_0} \simeq \mathbb{R}^{m_0} \to T_{D_\theta(z)}\mathcal{M}$ sends latent directions to tangent velocities at $D_\theta(z)$. The *decoder pullback matrix* (Arvanitidis et al., 2018) and its population average are

$$v^\top G_\theta(z) w := g_{D_\theta(z)}(dD_{\theta,z}[v], dD_{\theta,z}[w]), \qquad \overline{G}_{\phi,\theta} := \mathbb{E}\big[G_\theta(Z)\big], \tag{3}$$

for $v, w \in \mathbb{R}^{m_0}$. Thus $G_\theta(z)$ measures the local squared geodesic effect of latent perturbations.

We choose canonical coordinates with identity covariance (Kessy et al., 2018) and diagonal mean decoder pullback, ordered by decreasing sensitivity. For full-rank Σ_ϕ, generalized eigenpairs (η_j, w_j) define Ξ:

$$\begin{aligned} \overline{G}_{\phi,\theta} w_j &= \eta_j \Sigma_\phi^{-1} w_j, & w_a^\top \Sigma_\phi^{-1} w_b &= \delta_{ab}, \\ \Xi = T(Z) &= W^\top \Sigma_\phi^{-1}(Z - \mu_\phi), & T^{-1}(\xi) &= \mu_\phi + W\xi, \end{aligned} \tag{4}$$

where $W = [w_1, \ldots, w_{m_0}]$, so $\mathrm{Cov}(\Xi) = I$ and $W^\top \overline{G}_{\phi,\theta} W = \mathrm{diag}(\eta_1, \ldots, \eta_{m_0})$ with $\eta_1 \geq \cdots \geq \eta_{m_0}$.

For active coordinates, the population marginal CDFs $F = (F_1, \ldots, F_{m_0})$ give a common probability scale; $\mathcal{D}$ is the same decoder expressed in these coordinates:

$$U_j = F_j(\Xi_j), \qquad \mathcal{D}(u) = D_\theta\big[T^{-1}\big(F^{-1}(u)\big)\big]. \tag{5}$$

For continuous active marginals, each U_j is uniform on $[0,1]$, but the coordinates can remain dependent. In practice, train-only estimates $(\widehat{\mu}, \widehat{\Sigma}_\epsilon, \widehat{\overline{G}}, \widehat{T}, \widehat{F})$ replace the population quantities:

$$\widehat{\Xi} = \widehat{T}(Z), \qquad \widehat{U} = \widehat{F}(\widehat{\Xi}), \qquad \widehat{\mathcal{D}}(u) = D_\theta\Big[\widehat{T}^{-1}\Big(\widehat{F}^{-1}(u)\Big)\Big]. \tag{6}$$

The empirical map evenly spaces training ranks but need not make unseen observations uniform. Appendix B specifies empirical calibration and numerical conventions.

3.2 DATA-SUPPORTED DECODER RESPONSE

Uniform cells can waste resolution as sensitivity varies along each canonical axis. Let $a_{\mathcal{M}} > 0$ be a train-only intrinsic distance scale and define $\overline{d}_g = d_g / a_{\mathcal{M}}$. For the response analysis, assume that $\mathcal{D}$ is differentiable along the coordinate paths considered below. For the standard basis vector e_j, write $\partial_j \mathcal{D}(u) := d\mathcal{D}_u[e_j] \in T_{\mathcal{D}(u)}\mathcal{M}$; $\|\cdot\|_g$ abbreviates $\|\cdot\|_{g,\mathcal{D}(u)}$. The population conditional response is

$$\gamma_j(s) := \mathbb{E}\Big[a_{\mathcal{M}}^{-2} \, \|\partial_j \mathcal{D}(U)\|_g^2 \;\Big|\; U_j = s\Big]. \tag{7}$$

It measures squared decoded sensitivity to coordinate j at value s, averaging the remaining coordinates over data-supported contexts.

Paired finite secants of $\widehat{\mathcal{D}}$ estimate this response without a full Jacobian. For each pilot interval r, training observations with $\widehat{U}_j$ in that interval supply $\widehat{U}_{-j}$, which is held fixed while coordinate j is probed at two interior values. Averaging the resulting squared normalized geodesic secants gives $\widehat{\gamma}_{j,r}$. Appendix B specifies the probes, context budget, and numerical floor.

3.3 HIGH-RATE COMPANDING AND FINITE TOKENIZATION

Unlike F_j, which sets the probability scale, the compander A_j allocates resolution along it. Let $A_j : [0,1] \to [0,1]$ be increasing with $A_j(0) = 0$, $A_j(1) = 1$, and *point density* $\lambda_j(s) = A_j'(s)$. Equal-width bins of $A_j(U_j)$ have local width approximately $[K_0 \lambda_j(s)]^{-1}$ in U_j, so larger λ_j gives finer cells. Under the high-rate conditions in Appendix C, including the uniform remainder condition (53), uniform resampling in the selected U_j-cell with other coordinates fixed has leading normalized squared geodesic distortion between decoder outputs

$$\mathcal{E}_{j,K_0}(\lambda_j) = \frac{1}{6K_0^2} \int_0^1 \frac{\gamma_j(s)}{\lambda_j(s)^2} \, ds + o(K_0^{-2}). \tag{8}$$

For $0 < \int_0^1 \gamma_j^{1/3} < \infty$, the leading functional over unit-integral densities is minimized by

$$\lambda_j^{\mathrm{opt}}(s) = \frac{\gamma_j(s)^{1/3}}{\int_0^1 \gamma_j(v)^{1/3} \, dv}, \qquad A_j^{\mathrm{opt}}(u) = \int_0^u \lambda_j^{\mathrm{opt}}(s) \, ds. \tag{9}$$

This intrinsic decoder-response allocation follows the cube-root law of scalar and functional quantization (Bennett, 1948; Lloyd, 1982; Gersho & Gray, 1992; Gray & Neuhoff, 1998; Misra et al., 2011). If $\gamma_j \equiv 0$, use uniform allocation; high-rate admissibility is specified in Theorem 1.

The train-only estimate $\widehat{\gamma}$ yields a strictly increasing piecewise-linear $\widehat{A}_j$ with monotone generalized inverse $\widehat{A}_j^{\dagger}$. Its cell boundaries, widths, and hard token are

$$\begin{aligned} b_{j,k} &= \widehat{A}_j^{\dagger}(k/K_0), & \Delta_{j,k} &= b_{j,k+1} - b_{j,k}, \\ C_j &= \min\Big\{K_0 - 1, \Big\lfloor K_0 \widehat{A}_j(\widehat{U}_j) \Big\rfloor\Big\}. \end{aligned} \tag{10}$$

These boundaries define the probability-space intervals used for uniform reconstruction. Every active coordinate retains K_0 symbols, so geometry reallocates resolution rather than alphabet size; the Identity Compander sets $\widehat{A}_j(u) = u$ at the same budget.

3.4 TWO-STAGE LEARNING AND ONE-DECODE GENERATION

Phase 1 alternates train-only calibration of $\mathcal{I} = (\widehat{T}, \widehat{F}, \widehat{A}, b, \Delta)$ with encoder–decoder updates at fixed $\mathcal{I}$. A straight-through gradient trains the encoder while preserving hard-cell samples in the forward pass (Bengio et al., 2013). Given $C = Q_{\phi,\theta}(X)$ and independent $\varepsilon \sim \mathrm{Unif}([0,1)^{m_0})$, let b_C, Δ_C collect the selected lower boundaries and widths. With $\odot$ denoting coordinatewise multiplication and sg stop-gradient, reconstruction is

$$\begin{aligned} \widetilde{U} &= b_C + \varepsilon \odot \Delta_C, & \widetilde{\Xi} &= \widehat{F}^{-1}(\widetilde{U}), \\ \widetilde{\Xi}^{\mathrm{ST}} &= \widehat{\Xi} + \mathrm{sg}(\widetilde{\Xi} - \widehat{\Xi}), & X_{\mathrm{rec}} &= D_\theta\Big(\widehat{T}^{-1}(\widetilde{\Xi}^{\mathrm{ST}})\Big). \end{aligned} \tag{11}$$

The straight-through path passes identity gradients in the calibrated canonical latent. Phase 1 combines clean reconstruction with finite-channel reconstruction, weighted by λ_{qrec}:

$$\begin{aligned}\mathcal{L}_{\mathrm{P1}} &= \mathcal{L}_{\mathrm{clean}} + \lambda_{\mathrm{qrec}}\mathcal{L}_{\mathrm{qrec}},\\ \mathcal{L}_{\mathrm{clean}} &= \mathbb{E}\,\overline{d}_g^2(D_\theta(E_\phi(X)), X)\,,\\ \mathcal{L}_{\mathrm{qrec}} &= \mathbb{E}\,\overline{d}_g^2(X_{\mathrm{rec}}, X).\end{aligned} \tag{12}$$

Each candidate is recalibrated from its training latents before validation selection; Appendix D gives the procedure.

Phase 2 freezes the selected $(\phi^*, \theta^*, \mathcal{I}^*)$ and fits their joint code law $\Pi^* = (Q^*)_\sharp \rho^*$ using an autoregressive prior (Razavi et al., 2019; Minnen et al., 2018) in canonical-coordinate order, where Q^* is the selected code map:

$$\begin{aligned}P_\psi^C(C) &= \prod_{j=1}^{m_0} p_{\psi,j}(C_j \mid C_{<j}),\\ p_{\psi,j}(k \mid C_{<j}) &= \mathrm{softmax}_k\big(r_{\psi,j,k}(C_{<j}) + \log \Delta^*_{j,k}\big)\,,\\ \mathcal{L}_{\mathrm{P2}}(\psi) &= -\frac{1}{m_0 \log K_0}\mathbb{E}_{C\sim\Pi^*}\sum_{j=1}^{m_0}\log p_{\psi,j}(C_j \mid C_{<j}).\end{aligned} \tag{13}$$

Here $C_{<j} = (C_1, \ldots, C_{j-1})$ and $r_{\psi,j,k}$ is a learned residual logit. Cell widths define reference masses under the uniform probability coordinate; residuals learn conditional departures from this reference without restricting positive categorical probabilities.

After selecting ψ^*, generation draws $C^{\mathrm{gen}} \sim P^C_{\psi^*}$ and independent $\varepsilon \sim \mathrm{Unif}([0,1)^{m_0})$, then computes

$$\begin{aligned}U^{\mathrm{gen}} &= b^*_{C^{\mathrm{gen}}} + \varepsilon \odot \Delta^*_{C^{\mathrm{gen}}},\\ X_{\mathrm{gen}} &= D_{\theta^*}\left[(\widehat{T}^*)^{-1}\Big((\widehat{F}^*)^{-1}(U^{\mathrm{gen}})\Big)\right].\end{aligned} \tag{14}$$

The unconditional sampler performs m_0 sequential categorical updates, frozen inverse transforms, and one manifold-valued decoder evaluation, without trajectory integration or decoder refinement.

4 Theoretical Analysis

Our analysis links canonical coordinates, geometric allocation, and two-stage learning to guarantees under separate assumptions. The population generation bound requires no high-rate approximation. Appendix C provides proofs.

Proposition 1 (Canonical equivalence). Consider $Z' = MZ + b$ with invertible M and decoder $D_{\theta'}(z') = D_\theta(M^{-1}(z' - b))$. With positive-definite population covariance and distinct generalized eigenvalues, the unregularized coordinates in Eq. (4) satisfy

$$T_{\phi',\theta'}(Z') = S\,T_{\phi,\theta}(Z), \qquad S = \mathrm{diag}(s_1, \ldots, s_{m_0}), \quad s_j \in \{-1, 1\}. \tag{15}$$

Canonical coordinates thus agree up to eigenvector signs; equal eigenvalues also allow rotations within their eigenspaces. Finite ridge regularization need not preserve this population equivalence.

Theorem 1 (High-rate bound and optimal scalar allocation). Fix the law of U, decoder $\mathcal{D}$, and latent dimension m_0; let $\widetilde{U}$ be the exact-cell reconstruction of U. Under the coordinatewise high-rate and hybrid-context stability conditions of Theorems C.6–C.7, as $K_0 \to \infty$,

$$\mathbb{E}\,\overline{d}_g^2\Big(\mathcal{D}(U), \mathcal{D}(\widetilde{U})\Big) \le \frac{m_0}{6K_0^2}\sum_{j=1}^{m_0}\int_0^1 \frac{\gamma_j(s)}{\lambda_j(s)^2}\,ds + o(K_0^{-2}). \tag{16}$$

The density in Eq. (9) minimizes the leading term if bounded away from zero; otherwise positive regularizations approach its infimum. This infimum never exceeds the uniform-bin value; optimality concerns this coordinatewise bound, not finite-alphabet distortion or general vector quantization.

Theorem 2 (End-to-end generation control). Freeze the selected Phase 1 representation and its calibrated interface, and assume $(\mathcal{M}, g)$ is compact. Let W_1^g denote Wasserstein-1 distance induced by d_g and $\mathrm{diam}_g(\mathcal{M})$ the geodesic diameter (Villani, 2009). For $P_\psi^C(c) > 0$ whenever $\Pi^*(c) > 0$,

$$W_1^g(\rho^*, \rho_{\theta^*,\psi}) \le a_{\mathcal{M}} \sqrt{\mathcal{L}_{\mathrm{qrec}}} + \mathrm{diam}_g(\mathcal{M}) \sqrt{\tfrac{1}{2} \mathrm{KL}(\Pi^* \| P_\psi^C)}. \tag{17}$$

Here $\mathcal{L}_{\mathrm{qrec}}$ denotes the population exact-cell reconstruction loss of the frozen Phase 1 model. With the data-code entropy fixed, minimizing population cross-entropy in Eq. (13) is equivalent to minimizing the joint KL term (Appendix C.5). Appendix C.6 extends this bound to noncompact manifolds with bounded reconstruction range.

5 Related Work

Manifold generators. Continuous flows and diffusions generate samples through numerical integration (Lipman et al., 2023; Mathieu & Nickel, 2020; Lou et al., 2020; Rozen et al., 2021; Ben-Hamu et al., 2022; De Bortoli et al., 2022; Huang et al., 2022; Jo & Hwang, 2024; Chen & Lipman, 2024). Specialized flows use geometry-specific invertible maps (Rezende et al., 2020). RiCoTA instead models finite codes and decodes continuously without integration.

Accelerated and direct generators. Consistency models, flow maps, shortcuts, and MeanFlow accelerate sampling (Song et al., 2023; Boffi et al., 2025a;b; Frans et al., 2025; Geng et al., 2025), including Riemannian extensions (Cheng et al., 2025; Davis et al., 2026; Zhong et al., 2026; Woo et al., 2026). M-FFF combines projected encoder–decoder maps, likelihood training, and reconstruction (Sorrenson et al., 2024). RiCoTA instead models interval indices autoregressively and samples continuously within cells.

Discrete latent models and quantization. High-rate and functional quantization allocate resolution to reduce distortion (Bennett, 1948; Gersho & Gray, 1992; Gray & Neuhoff, 1998; Misra et al., 2011). Quantized autoencoders and learned compression model finite codes (van den Oord et al., 2017; Razavi et al., 2019; Ballé et al., 2017; 2018; Minnen et al., 2018; Mentzer et al., 2024; Agustsson et al., 2017); universal quantization matches training and deployment noise channels (Agustsson & Theis, 2020). RiCoTA places cells using intrinsic decoder response in canonical probability coordinates, sharing uniform within-cell reconstruction across training and generation.

6 Experiments

We evaluate generation quality, sampling cost, and geometry-adapted latent resolution on the manifold benchmarks.

6.1 Experimental setup

The main suite contains 14 benchmarks: four Earth distributions on $\mathbb{S}^2$, four protein Ramachandran distributions on $\mathbb{T}^2$, RNA torsions on $\mathbb{T}^7$, four $\mathrm{SO}(3)$ distributions, and one on $\mathbb{H}^2$. Cross-method macro averages use the 13 tasks shared by all methods; $\mathbb{H}^2$ and promoter DNA (Appendix J) are reported separately. Baselines are 1000-step RFM (Chen & Lipman, 2024), one-step G-LSD (Davis et al., 2026), two-step RCM/RCT (Cheng et al., 2025), one-step RMF-MT (Zhong et al., 2026), and one-step M-FFF (Sorrenson et al., 2024). The geometric suite uses fixed splits, validation-only selection, and test-count quality evaluation. Generation-quality results average five independent training seeds per method, dataset, and variant, except promoter DNA, which averages three. MMD summarizes kernel discrepancy, sF1 multi-scale support fidelity, and HD95 geodesic tails. Table 5 uses local $F_1(0.5)$ for stricter support matching. Times are measured medians for 100,000 samples unless stated otherwise. Bold and underline rank RiCoTA and the one- or few-step baselines using unrounded values; RFM is the iterative reference above the rule. Appendices F–G specify the evaluation and timing protocols.

6.2 GENERATION QUALITY AND EFFICIENCY

Sphere. The $\mathbb{S}^2$ tasks contain earthquake, fire, flood, and volcano locations on the Earth's surface, testing sparse multimodal support under a globally curved geometry.

Table 1: Generation quality and efficiency on $\mathbb{S}^2$. HD95 is in degrees; time is the mean of per-dataset median times (ms) for 100k samples across the four datasets.

Method	Eq.			Fire			Flood			Volc.			Avg. E2E
	sF1↑	HD95↓	MMD↓	sF1↑	HD95↓	MMD↓	sF1↑	HD95↓	MMD↓	sF1↑	HD95↓	MMD↓	100k ms↓
RFM	0.907	4.881	0.025	0.908	2.016	0.025	0.911	6.360	0.067	0.922	19.42	0.113	382141
G-LSD	0.857	5.556	0.041	0.806	4.219	0.031	0.840	9.132	**0.038**	0.886	20.58	0.142	468.90
RCM/RCT	0.850	9.841	0.052	0.791	8.606	0.046	0.868	9.377	0.073	0.890	21.10	0.152	130.16
RMF-MT	0.794	9.310	0.051	0.617	12.09	0.052	0.808	12.30	0.043	0.729	21.88	0.156	469.40
M-FFF	0.856	9.860	0.042	0.825	8.003	0.030	0.878	10.74	0.069	0.890	19.16	0.147	18.55
RiCoTA	**0.898**	**5.221**	**0.027**	**0.866**	**2.210**	**0.023**	**0.900**	**7.626**	0.044	**0.927**	**15.66**	**0.084**	**9.33**

Compared with the one- and few-step baselines, RiCoTA leads all four sphere tasks in sF1 and HD95 and three in MMD; G-LSD and RMF-MT have lower Flood MMD. The gains therefore also appear in support and geodesic-tail metrics. Mean generation time is 9.33 ms, versus 18.55 ms for M-FFF. Figure 2 compares Earthquake samples using a shared Mediterranean/West Asia zoom.

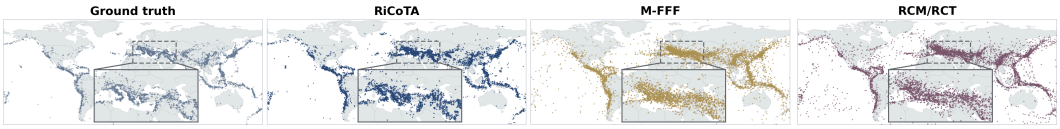


Figure 2: Earthquake on $\mathbb{S}^2$. Ground truth and $4{,}896$ generations per method. Every panel magnifies the same Mediterranean/West Asia region; RiCoTA closely follows the observed support.

Special orthogonal group $\mathrm{SO}(3)$. These tasks test separated rotational modes and thin supports. Support and tail metrics complement the global kernel score.

Table 2: Generation quality and efficiency on $\mathrm{SO}(3)$. HD95 is in degrees; time is the mean of per-dataset median times (ms) for 100k samples across the four datasets.

Method	Cone			Fisher24			Line			Swiss			Avg. E2E
	sF1↑	HD95↓	MMD↓	sF1↑	HD95↓	MMD↓	sF1↑	HD95↓	MMD↓	sF1↑	HD95↓	MMD↓	100k ms↓
RFM	0.090	34.42	0.105	0.150	50.89	0.129	0.657	27.88	0.080	0.257	96.58	0.201	13892
G-LSD	0.055	64.47	0.152	0.161	46.24	0.087	0.667	26.28	0.076	0.127	129.40	0.523	77.12
RCM/RCT	0.520	49.40	0.098	0.329	43.73	0.087	0.688	22.54	0.063	0.856	2.461	0.131	137.68
RMF-MT	0.090	54.67	0.096	0.163	46.23	0.091	0.674	24.87	0.055	0.234	74.97	0.229	101.58
M-FFF	0.784	27.35	0.039	0.518	36.09	0.040	0.691	23.89	0.025	0.051	149.87	0.765	86.93
RiCoTA	**0.809**	**1.237**	**0.019**	**0.836**	**4.020**	**0.022**	**0.844**	**13.09**	**0.024**	**0.906**	**1.219**	**0.007**	**30.54**

RiCoTA has the best sF1, HD95, and MMD on all four rotation tasks in Table 2. On Fisher24, RiCoTA achieves sF1 0.836 and HD95 $4.020°$, compared with 0.518 and $36.09°$ for M-FFF, the strongest fast baseline on these metrics. Mean 100k time is 30.54 ms versus 77.12 ms for the next-fastest method, G-LSD. Figure 3 visualizes the separated target modes.

Product tori and $\mathbb{H}^2$. Protein Ramachandran angles and RNA torsions test periodic ridges and multi-angle dependence on $\mathbb{T}^2$ and $\mathbb{T}^7$. The $\mathbb{H}^2$ benchmark tests generation on the negatively curved hyperbolic plane, extending evaluation beyond compact manifolds.

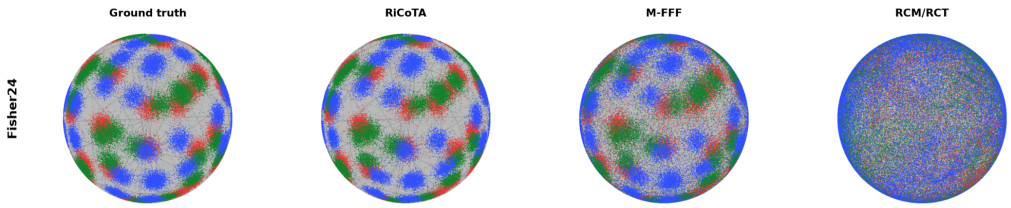


Figure 3: Fisher24 on $\mathrm{SO}(3)$. Rotated-axis projections compare the target and generated samples. Table 2 quantifies fidelity in the full rotation space.

Table 3: Generation quality and efficiency on product tori. HD95 is in degrees; time is the mean of per-dataset medians for four protein tasks in (a) and the RNA median in (b), in ms for 100k samples.

(a) Protein $\mathbb{T}^2$

Method	General			Gly.			Pro.			Pre-Pro			Avg. E2E
	sF1↑	HD95↓	MMD↓	sF1↑	HD95↓	MMD↓	sF1↑	HD95↓	MMD↓	sF1↑	HD95↓	MMD↓	100k ms↓
RFM	0.909	1.791	0.007	0.900	9.161	0.040	0.904	4.654	0.041	0.910	5.480	0.037	16082
G-LSD	0.890	2.753	0.015	0.885	11.58	0.038	0.855	7.309	0.041	0.858	19.80	0.053	32.71
RCM/RCT	0.876	16.92	0.124	0.857	24.66	0.099	0.762	45.56	0.223	0.761	60.07	0.280	129.59
RMF-MT	0.816	13.66	0.053	0.868	12.79	0.040	0.814	16.09	0.055	0.822	34.47	0.076	32.61
M-FFF	0.906	2.285	0.014	0.902	8.728	0.038	0.890	8.332	0.023	0.891	14.25	0.034	22.72
RiCoTA	**0.906**	**1.893**	**0.009**	**0.904**	**8.711**	**0.023**	**0.894**	**5.511**	**0.018**	**0.902**	**5.492**	**0.031**	**6.00**

(b) RNA $\mathbb{T}^7$

Method	sF1↑	HD95↓	MMD↓	E2E 100k ms↓
RFM	0.757	60.41	0.033	18023
G-LSD	0.615	98.88	0.217	36.35
RCM/RCT	0.545	162.16	0.547	130.07
RMF-MT	0.525	152.80	0.196	36.37
M-FFF	0.580	182.86	0.425	23.01
RiCoTA	**0.743**	**64.57**	**0.040**	**7.19**

Table 4: Generation quality and efficiency on $\mathbb{H}^2$. HD95 is in curvature-radius units; time is for 100k samples (ms).

Method	sF1↑	HD95↓	MMD↓	E2E 100k ms↓
RFM	0.898	0.183	0.091	51882
G-LSD	0.899	0.146	0.099	30.65
RiCoTA	**0.905**	**0.132**	**0.023**	**7.60**

Among fast methods, RiCoTA leads the protein and RNA tasks in all three quality metrics and generation time. RNA sF1 reaches 0.743 versus 0.615 for the best fast baseline, approaching the iterative RFM reference (0.757). RFM retains lower HD95 on three protein tasks and RNA. RiCoTA also improves all reported $\mathbb{H}^2$ quantities over its available baselines. Appendix I gives qualitative comparisons.

6.3 Component analysis across manifolds

Full–Identity tests the end-to-end effect of response-aware allocation at fixed m_0 and K_0; Full–Clean-Only tests training through the finite reconstruction channel.

Table 5: Geometry-family component comparison with local F_1 at $\lambda = 0.5$. Dataset means are equally weighted within families; product tori combine protein and RNA. HD95 uses degrees on angular manifolds and curvature-radius units on $\mathbb{H}^2$.

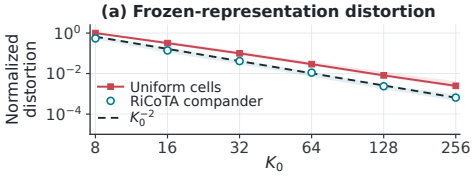
Variant	$\mathbb{S}^2$			$\mathrm{SO}(3)$			Product tori			$\mathbb{H}^2$		
	$F_1(0.5)$ ↑	HD95↓	MMD↓	$F_1(0.5)$ ↑	HD95↓	MMD↓	$F_1(0.5)$ ↑	HD95↓	MMD↓	$F_1(0.5)$ ↑	HD95↓	MMD↓
Full RiCoTA	**0.700**	**7.678**	**0.0445**	**0.478**	**4.893**	**0.0180**	**0.550**	**17.236**	**0.0238**	**0.659**	**0.132**	**0.0228**
Clean-Only	0.671	8.426	0.0470	0.469	5.115	0.0207	0.544	18.488	0.0298	0.652	0.145	0.0273
Identity	0.675	9.495	0.0516	0.465	5.180	0.0198	0.540	19.379	0.0313	0.650	0.140	0.0294

Full improves all three family-averaged metrics over both controls, supporting cell allocation and finite-channel training without increasing the code budget (Appendix H.1).

6.4 LATENT-RESOLUTION SCALING

Train-fitted response predicts held-out coordinatewise intrinsic risk (Appendix H.2). We freeze the encoder, decoder, gauge, and CDFs and compare response-aware with uniform cells at $K_0 = 8$–256, without Phase 2. Response-aware cells lower mean coordinatewise intrinsic distortion in all 81 settings with training support in every active calibration interval, at matched m_0, K_0 (Appendix H.4). The mean fitted decay exponent is 1.92, consistent with the high-rate prediction in Eq. (8).

A separate sweep retrains both stages on four representative datasets with all non-K_0 hyperparameters fixed. Mean sF1 improves from small alphabets and largely saturates at larger K_0 (Figure 4(b)); other generation metrics need not improve monotonically. The frozen comparison supports improved allocation at a fixed alphabet size, while the retrained sweep shows how generation quality changes with resolution. Precomputed boundaries preserve the allocation without response evaluation or additional decoder calls. Appendices H.4 and H.5 give full resolution and exponent sweeps.



(b) End-to-end support fidelity

K_0	8	16	32	64	128	256
Fire	0.722	0.792	0.837	0.866	0.886	0.885
General ($\mathbb{T}^2$)	0.761	0.868	0.891	0.907	0.906	0.906
Swiss (SO(3))	0.855	0.892	0.903	0.904	0.905	0.905
$\mathbb{H}^2$	0.854	0.894	0.906	0.900	0.907	0.903

Figure 4: Latent-resolution scaling. (a) Normalized coordinatewise distortion over the 13 datasets supported at every resolution, excluding Volcano and including $\mathbb{H}^2$ (median and interquartile range). (b) End-to-end support fidelity.

7 LIMITATIONS AND FUTURE WORK

Our distortion and generation-error bounds apply under the fixed-representation and population assumptions stated with each result. Extending them to jointly learned representations and finite-sample interface estimates remains an open direction. The interface uses coordinatewise cells after canonicalization, while the prior models their dependence. Code-generation cost grows with latent width and alphabet size. The geometric efficiency results concern complete sampling under the stated benchmark configurations. Generation includes sequential code sampling, inverse transforms, and decoding, while interface calibration is performed during training.

The promoter experiment in Appendix J explores a task-specific conditional adaptation that retains the finite-code interface, within-cell sampling, and one decoder evaluation. Its measured sequence quality and generation time support the feasibility of using this interface within a structured conditional model for high-dimensional sequences. Building on this case study, future work can compare additional allocation schemes and prior designs under matched architectures and training budgets. Another direction is to examine how generation quality, calibration cost, and code-generation depth vary with latent dimensionality.

8 CONCLUSION

RiCoTA treats finite latent resolution as a geometric design choice for continuous manifold generation. Canonical probability coordinates and response-aware cells allocate a fixed code budget according to intrinsic decoder sensitivity. An autoregressive prior models joint code dependence, and a shared exact-cell reconstruction rule aligns training and generation. Sampling combines sequential code generation with one manifold-valued decoder evaluation, avoiding trajectory integration. For a fixed representation, our analysis identifies the allocation minimizing the leading coordinatewise high-rate distortion bound under stated conditions. A separate population bound connects generation error to quantized reconstruction and joint code-law mismatch. Frozen-representation diagnostics isolate cell-placement effects, while end-to-end controls evaluate allocation and finite-channel training. The alphabet-size study links observed reconstruction trends to the high-rate prediction across finite code budgets. Across 14 manifold benchmarks, the results support geometry-adapted resolution for strong support fidelity at low measured sampling cost.

AI USE STATEMENT

The authors used generative AI tools to assist with language editing and polishing, literature retrieval and discovery, drafting and revising portions of the manuscript, and reviewing mathematical proofs and derivations. These tools were also used to support the authors in reviewing datasets and evaluation protocols, interpreting experimental results, examining baseline-comparison settings, and developing and reviewing implementation code. All AI-assisted content retained in this work was reviewed by the authors, who take full responsibility for the final manuscript, including its claims, results, and accompanying artifacts.

REPRODUCIBILITY STATEMENT

We provide the finite-interface construction and numerical details in Appendix B, the assumptions and proofs of the theoretical results in Appendix C, detailed training and sampling procedures in Appendix D, and manifold-specific implementation details in Appendix E. Appendix F specifies the datasets, preprocessing, model-selection procedures, hyperparameter settings, evaluation metrics, and timing protocol. Additional quantitative results, diagnostic experiments, and qualitative comparisons are reported in Appendices G–I, and Appendix J describes the conditional promoter extension. Anonymous code is provided with the submission.

ETHICS STATEMENT

All experiments are computational and use publicly available benchmark datasets or synthetic benchmark data generated according to published protocols. The conditional promoter experiment evaluates generative modeling on biological sequences using computational metrics. No wet-lab experiments were conducted, and we make no claims about the biological functionality or clinical utility of the generated sequences. Any downstream biological application would require appropriate domain-specific validation, safety assessment, and oversight.

REFERENCES

Eirikur Agustsson and Lucas Theis. Universally quantized neural compression. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12367–12376, 2020.

Eirikur Agustsson, Fabian Mentzer, Michael Tschannen, Lukas Cavigelli, Radu Timofte, Luca Benini, and Luc Van Gool. Soft-to-hard vector quantization for end-to-end learning compressible representations. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: On the curvature of deep generative models. In *International Conference on Learning Representations*, 2018.

Pavel Avdeyev, Chenlai Shi, Yuhao Tan, Kseniia Dudnyk, and Jian Zhou. Dirichlet diffusion score model for biological sequence generation. In *International Conference on Machine Learning*, pp. 1276–1301. PMLR, 2023.

Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. End-to-end optimized image compression. In *International Conference on Learning Representations*, 2017.

Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018.

Heli Ben-Hamu, Samuel Cohen, Joey Bose, Brandon Amos, Aditya Grover, Maximilian Nickel, Ricky T. Q. Chen, and Yaron Lipman. Matching normalizing flows and probability paths on manifolds. In *International Conference on Machine Learning*, 2022.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

William Ralph Bennett. Spectra of quantized signals. *The Bell System Technical Journal*, 27(3): 446–472, 1948.

Nicholas M. Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. Flow map matching with stochastic interpolants: A mathematical framework for consistency models. *Transactions on Machine Learning Research*, 2025a. ISSN 2835-8856.

Nicholas M. Boffi, Michael S. Albergo, and Eric Vanden-Eijnden. How to build a consistency model: Learning flow maps via self-distillation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 33346–33382, 2025b.

G. R. Brakenridge. Global active archive of large flood events. Dartmouth Flood Observatory, University of Colorado, 2017. URL `https://floodobservatory.colorado.edu/Archives/index.html`.

Kathleen M. Chen, Aaron K. Wong, Olga G. Troyanskaya, and Jian Zhou. A sequence-based global map of regulatory activity for deciphering human genetics. *Nature Genetics*, 54(7):940–949, 2022.

Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries. In *International Conference on Learning Representations*, volume 2024, pp. 47922–47945, 2024.

Chaoran Cheng, Yusong Wang, Yuxin Chen, Xiangxin Zhou, Nanning Zheng, and Ge Liu. Riemannian consistency model. In *Advances in Neural Information Processing Systems*, 2025.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2nd edition, 2006.

Nathaël Da Costa, Cyrus Mostajeran, and Juan-Pablo Ortega. The Gaussian kernel on the circle and spaces that admit isometric embeddings of the circle. In *International Conference on Geometric Science of Information*, pp. 426–435. Springer, 2023.

Oscar Davis, Samuel Kessler, Mircea Petrache, İsmail İlkan Ceylan, Michael Bronstein, and Avishek Joey Bose. Fisher flow matching for generative modeling over discrete data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 139054–139084, 2024.

Oscar Davis, Michael S. Albergo, Nicholas M. Boffi, Michael M. Bronstein, and Avishek Joey Bose. Generalised flow maps for few-step generative modelling on Riemannian manifolds. In *International Conference on Learning Representations*, volume 2026, pp. 31119–31140, 2026.

Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2406–2422, 2022.

Luca Falorsi, Pim de Haan, Tim R. Davidson, and Patrick Forré. Reparameterizing distributions on Lie groups. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3244–3253. PMLR, 2019.

Haoqiang Fan, Hao Su, and Leonidas J. Guibas. A point set generation network for 3D object reconstruction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 605–613, 2017.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *International Conference on Learning Representations*, volume 2025, pp. 34668–34684, 2025.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *Advances in Neural Information Processing Systems*, 2025.

Allen Gersho and Robert M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic Publishers, 1992.

Robert M. Gray and David L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44 (6):2325–2383, 1998.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13:723–773, 2012.

Nicholas J. Higham. Stable iterations for the matrix square root. *Numerical Algorithms*, 15(2): 227–242, 1997.

Chin-Wei Huang, Milad Aghajohari, Joey Bose, Prakash Panangaden, and Aaron Courville. Riemannian diffusion models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2750–2761, 2022.

Jaehyeong Jo and Sung Ju Hwang. Generative modeling on manifolds through mixture of Riemannian diffusion processes. In *International Conference on Machine Learning*. PMLR, 2024.

Agnan Kessy, Alex Lewin, and Korbinian Strimmer. Optimal whitening and decorrelation. *The American Statistician*, 72(4):309–314, 2018.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Yulin Liu, Haoran Liu, Yingda Yin, Yang Wang, Baoquan Chen, and He Wang. Delving into discrete normalizing flows on SO(3) manifold for probabilistic rotation modeling. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21264–21273. IEEE, 2023.

Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28 (2):129–137, 1982.

Aaron Lou, Derek Lim, Isay Katsman, Leo Huang, Qingxuan Jiang, Ser Nam Lim, and Christopher M. De Sa. Neural manifold ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 17548–17558, 2020.

Simon C. Lovell, Ian W. Davis, W. Bryan Arendall, III, Paul I. W. de Bakker, J. Michael Word, Michael G. Prisant, Jane S. Richardson, and David C. Richardson. Structure validation by Cα geometry: ϕ, ψ and Cβ deviation. *Proteins*, 50(3):437–450, 2003.

Kanti V. Mardia and Peter E. Jupp. *Directional Statistics*. Wiley, 2000.

Emile Mathieu and Maximilian Nickel. Riemannian continuous normalizing flows. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2503–2515, 2020.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: VQ-VAE made simple. In *International Conference on Learning Representations*, volume 2024, pp. 51772–51783, 2024.

David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Vinith Misra, Vivek K Goyal, and Lav R Varshney. Distributed scalar quantization for computing: High-resolution analysis and extensions. *IEEE Transactions on Information Theory*, 57(8):5298–5325, 2011.

Laura J. W. Murray, W. Bryan Arendall, III, David C. Richardson, and Jane S. Richardson. RNA backbone is rotameric. *Proceedings of the National Academy of Sciences*, 100(24):13904–13909, 2003.

NASA Earth Science Data and Information System. Active fire data. Land, Atmosphere Near real-time Capability for EOS, 2020.

NOAA National Centers for Environmental Information. Global significant earthquake database. National Geophysical Data Center / World Data Service, 2020a.

NOAA National Centers for Environmental Information. Global significant volcanic eruptions database. National Geophysical Data Center / World Data Service, 2020b.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with VQ-VAE-2. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Danilo Jimenez Rezende, George Papamakarios, Sébastien Racaniere, Michael Albergo, Gurtej Kanwar, Phiala Shanahan, and Kyle Cranmer. Normalizing flows on tori and spheres. In *International Conference on Machine Learning*, pp. 8083–8092. PMLR, 2020.

Noam Rozen, Aditya Grover, Maximilian Nickel, and Yaron Lipman. Moser flow: Divergence-based generative modeling on manifolds. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17669–17680, 2021.

Peter H Schönemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 2023.

Peter Sorrenson, Felix Draxler, Armand Rousselot, Sander Hummerich, and Ullrich Köthe. Learning distributions on manifolds with free-form flows. In *Advances in Neural Information Processing Systems*, volume 37, pp. 66961–66994, 2024.

Hannes Stark, Bowen Jing, Chenyu Wang, Gabriele Corso, Bonnie Berger, Regina Barzilay, and Tommi Jaakkola. Dirichlet flow matching with applications to DNA sequence design. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 2024.

Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging*, 15(1):29, 2015.

The FANTOM Consortium and the RIKEN PMI and CLST (DGT). A promoter-level mammalian expression atlas. *Nature*, 507:462–470, 2014.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Cédric Villani. *Optimal Transport: Old and New*, volume 338. Springer, 2009.

Dongyeop Woo, Marta Skreta, Seonghyun Park, Kirill Neklyudov, and Sungsoo Ahn. Riemannian MeanFlow. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

Zichen Zhong, Haoliang Sun, Yukun Zhao, Yongshun Gong, and Yilong Yin. Riemannian MeanFlow for one-step generation on manifolds. In *International Conference on Machine Learning*, 2026.

APPENDIX CONTENTS

A Appendix Organization and Notation **16**

B Finite-Interface Construction and Numerical Details **16**

B.1 Training statistics and decoder pullback . . . 17
B.2 Canonical frame and alignment . . . 17
B.3 Empirical probability coordinates . . . 17
B.4 Paired response estimation . . . 18
B.5 Companding and exact-cell reconstruction . . . 18
B.6 Numerical conventions . . . 19

C Proofs and Additional Theoretical Results **19**

C.1 Local geometry and canonical equivalence . . . 19
C.2 Probability coordinates and response estimation . . . 20
C.3 Scalar high-rate allocation . . . 21
C.4 Joint reconstruction control . . . 23
C.5 Code-law learning and generation control . . . 24
C.6 Noncompact extension with bounded reconstruction range . . . 26

D Algorithms and Training Details **27**

D.1 Phase 1 training and selection . . . 27
D.2 Phase 2 training and selection . . . 27
D.3 Generation and computational cost . . . 28

E Manifold-Specific Implementations **28**

E.1 Sphere . . . 28
E.2 Flat torus . . . 29
E.3 Rotation group . . . 29
E.4 Hyperbolic plane . . . 29

F Experimental Details and Reproducibility **29**

F.1 Datasets and preprocessing . . . 29
F.2 Validation-based selection and controls . . . 30
F.3 Architecture and optimization settings . . . 30
F.4 Baseline implementations . . . 30
F.5 Evaluation metrics and statistical reporting . . . 31
F.6 Generation-time measurement . . . 32

G Additional Quantitative Results and Efficiency Analysis **33**

G.1 Distribution and support . . . 33
G.2 Geodesic shape and tail coverage . . . 34
G.3 Per-dataset latency and hyperbolic results . . . 34

H Ablations and Interface Diagnostics **35**

H.1 Component comparisons across geometry families . . . 35
H.2 Held-out response and allocation diagnostics . . . 36
H.3 Total coordinatewise diagnostic distortion . . . 38
H.4 Latent-resolution scaling . . . 39
H.5 Sensitivity to the companding exponent . . . 43

I Qualitative Comparisons **45**

I.1 Spherical Earth data . . . 45
I.2 Protein torsion distributions . . . 47
I.3 Rotation distributions . . . 49

J Case study: high-dimensional conditional promoter generation **51**

J.1 Task, data, and representation . . . 51
J.2 Conditioning and spatial-grid autoencoder . . . 51
J.3 Conditional geometry and interface calibration . . . 51
J.4 Spectral-prefix conditional prior and training objective . . . 52
J.5 Sequence evaluation and generation time . . . 53

A Appendix Organization and Notation

Implementation details are in Appendices B, D, and E; proofs are in Appendix C. Appendices F–I cover evaluation and additional results, followed by the conditional DNA case study in Appendix J. Section 7 discusses limitations and future directions.

For the core interface quantities, hats indicate training estimates or maps constructed from them, and tildes indicate within-cell reconstructions. The labels gen and opt denote generated quantities and variational optima. For learned model and interface quantities, $*$ denotes the selected, frozen checkpoint; ρ^* is the data distribution. In the interface construction, i, j, r, k index observations, latent coordinates, pilot intervals, and tokens; t indexes interface refreshes.

Table 6: Core notation used in the main text and appendices.

Symbol	Meaning
$(\mathcal{M}, g)$; $T_x\mathcal{M}$	Connected Riemannian space; tangent space at x
$g_x, \lVert\cdot\rVert_{g,x}$	Tangent inner product and norm
$d_g, \bar{d}_g = d_g/a_{\mathcal{M}}$	Intrinsic and normalized distance
$a_{\mathcal{M}} > 0$	Training-data intrinsic distance scale (Appendix B.1)
$W_1^g, \mathrm{diam}_g, \mathrm{TV}$	Wasserstein-1 distance, diameter, total variation
$X \sim \rho^*$; E_ϕ, D_θ	Data law; encoder and manifold-valued decoder
$Z = E_\phi(X)$; m_0, K_0	Latent; coordinate count and symbols per active coordinate
$\mathcal{C} = \{0, \dots, K_0 - 1\}^{m_0}$	Finite code space
μ_ϕ, Σ_ϕ	Population latent mean and covariance
$G_\theta(z), \overline{G}_{\phi,\theta}$	Decoder pullback and its population average
$T, \Xi = T(Z)$	Population canonical map and coordinates
F, $U = F(\Xi)$	Population marginal CDFs and probability coordinates
$\mathcal{D}$	The decoder D_θ expressed in population probability coordinates
$\widehat{\mu}$	Training-sample mean of the encoder latents
$\widehat{\Sigma}_\epsilon$	Regularized training-sample latent covariance
$\widehat{\overline{G}}$	Training estimate of the mean decoder pullback
$\widehat{T}$	Train-fitted, aligned canonical-coordinate map
$\widehat{F}$	Coordinatewise interpolated training mid-rank maps
$\widehat{\mathcal{D}}$	The same D_θ composed with $\widehat{T}^{-1}$ and $\widehat{F}^{-1}$, as in Eq. (6)
$\gamma_j(s), \widehat{\gamma}_{j,r}$	Conditional response and paired-secant estimate
$\lambda_j^{\mathrm{opt}}, A_j^{\mathrm{opt}}$	Optimal point density and compander
$\widehat{A}_j, b_{j,k}, \Delta_{j,k}$	Fitted compander, boundary, and width
C, Q, Rec	Hard code, code map, and stochastic reconstruction
$\widetilde{U}$	Reconstructed probability coordinates
$\widetilde{\Xi}$	Reconstructed canonical coordinates
X_{rec}	Decoded manifold reconstruction
$\mathcal{I}$	Fitted coordinate maps, responses, cells, active mask, and distance scale
Π^*, P_ψ^C	Frozen data-code law and learned joint prior
$\mathcal{L}_{\mathrm{clean}}, \mathcal{L}_{\mathrm{qrec}}$	Clean and exact-cell reconstruction losses
$\mathcal{L}_{\mathrm{P1}}, \mathcal{L}_{\mathrm{P2}}$	Representation and normalized prior objectives

Conventions. All analytical maps are measurable; expectations are taken where the stated moment conditions hold. Logarithms are natural and $\mathrm{TV}(P, Q) = \frac{1}{2}\sum_c |P(c) - Q(c)|$ for code laws. We take $m_0 \geq 1$ and $K_0 \geq 2$. Each result states its own regularity assumptions. Numerical tolerances in Appendix B are implementation rules, not assumptions on ρ^*.

B Finite-Interface Construction and Numerical Details

For a fixed encoder–decoder pair, calibration fits the gauge, probability maps, response, and cells on training data only. Validation selects candidates; test data is reserved for evaluation. The selected interface is reused without refitting at generation.

B.1 Training statistics and decoder pullback

For the geometric benchmarks, let $x_1, \dots, x_N$ be the training observations and $z_i = E_\phi(x_i)$. The training statistics are

$$\widehat{\mu} = \frac{1}{N}\sum_i z_i, \qquad \widehat{\Sigma} = \frac{1}{N-1}\sum_i (z_i - \widehat{\mu})(z_i - \widehat{\mu})^\top, \qquad \widehat{\Sigma}_\epsilon = \widehat{\Sigma} + \epsilon_\Sigma \frac{\mathrm{tr}(\widehat{\Sigma})}{m_0} I + \epsilon_0 I. \quad (18)$$

The fixed ridges in Appendix B.6 make the estimate positive definite. The intrinsic scale and normalized distance are

$$a_{\mathcal{M}} = \mathrm{median}_i \, d_g(x_i, \mathrm{NN}_5(x_i; \{x_\ell\}_{\ell \neq i})) > 0, \qquad \overline{d}_g = d_g / a_{\mathcal{M}}. \quad (19)$$

The fifth neighbor excludes the query itself and requires $N \geq 6$. The normalization is defined only for positive $a_{\mathcal{M}}$; the same intrinsic distance is used for reconstruction and reporting. A common positive rescaling of d_g and $a_{\mathcal{M}}$ leaves $\overline{d}_g$ and the compander unchanged.

Using the pullback in Eq. (3), the geometric gauge-bank average is

$$\widehat{\overline{G}}_{\phi,\theta} = \frac{1}{N_G}\sum_{i \in \mathcal{G}} G_\theta(z_i), \qquad |\mathcal{G}| = N_G. \quad (20)$$

Automatic differentiation forms the intrinsic Gram matrix, including the manifold-specific support map or tangent-coordinate conversion. Gauge points and response contexts are drawn from training data using predeclared refresh-specific seeds. The interface is fixed within each representation-update block and recalibrated for the next block (Appendix D.1). The DNA extension uses the sampled scale and stochastic pullback estimators described in Appendix J.3.

B.2 Canonical frame and alignment

Let $\widehat{\Sigma}_\epsilon = LL^\top$. Diagonalize

$$L^\top \widehat{\overline{G}}_{\phi,\theta} L = V \Lambda V^\top, \quad V^\top V = I, \quad \Lambda = \mathrm{diag}(\eta_1, \dots, \eta_{m_0}), \quad \eta_1 \geq \cdots \geq \eta_{m_0}. \quad (21)$$

Before alignment, $W = LV$ satisfies $W^\top \widehat{\Sigma}_\epsilon^{-1} W = I$ and $\widehat{\overline{G}}_{\phi,\theta} w_j = \eta_j \widehat{\Sigma}_\epsilon^{-1} w_j$.

A fixed relative tolerance groups adjacent near-tied eigenvalues. At refresh $t \geq 1$, let $W_t^{(\mathcal{B})}$ be the newly computed frame for block $\mathcal{B}$ and $W_{t-1,\mathrm{al}}^{(\mathcal{B})}$ the corresponding previously aligned frame. Align them by

$$R_{\mathcal{B}}^{\mathrm{opt}} = \operatorname*{arg\,min}_{R^\top R = I} \|W_t^{(\mathcal{B})} R - W_{t-1,\mathrm{al}}^{(\mathcal{B})}\|_F^2. \quad (22)$$

An SVD $(W_t^{(\mathcal{B})})^\top W_{t-1,\mathrm{al}}^{(\mathcal{B})} = U_B S_B V_B^\top$ gives $R_{\mathcal{B}}^{\mathrm{opt}} = U_B V_B^\top$ (Schönemann, 1966). Singleton signs align with their predecessors. At the first calibration, every axis, including those within tied blocks, is oriented to nonnegative empirical skewness; when the skewness magnitude is below tolerance, its largest-magnitude loading is made positive. The resulting sign-oriented frame is stored as the initial reference.

Let R collect these block rotations and signs, and use the aligned factors $V_{\mathrm{al}} = VR$ and $W_{\mathrm{al}} = LV_{\mathrm{al}}$ in both directions:

$$\widehat{T}(z) = V_{\mathrm{al}}^\top L^{-1}(z - \widehat{\mu}), \qquad \widehat{T}^{-1}(\xi) = \widehat{\mu} + LV_{\mathrm{al}}\xi. \quad (23)$$

Triangular solves implement L^{-1}. Alignment preserves regularized covariance-metric orthonormality, whereas $W_{\mathrm{al}}^\top \widehat{\overline{G}} W_{\mathrm{al}} = R^\top \Lambda R$ remains exactly diagonal for exact ties but need not do so for numerical near-ties. Blocks retain decreasing eigenvalue order; coordinates within a near-tied block follow the stored alignment. This is not exact whitening of the unregularized sample covariance. Population affine equivalence is treated separately in Appendix C.1.

B.3 Empirical probability coordinates

Sort each canonical training coordinate and assign mid-ranks $(i - \frac{1}{2})/N$. Average ranks for duplicate values to obtain M_j unique knots $(a_{j,\ell}, p_{j,\ell})$ with both knot sequences strictly increasing. The fitted

$\widehat{F}_j$ is a piecewise-linear mid-rank map, not a step empirical CDF: interpolate between knots and clamp outside to the endpoint probabilities. Its inverse swaps the knot axes and clips probabilities:

$$\widehat{F}_j^{-1}(u) = \mathrm{interp}\Big(\mathrm{clip}(u, p_{j,1}, p_{j,M_j}); (p_{j,\ell}, a_{j,\ell})_{\ell=1}^{M_j}\Big). \tag{24}$$

For active coordinates, the two maps are inverses on the knot range in exact arithmetic. No exact uniformity on unseen data is assumed.

If $a_{j,M_j} - a_{j,1} \le \tau_{\mathrm{range}}$, mark coordinate j inactive and bypass response estimation. Its token is the sentinel zero, distinguished from an active coordinate's first cell by the mask. Inactive reconstruction uses $U_j = 1/2$ and $\Xi_j = \widehat{F}_j^{-1}(1/2)$ without noise. Inactive prior factors are deterministic and omitted from the loss sum; the denominator $m_0 \log K_0$ is retained.

B.4 Paired response estimation

Each active coordinate uses $R_0 = K_0$ uniform pilot intervals and paired probes:

$$I_r = [r/K_0, (r+1)/K_0), \quad s_r^- = (r + \tfrac{1}{4})/K_0, \quad s_r^+ = (r + \tfrac{3}{4})/K_0, \quad h_r = s_r^+ - s_r^- = \frac{1}{2K_0}. \tag{25}$$

With $u^{(i)} = \widehat{F}(\widehat{T}(z_i))$, set $\mathcal{B}_{j,r} = \{i : u_j^{(i)} \in I_r\}$ and select $\mathcal{S}_{j,r} \subseteq \mathcal{B}_{j,r}$ uniformly without replacement, using all contexts when there are at most M_C. Writing $n_{j,r} = |\mathcal{S}_{j,r}| = \min(M_C, |\mathcal{B}_{j,r}|)$,

$$\widehat{\gamma}_{j,r} = \frac{1}{n_{j,r}} \sum_{i \in \mathcal{S}_{j,r}} \left[\frac{\overline{d}_g\Big(\widehat{\mathcal{D}}(u_{-j}^{(i)} \oplus_j s_r^-), \widehat{\mathcal{D}}(u_{-j}^{(i)} \oplus_j s_r^+)\Big)}{h_r} \right]^2. \tag{26}$$

Here $\oplus_j$ inserts coordinate j. Both probes retain the same non-j context, so the secant measures only coordinate-j variation. The response-probe decoder evaluations do not retain a gradient graph for parameter updates. An unsupported active interval invalidates calibration; neighboring responses are not imputed.

Let $\bar{\gamma}_j = K_0^{-1} \sum_r \widehat{\gamma}_{j,r}$. If $\bar{\gamma}_j \le \tau_{\mathrm{response}}$, apply the same inactive rule as above. Otherwise use

$$\gamma_{j,r}^{\mathrm{floor}} = \widehat{\gamma}_{j,r} + \epsilon_{\mathrm{num}} \bar{\gamma}_j, \qquad \epsilon_{\mathrm{num}} = 64\epsilon_{\mathrm{fp64}}, \tag{27}$$

which gives positive compander slopes in exact arithmetic.

B.5 Companding and exact-cell reconstruction

Uniform pilot intervals estimate response; the final token cells are obtained by inverting the fitted compander. Define $q_{j,r} = (\gamma_{j,r}^{\mathrm{floor}})^{1/3}$, $Z_j^{\mathrm{cmp}} = K_0^{-1} \sum_r q_{j,r}$, and $\alpha_{j,r} = (K_0 Z_j^{\mathrm{cmp}})^{-1} \sum_{\ell<r} q_{j,\ell}$. For $u \in I_r$ and $y \in [\alpha_{j,r}, \alpha_{j,r+1}]$,

$$\widehat{A}_j(u) = \alpha_{j,r} + \frac{(u - r/K_0) q_{j,r}}{Z_j^{\mathrm{cmp}}}, \qquad \widehat{A}_j^{\dagger}(y) = \frac{r}{K_0} + \frac{(y - \alpha_{j,r}) Z_j^{\mathrm{cmp}}}{q_{j,r}}. \tag{28}$$

Set $b_{j,k} = \widehat{A}_j^{\dagger}(k/K_0)$ and $\Delta_{j,k} = b_{j,k+1} - b_{j,k}$. Overwrite $\widehat{A}_j(0) = b_{j,0} = 0$ and $\widehat{A}_j(1) = b_{j,K_0} = 1$ exactly, and reject nonfinite or nonpositive active widths. The Identity Compander uses $\widehat{A}_j(u) = u$ with all other components unchanged.

For independent $\varepsilon_j \sim \mathrm{Unif}[0,1)$, hard tokens and dequantization are

$$C_j = \min\{K_0 - 1, \lfloor K_0 \widehat{A}_j(\widehat{U}_j) \rfloor\}, \quad \widetilde{U}_j = b_{j,C_j} + \varepsilon_j \Delta_{j,C_j}, \quad \widetilde{\Xi}_j = \widehat{F}_j^{-1}(\widetilde{U}_j). \tag{29}$$

Cells are left-closed and right-open, except the last includes one. The Phase 1 straight-through value

$$\widetilde{\Xi}^{\mathrm{ST}} = \widehat{\Xi} + \mathrm{sg}(\widetilde{\Xi} - \widehat{\Xi}) \tag{30}$$

has the exact hard-cell forward value and identity derivative in $\widehat{\Xi}$; fitted interface parameters and boundaries are held constant within each update block. The shared reconstruction map is

$$\mathrm{Rec}_{\theta,\mathcal{I}}(c, \varepsilon) = D_\theta\Big[\widehat{T}^{-1}\Big(\widehat{F}^{-1}(b_c + \varepsilon \odot \Delta_c)\Big)\Big]. \tag{31}$$

Inactive entries are restored to $\widehat{F}_j^{-1}(1/2)$ before applying the inverse gauge. Quantized training, validation qrec, and generation use this same forward rule without soft assignments. Cell boundaries are recalibrated from training data rather than optimized directly by backpropagation.

B.6 NUMERICAL CONVENTIONS

The calibrated interface is stored and reused with the numerical conventions below. Gauge-bank size N_G and response-context budget M_C are dataset-specific; their values are given in the accompanying configurations (Appendix F.3).

Quantity	Fixed value
Covariance ridges $(\epsilon_\Sigma, \epsilon_0)$	$(10^{-5}, 10^{-8})$
Relative near-tie tolerance	0.05
Sign-skewness tolerance	10^{-6}
CDF interpolation tolerance; range threshold τ_{range}	10^{-8}
Response threshold τ_{response}	10^{-12}

These numerical conventions are fixed independently of test data.

C PROOFS AND ADDITIONAL THEORETICAL RESULTS

The proofs address canonical equivalence, fixed-representation allocation, and generation error. High-rate statements hold inactive entries fixed in both decoder outputs. Generation control uses the selected reconstruction kernel and population losses, without a high-rate assumption. Within each proof, subscripts identifying a fixed representation are suppressed when unambiguous.

C.1 LOCAL GEOMETRY AND CANONICAL EQUIVALENCE

Lemma C.1 (Pullback expansion)**.** *If D_θ is C^2 near z and its nearby images lie in a normal neighborhood of $x = D_\theta(z)$, then, as $\delta z \to 0$,*

$$d_g^2(D_\theta(z+\delta z), D_\theta(z)) = \delta z^\top G_\theta(z)\delta z + o(\|\delta z\|_2^2). \tag{32}$$

Proof. Use normal coordinates centered at x, and write

$$v(\delta z) := \log_x D_\theta(z+\delta z) = dD_{\theta,z}[\delta z] + r(\delta z), \qquad \|r(\delta z)\|_{g,x} = o(\|\delta z\|_2). \tag{33}$$

Within this neighborhood, $d_g(x, \exp_x v) = \|v\|_{g,x}$. Therefore

$$\begin{aligned} d_g^2(D_\theta(z+\delta z), x) &= \|dD_{\theta,z}[\delta z] + r(\delta z)\|_{g,x}^2 \\ &= g_x(dD_{\theta,z}[\delta z], dD_{\theta,z}[\delta z]) + o(\|\delta z\|_2^2) \\ &= \delta z^\top G_\theta(z)\delta z + o(\|\delta z\|_2^2). \end{aligned} \tag{34}$$

The remainder follows from $\|dD_{\theta,z}[\delta z]\|_{g,x} = O(\|\delta z\|_2)$ and Cauchy–Schwarz; the last equality is the pullback definition in Eq. (3). □

For the decoder $\mathcal{D}$ in probability coordinates, assume differentiability at the points u on the coordinate paths considered. This is a separate assumption: continuity of the marginal CDFs alone need not make their inverses, or $\mathcal{D}$, differentiable. With the tangent norm evaluated at $\mathcal{D}(u)$, the pointwise expansion is

$$\Gamma_j(u) := a_{\mathcal{M}}^{-2}\|\partial_j \mathcal{D}(u)\|_g^2, \qquad \overline{d}_g^2(\mathcal{D}(u+he_j), \mathcal{D}(u)) = \Gamma_j(u)h^2 + o(h^2). \tag{35}$$

Thus Eq. (7) is $\gamma_j(s) = \mathbb{E}[\Gamma_j(U) \mid U_j = s]$. The high-rate results further assume the uniform remainder control in Eq. (53); pointwise differentiability alone does not imply it.

Proposition C.2 (Population affine equivalence)**.** *Assume positive-definite unregularized population covariance and finite mean pullback. As in Proposition 1, let*

$$Z' = MZ + b, \qquad M \in \mathrm{GL}(m_0), \qquad D_{\theta'}(z') = D_\theta(M^{-1}(z'-b)). \tag{36}$$

The generalized eigenvalues are unchanged. For every exact eigenvalue block $\mathcal{B}$, the normalized frames satisfy

$$W'_{\mathcal{B}} = MW_{\mathcal{B}}R_{\mathcal{B}}, \qquad T'(Z')_{\mathcal{B}} = R_{\mathcal{B}}^\top T(Z)_{\mathcal{B}}, \qquad R_{\mathcal{B}}^\top R_{\mathcal{B}} = I. \tag{37}$$

For simple eigenvalues each block rotation is a sign, yielding Proposition 1.

Proof. The mean and covariance transform as

$$\mu' = M\mu + b, \qquad \Sigma' = M\Sigma M^\top, \qquad (\Sigma')^{-1} = M^{-\top}\Sigma^{-1}M^{-1}. \tag{38}$$

The chain rule gives $dD_{\theta',z'} = dD_{\theta,z} \circ M^{-1}$, hence

$$\begin{aligned} v^\top G_{\theta'}(z')w &= g_{D_\theta(z)}(dD_{\theta,z}[M^{-1}v], dD_{\theta,z}[M^{-1}w]) \\ &= v^\top M^{-\top} G_\theta(z) M^{-1} w, \\ G_{\theta'}(z') &= M^{-\top} G_\theta(z) M^{-1}, \qquad \overline{G}' = M^{-\top}\overline{G}M^{-1}. \end{aligned} \tag{39}$$

For a generalized eigenvector w_j in Eq. (4),

$$\begin{aligned} \overline{G}'(Mw_j) &= M^{-\top}\overline{G}w_j = \eta_j M^{-\top}\Sigma^{-1}w_j \\ &= \eta_j (M\Sigma M^\top)^{-1}(Mw_j), \end{aligned} \tag{40}$$

$$(Mw_a)^\top (\Sigma')^{-1}(Mw_b) = w_a^\top \Sigma^{-1} w_b = \delta_{ab}. \tag{41}$$

Applying the same argument to M^{-1} shows equality of the spectra, including multiplicities. The columns of $MW_\mathcal{B}$ and $W'_\mathcal{B}$ are orthonormal bases of the same transformed eigenspace in the $(\Sigma')^{-1}$ inner product; thus $W'_\mathcal{B} = MW_\mathcal{B}R_\mathcal{B}$ for an orthogonal $R_\mathcal{B}$. Substitution into the canonical map yields

$$\begin{aligned} T'(Z')_\mathcal{B} &= (MW_\mathcal{B}R_\mathcal{B})^\top (M\Sigma M^\top)^{-1} M(Z-\mu) \\ &= R_\mathcal{B}^\top W_\mathcal{B}^\top \Sigma^{-1}(Z-\mu) = R_\mathcal{B}^\top T(Z)_\mathcal{B}. \end{aligned} \tag{42}$$

A one-dimensional orthogonal block is $+1$ or -1, so in the simple-spectrum case $T'(Z') = S\,T(Z)$ with $S = \mathrm{diag}(s_1, \ldots, s_{m_0})$, $s_j \in \{-1, 1\}$. □

Orientation and finite cells. A nonzero finite third moment fixes a simple coordinate's sign by positive skewness. Otherwise the sign ambiguity remains. For $\Xi'_j = -\Xi_j$ and continuous marginals,

$$\begin{aligned} &F'_j(t) = 1 - F_j(-t), && U'_j = 1 - U_j \quad \text{a.s.}, \\ &\gamma'_j(s) = \gamma_j(1-s), && A'_j(s) = 1 - A_j(1-s), && b'_{j,k} = 1 - b_{j,K_0-k}. \end{aligned} \tag{43}$$

Here primes denote the reparameterized objects, not derivatives; the response identity holds for Lebesgue-almost every s, since conditional expectations are defined up to null sets. Consistent relabeling $k \mapsto K_0 - 1 - k$ preserves the output law apart from null boundaries. General tied-block rotations need not induce token permutations; finite regularization and alignment follow Appendix B.2, not an exact affine-invariance claim for fitted models.

C.2 Probability coordinates and response estimation

Lemma C.3 (Probability-integral transform)**.** *If F_j is the continuous population CDF of Ξ_j, then $U_j = F_j(\Xi_j) \sim \mathrm{Unif}(0,1)$. This imposes no independence among coordinates.*

Proof. For $0 < u < 1$, continuity gives $F_j(F_j^{-1}(u)) = u$. Flat intervals have zero probability, so the events below agree up to null sets:

$$\Pr(F_j(\Xi_j) \le u) = \Pr(\Xi_j \le F_j^{-1}(u)) = F_j(F_j^{-1}(u)) = u. \tag{44}$$

The endpoint probabilities follow by limits. Each transform acts marginally, without altering the copula. □

The interpolated empirical map in Appendix B.3 is an estimator; exact uniformity is a population statement.

For fixed smooth $\mathcal{D}$, fixed $a_\mathcal{M} > 0$, and interior s, let $I_h = [s - h/2, s + h/2] \subset (0,1)$ have positive probability. Define

$$q_h(v) = \frac{\overline{d}_g^2(\mathcal{D}(v \oplus_j (s - h/4)), \mathcal{D}(v \oplus_j (s + h/4)))}{(h/2)^2}, \qquad \Gamma_s(v) = \Gamma_j(v \oplus_j s). \tag{45}$$

The target is $\gamma_j(s) = \mathbb{E}[\Gamma_s(U_{-j}) \mid U_j = s]$ for a fixed version of the conditional law.

Proposition C.4 (Fixed-interface secant consistency). *Assume $q_h \to \Gamma_s$ uniformly on the supported contexts, $t \mapsto \mathbb{E}[\Gamma_s(U_{-j}) \mid U_j = t]$ is continuous at s, and $\sup_{0<h<h_0} \mathbb{E}[q_h(U_{-j})^2 \mid U_j \in I_h] \le C_s < \infty$. For independent contexts $V_1, \dots, V_n \sim \mathcal{L}(U_{-j} \mid U_j \in I_h)$,*

$$\widehat{\gamma}^{\mathcal{D}}_{j,h,n}(s) = \frac{1}{n}\sum_{i=1}^{n} q_h(V_i) \xrightarrow{\Pr} \gamma_j(s) \qquad (h \to 0,\ n \to \infty). \tag{46}$$

Proof. Along $c_v(t) = \mathcal{D}(v \oplus_j t)$, the local pullback expansion identifies

$$q_h(v) \longrightarrow a_{\mathcal{M}}^{-2} \|\dot{c}_v(s)\|_g^2 = \Gamma_s(v). \tag{47}$$

Write $m_h = \mathbb{E}[q_h(U_{-j}) \mid U_j \in I_h]$ and $g_s(t) = \mathbb{E}[\Gamma_s(U_{-j}) \mid U_j = t]$. The bias satisfies

$$\begin{aligned} |m_h - \gamma_j(s)| &\le \mathbb{E}[|q_h(U_{-j}) - \Gamma_s(U_{-j})| \mid U_j \in I_h] \\ &\quad + |\mathbb{E}[g_s(U_j) \mid U_j \in I_h] - g_s(s)| \\ &\le \sup_v |q_h(v) - \Gamma_s(v)| + \sup_{t \in I_h} |g_s(t) - g_s(s)| \longrightarrow 0. \end{aligned} \tag{48}$$

Independence and the conditional second-moment bound give

$$\mathrm{Var}(\widehat{\gamma}^{\mathcal{D}}_{j,h,n}) = \frac{\mathrm{Var}(q_h(V_1))}{n} \le \frac{\mathbb{E}[q_h(V_1)^2]}{n} \le \frac{C_s}{n}. \tag{49}$$

For any $\delta > 0$, choose h small enough that $|m_h - \gamma_j(s)| \le \delta/2$; then Chebyshev gives

$$\Pr\big(|\widehat{\gamma}^{\mathcal{D}}_{j,h,n}(s) - \gamma_j(s)| > \delta\big) \le \Pr\big(|\widehat{\gamma}^{\mathcal{D}}_{j,h,n}(s) - m_h| > \delta/2\big) \le \frac{4C_s}{n\delta^2} \longrightarrow 0. \tag{50}$$

□

Uniform local convergence is ensured, for example, by C^2 regularity on a compact path region with uniform normal-neighborhood control. Consistency requires a diverging retained context count and, for the floored estimator, a vanishing floor coefficient. It does not follow for the fixed-cap, reused-training-bank implementation merely by freezing its learned maps; its empirical behavior is assessed in Appendix H.2.

C.3 Scalar high-rate allocation

Lemma C.5 (Within-cell second moment). *If V and $\widetilde{V}$ are independent and uniform on an interval B of width Δ and center c_B, then*

$$\mathbb{E}[(\widetilde{V} - V)^2] = \frac{\Delta^2}{6}. \tag{51}$$

Proof. Independence gives $\mathbb{E}[(\widetilde{V} - V)^2] = \mathrm{Var}(\widetilde{V}) + \mathrm{Var}(V) = \Delta^2/6$. For fixed $t \in B$,

$$w_B(t) := \mathbb{E}[(\widetilde{V} - t)^2] = \frac{\Delta^2}{12} + (t - c_B)^2, \qquad \int_B w_B(t)\, dt = \frac{\Delta^3}{6}. \tag{52}$$

□

For the coordinate paths used below, assume the uniform remainder condition

$$\left|\overline{d}_g^2(\mathcal{D}(u + he_j), \mathcal{D}(u)) - \Gamma_j(u)h^2\right| \le \omega_j(\delta)h^2 \quad (|h| \le \delta), \qquad \omega_j(\delta) \longrightarrow 0. \tag{53}$$

A C^2 decoder near a compact path region with uniform normal-neighborhood control is one sufficient case.

Theorem C.6 (Scalar Riemannian high-rate companding). *Fix $\mathcal{D}$ and the law of U. Suppose $U_j \sim \mathrm{Unif}(0,1)$, $\gamma_j(s) = \mathbb{E}[\Gamma_j(U) \mid U_j = s]$ has a continuous version on $[0,1]$, and λ_j is continuous with $\int_0^1 \lambda_j = 1$ and $\lambda_j \ge \lambda_{\min} > 0$. Assume Eq. (53). Partition by $A_j(u) = \int_0^u \lambda_j(s)\, ds$*

and draw $\widetilde{U}_j$ uniformly in the assigned cell, independently of the full original vector U conditional on that cell. Then

$$\begin{aligned}\mathcal{E}_{j,K_0}(\lambda_j) &:= \mathbb{E}\,\overline{d}_g^2(\mathcal{D}(U_{-j} \oplus_j \widetilde{U}_j), \mathcal{D}(U)) \\ &= \frac{1}{6K_0^2}\int_0^1 \frac{\gamma_j(s)}{\lambda_j(s)^2}\,ds + o(K_0^{-2}).\end{aligned} \tag{54}$$

For any nonnegative response with $0 < \int_0^1 \gamma_j^{1/3} < \infty$, the leading functional over nonnegative unit-integral densities satisfies

$$\inf_{\lambda_j} \int_0^1 \frac{\gamma_j(s)}{\lambda_j(s)^2}\,ds = \left(\int_0^1 \gamma_j(s)^{1/3}\,ds\right)^3, \tag{55}$$

$$\lambda_j^{\mathrm{opt}}(s) = \frac{\gamma_j(s)^{1/3}}{\int_0^1 \gamma_j(v)^{1/3}\,dv}. \tag{56}$$

The functional contributes zero where $\gamma_j = \lambda_j = 0$ and is infinite where $\gamma_j > 0$ but $\lambda_j = 0$.

Proof. Cell contribution. Write $B_k = [b_{j,k}, b_{j,k+1})$ and $\Delta_k = |B_k|$. Since $\int_{B_k} \lambda_j = 1/K_0$, continuity gives a point $s_k \in B_k$ (or its closure) with

$$\lambda_j(s_k)\Delta_k = \frac{1}{K_0}, \qquad \Delta_{\max} := \max_k \Delta_k \le \frac{1}{K_0\lambda_{\min}}. \tag{57}$$

Condition on $U_j = t \in B_k$. The new draw is independent of U_{-j} given t, so Eq. (53) and the conditional moment in Eq. (52) imply

$$\begin{aligned}&\mathbb{E}\Big[\overline{d}_g^2(\mathcal{D}(U_{-j} \oplus_j \widetilde{U}_j), \mathcal{D}(U))\mathbf{1}_{\{U_j \in B_k\}}\Big] \\ &\qquad = \int_{B_k} \gamma_j(t) w_{B_k}(t)\,dt + e_k, \qquad |e_k| \le \omega_j(\Delta_{\max})\frac{\Delta_k^3}{6}.\end{aligned} \tag{58}$$

Let $\kappa_j(\delta) = \sup_{|s-t|\le\delta} |\gamma_j(s) - \gamma_j(t)| \to 0$. Then

$$\left|\int_{B_k} \gamma_j(t) w_{B_k}(t)\,dt - \frac{\gamma_j(s_k)\Delta_k^3}{6}\right| \le \kappa_j(\Delta_{\max})\frac{\Delta_k^3}{6}. \tag{59}$$

Dependence among coordinates is contained in $\gamma_j(t)$; no product law for U is used.

Summed remainder and limit. Since $\sum_k \Delta_k = 1$,

$$\sum_k \Delta_k^3 \le \Delta_{\max}^2 \le (K_0\lambda_{\min})^{-2}. \tag{60}$$

The total error from Eqs. (58)–(59) is therefore bounded by

$$\frac{\omega_j(\Delta_{\max}) + \kappa_j(\Delta_{\max})}{6K_0^2\lambda_{\min}^2} = o(K_0^{-2}). \tag{61}$$

Using Eq. (57) and the Riemann sum for the continuous function γ_j/λ_j^2,

$$\begin{aligned}\mathcal{E}_{j,K_0}(\lambda_j) &= \frac{1}{6}\sum_k \gamma_j(s_k)\Delta_k^3 + o(K_0^{-2}) \\ &= \frac{1}{6K_0^2}\sum_k \frac{\gamma_j(s_k)}{\lambda_j(s_k)^2}\Delta_k + o(K_0^{-2}) \\ &= \frac{1}{6K_0^2}\int_0^1 \frac{\gamma_j(s)}{\lambda_j(s)^2}\,ds + o(K_0^{-2}).\end{aligned} \tag{62}$$

Optimal density. For any unit-integral density with finite functional, Hölder's inequality gives

$$\begin{aligned}\int_0^1 \gamma_j(s)^{1/3}\,ds &= \int_0^1 \left(\frac{\gamma_j(s)}{\lambda_j(s)^2}\right)^{1/3} \lambda_j(s)^{2/3}\,ds \\ &\le \left(\int_0^1 \frac{\gamma_j(s)}{\lambda_j(s)^2}\,ds\right)^{1/3} \left(\int_0^1 \lambda_j(s)\,ds\right)^{2/3}.\end{aligned} \tag{63}$$

The second factor is one. Cubing gives Eq. (55). Equality requires γ_j/λ_j^2 proportional to λ_j, so $\lambda_j^3 \propto \gamma_j$; normalization gives Eq. (56) and attains the displayed functional. □

Uniform allocation and zero response. Jensen yields

$$\left(\int_0^1 \gamma_j(s)^{1/3}\,ds\right)^3 \le \int_0^1 \gamma_j(s)\,ds, \tag{64}$$

with equality only for an almost-everywhere constant response. This compares leading coefficients, not arbitrary finite-K_0 errors. If continuous γ_j vanishes, the optimal density may lack a positive lower bound. For fixed $\varepsilon > 0$, set

$$\lambda_{j,\varepsilon}(s) = \frac{(\gamma_j(s)+\varepsilon)^{1/3}}{Z_\varepsilon}, \qquad Z_\varepsilon = \int_0^1 (\gamma_j(s)+\varepsilon)^{1/3}\,ds. \tag{65}$$

Then

$$\int_0^1 \frac{\gamma_j(s)}{\lambda_{j,\varepsilon}(s)^2}\,ds = Z_\varepsilon^2 \int_0^1 \frac{\gamma_j(s)}{(\gamma_j(s)+\varepsilon)^{2/3}}\,ds \longrightarrow \left(\int_0^1 \gamma_j(s)^{1/3}\,ds\right)^3 \tag{66}$$

by dominated convergence. The high-rate expansion holds for each fixed positive ε under the stated assumptions, not automatically in a simultaneous ε, K_0 limit. An identically zero response has zero functional and needs no normalized cube-root formula.

C.4 Joint reconstruction control

Let

$$U^{[j]} = (\widetilde{U}_1, \dots, \widetilde{U}_j, U_{j+1}, \dots, U_{m_0}), \quad U^{[0]} = U, \quad U^{[m_0]} = \widetilde{U}, \tag{67}$$

with within-cell noises independent of one another and of U. Define the normalized hybrid increments and responses by

$$\delta_j = \overline{d}_g(\mathcal{D}(U^{[j-1]}), \mathcal{D}(U^{[j]})), \tag{68}$$

$$\gamma_{j,K_0}^{\mathrm{hyb}}(s) = \mathbb{E}[\Gamma_j(U^{[j-1]}) \mid U_j = s]. \tag{69}$$

Theorem C.7 (Joint high-rate upper bound)**.** *For fixed m_0 and $\mathcal{D}$, assume the scalar conditions of Theorem C.6 for every active coordinate, Eq.* (53) *uniformly along the hybrid paths, and*

$$\epsilon_{j,K_0} := \|\gamma_{j,K_0}^{\mathrm{hyb}} - \gamma_j\|_{L^\infty(0,1)} = o(1). \tag{70}$$

Then

$$\mathbb{E}\,\overline{d}_g^2(\mathcal{D}(U), \mathcal{D}(\widetilde{U})) \le \frac{m_0}{6K_0^2} \sum_{j=1}^{m_0} \int_0^1 \frac{\gamma_j(s)}{\lambda_j(s)^2}\,ds + o(K_0^{-2}). \tag{71}$$

The infimum of the leading term is

$$\frac{m_0}{6K_0^2} \sum_{j=1}^{m_0} \left(\int_0^1 \gamma_j(s)^{1/3}\,ds\right)^3, \tag{72}$$

attained by cube-root densities when admissible, or approached by positive allocations as above.

Proof. The triangle inequality along $U^{[0]}, \dots, U^{[m_0]}$ and Cauchy–Schwarz give

$$\overline{d}_g^2(\mathcal{D}(U), \mathcal{D}(\widetilde{U})) \le \left(\sum_{j=1}^{m_0} \delta_j\right)^2 \le m_0 \sum_{j=1}^{m_0} \delta_j^2. \tag{73}$$

For coordinate j, repeat the cell calculation with the earlier coordinates already reconstructed:

$$\mathbb{E}\delta_j^2 = \sum_k \int_{B_{j,k}} \gamma_{j,K_0}^{\mathrm{hyb}}(t) w_{B_{j,k}}(t)\,dt + o(K_0^{-2}). \tag{74}$$

The same uniform remainder and cell-width bounds control the last term. Replacing the hybrid response by the original one changes this expression by at most

$$\left|\sum_k \int_{B_{j,k}} (\gamma_{j,K_0}^{\mathrm{hyb}}(t) - \gamma_j(t)) w_{B_{j,k}}(t)\,dt\right|$$

$$\leq \frac{\epsilon_{j,K_0}}{6} \sum_k \Delta_{j,k}^3 \leq \frac{\epsilon_{j,K_0}}{6K_0^2 \lambda_{\min,j}^2} = o(K_0^{-2}), \tag{75}$$

where $\lambda_{\min,j} > 0$ is coordinate j's density lower bound. Applying the scalar Riemann-sum calculation gives

$$\mathbb{E}\delta_j^2 = \frac{1}{6K_0^2} \int_0^1 \frac{\gamma_j(s)}{\lambda_j(s)^2}\, ds + o(K_0^{-2}). \tag{76}$$

Sum over the fixed number of coordinates in Eq. (73), then minimize each integral using Eq. (55). This proves the bound in Theorem 1. □

The factor m_0 comes from Cauchy–Schwarz, not vanishing coordinate interactions. Stability outside the original context support is an additional assumption.

Proposition C.8 (Continuous-to-finite reconstruction)**.** *For any observation X and exact-cell reconstruction $\widetilde{U}$ with finite displayed moments,*

$$\mathbb{E}d_g^2(X, \mathcal{D}(\widetilde{U})) \leq 2\mathbb{E}d_g^2(X, \mathcal{D}(U)) + 2\mathbb{E}d_g^2(\mathcal{D}(U), \mathcal{D}(\widetilde{U})). \tag{77}$$

Proof. Apply the triangle inequality and $(a+b)^2 \leq 2a^2 + 2b^2$, then take expectations. Division by $a_{\mathcal{M}}^2$ gives the normalized form. □

The same inequality applies to the fitted reconstruction map, including its clamping and inactive-coordinate conventions; its normalized left-hand side is the Phase 1 qrec loss. Phase 1 optimizes this loss directly together with clean reconstruction.

Marginal preservation.

Proposition C.9 (Coordinatewise marginal preservation)**.** *Suppose $U_j \sim \mathrm{Unif}(0,1)$, $C_j = k$ when $U_j \in B_{j,k} = [b_{j,k}, b_{j,k+1})$, and the token marginal is exact. If $\widetilde{U}_j \mid C_j = k \sim \mathrm{Unif}(B_{j,k})$, then $\widetilde{U}_j \sim \mathrm{Unif}(0,1)$.*

Proof. Uniformity of U_j gives $\Pr(C_j = k) = \Delta_{j,k}$. For any measurable $A \subseteq [0,1]$, with $|\cdot|$ denoting Lebesgue measure,

$$\begin{aligned}\Pr(\widetilde{U}_j \in A) &= \sum_{k=0}^{K_0-1} \Pr(C_j = k) \Pr(\widetilde{U}_j \in A \mid C_j = k) \\ &= \sum_{k=0}^{K_0-1} \Delta_{j,k} \frac{|A \cap B_{j,k}|}{\Delta_{j,k}} = \sum_{k=0}^{K_0-1} |A \cap B_{j,k}| = |A|. \end{aligned} \tag{78}$$

□

Marginals are preserved, not the conditional copula inside each joint cell; qrec measures the decoded reconstruction error and the prior models discrete dependence.

C.5 Code-law learning and generation control

Freeze the selected encoder, decoder, and interface. All losses below are population expectations over fresh observations conditional on these maps. Let

$$H(\Pi^*) = -\mathbb{E}_{\Pi^*} \log \Pi^*(C), \qquad \mathcal{L}_{\mathrm{P2}}^{\mathrm{oracle}} = \frac{H(\Pi^*)}{m_0 \log K_0}. \tag{79}$$

The oracle uses unrestricted $P^C = \Pi^*$, not an assumed optimum within the finite prior class.

Proposition C.10 (Cross-entropy and joint code KL)**.** *Suppose $P_\psi^C(c) > 0$ whenever $\Pi^*(c) > 0$. Then*

$$\mathcal{L}_{\mathrm{P2}}(\psi) = \frac{H(\Pi^*) + \mathrm{KL}(\Pi^* \| P_\psi^C)}{m_0 \log K_0}, \tag{80}$$

$$\mathcal{L}_{\mathrm{P2}}(\psi) - \mathcal{L}_{\mathrm{P2}}^{\mathrm{oracle}} = \frac{\mathrm{KL}(\Pi^* \| P_\psi^C)}{m_0 \log K_0}, \tag{81}$$

$$\mathrm{KL}(\Pi^* \| P_\psi^C) = \sum_{j=1}^{m_0} \mathbb{E}_{\Pi^*} \mathrm{KL}\big(\Pi_j^*(\cdot \mid C_{<j}) \| p_{\psi,j}(\cdot \mid C_{<j})\big)\,. \tag{82}$$

Proof. Add and subtract $\log \Pi^*(C)$ in the cross-entropy:

$$\begin{aligned} -\mathbb{E}_{\Pi^*} \log P_\psi^C(C) &= -\mathbb{E}_{\Pi^*} \log \Pi^*(C) + \mathbb{E}_{\Pi^*} \log \frac{\Pi^*(C)}{P_\psi^C(C)} \\ &= H(\Pi^*) + \mathrm{KL}(\Pi^* \| P_\psi^C). \end{aligned} \tag{83}$$

Divide by $m_0 \log K_0$ and subtract Eq. (79). To obtain the final identity, expand both joint distributions using the same coordinate order:

$$\begin{aligned} \mathrm{KL}(\Pi^* \| P_\psi^C) &= \mathbb{E}_{\Pi^*} \log \frac{\prod_j \Pi_j^*(C_j \mid C_{<j})}{\prod_j p_{\psi,j}(C_j \mid C_{<j})} \\ &= \sum_{j=1}^{m_0} \mathbb{E}_{C_{<j} \sim \Pi^*} \sum_k \Pi_j^*(k \mid C_{<j}) \log \frac{\Pi_j^*(k \mid C_{<j})}{p_{\psi,j}(k \mid C_{<j})} \\ &= \sum_{j=1}^{m_0} \mathbb{E}_{\Pi^*} \mathrm{KL}\big(\Pi_j^*(\cdot \mid C_{<j}) \| p_{\psi,j}(\cdot \mid C_{<j})\big)\,. \end{aligned} \tag{84}$$

Inactive factors are deterministic and contribute zero, while the normalization remains $m_0 \log K_0$ as in Eq. (13). □

Lemma C.11 (Cell-width bias at the logit level)**.** *If $\Delta_{j,k} > 0$ for every active symbol, then unrestricted residual logits in*

$$p_k = \frac{\Delta_{j,k} e^{r_k}}{\sum_\ell \Delta_{j,\ell} e^{r_\ell}} \tag{85}$$

can represent every strictly positive categorical distribution.

Proof. For a strictly positive target $(p_k)_k$, choose $r_k = \log p_k - \log \Delta_{j,k}$. The numerator in Eq. (85) is then p_k and the denominator is $\sum_\ell p_\ell = 1$. □

Representing the required context-dependent residuals with a finite neural prior is a separate approximation question.

Shared reconstruction kernel. Let $K^*(c, \cdot)$ be the frozen within-cell reconstruction kernel. The true-code reconstruction and generated laws are

$$\rho_{\mathrm{rec}}^* = \Pi^* K^* = (\mathrm{Rec}_{\theta^*, \mathcal{I}^*})_\sharp \left(\Pi^* \otimes \mathrm{Unif}([0,1)^{m_0})\right), \tag{86}$$

$$\rho_{\theta^*,\psi} = P_\psi^C K^*. \tag{87}$$

Lemma C.12 (True-code reconstruction coupling)**.** *For the selected model with finite population qrec and finite first moments,*

$$W_1^g(\rho^*, \rho_{\mathrm{rec}}^*) \le a_{\mathcal{M}} \sqrt{\mathcal{L}_{\mathrm{qrec}}}. \tag{88}$$

Proof. Draw $X \sim \rho^*$ and independently $\varepsilon \sim \mathrm{Unif}([0,1)^{m_0})$, and set

$$Y = \mathrm{Rec}_{\theta^*, \mathcal{I}^*}(Q^*(X), \varepsilon). \tag{89}$$

Because $Q^*(X) \sim \Pi^*$, the law of Y is ρ_{rec}^*. This coupling is admissible for Wasserstein distance, and

$$\begin{aligned} W_1^g(\rho^*, \rho_{\mathrm{rec}}^*) &\le \mathbb{E} d_g(X, Y) \le \sqrt{\mathbb{E} d_g^2(X, Y)} \\ &= a_{\mathcal{M}} \sqrt{\mathbb{E}\, \bar{d}_g^2(X, Y)} = a_{\mathcal{M}} \sqrt{\mathcal{L}_{\mathrm{qrec}}}. \end{aligned} \tag{90}$$

□

Lemma C.13 (Shared-kernel contraction)**.** *For two code laws P, P' and a common Markov kernel K,*

$$\mathrm{TV}(PK, P'K) \le \mathrm{TV}(P, P'). \tag{91}$$

If both output laws are supported in a metric space of diameter D, then

$$W_1(PK, P'K) \le D\ \mathrm{TV}(P, P'). \tag{92}$$

Proof. For any measurable event A, $K(\cdot, A)$ takes values in $[0, 1]$. By the variational characterization of total variation,

$$\begin{aligned} |PK(A) - P'K(A)| &= \left| \sum_c K(c, A)(P(c) - P'(c)) \right| \\ &\le \sup_{0 \le f \le 1} \left| \sum_c f(c)(P(c) - P'(c)) \right| = \mathrm{TV}(P, P'). \end{aligned} \tag{93}$$

Taking the supremum over A proves contraction. A maximal coupling (Y, Y') of the output laws satisfies $\Pr(Y \ne Y') = \mathrm{TV}(PK, P'K)$, so

$$W_1(PK, P'K) \le \mathbb{E}d(Y, Y') \le D \Pr(Y \ne Y') \le D\ \mathrm{TV}(P, P'). \tag{94}$$

□

Theorem C.14 (End-to-end generation bound)**.** *If $(\mathcal{M}, g)$ is compact and $P_\psi^C(c) > 0$ whenever $\Pi^*(c) > 0$, then*

$$W_1^g(\rho^*, \rho_{\theta^*,\psi}) \le a_{\mathcal{M}} \sqrt{\mathcal{L}_{\mathrm{qrec}}} + \mathrm{diam}_g(\mathcal{M}) \sqrt{\tfrac{1}{2} \mathrm{KL}(\Pi^* \| P_\psi^C)} \tag{95}$$

$$= a_{\mathcal{M}} \sqrt{\mathcal{L}_{\mathrm{qrec}}} + \mathrm{diam}_g(\mathcal{M}) \sqrt{\frac{m_0 \log K_0}{2} (\mathcal{L}_{\mathrm{P2}}(\psi) - \mathcal{L}_{\mathrm{P2}}^{\mathrm{oracle}})}. \tag{96}$$

Proof. Insert the true-code reconstruction law into the Wasserstein triangle inequality:

$$\begin{aligned} W_1^g(\rho^*, \rho_{\theta^*,\psi}) &\le W_1^g(\rho^*, \rho_{\mathrm{rec}}^*) + W_1^g(\Pi^* K^*, P_\psi^C K^*) \\ &\le a_{\mathcal{M}} \sqrt{\mathcal{L}_{\mathrm{qrec}}} + \mathrm{diam}_g(\mathcal{M})\, \mathrm{TV}(\Pi^*, P_\psi^C). \end{aligned} \tag{97}$$

The second line uses Lemmas C.12 and C.13. With natural logarithms, Pinsker's inequality (Cover & Thomas, 2006) gives

$$\mathrm{TV}(\Pi^*, P_\psi^C) \le \sqrt{\tfrac{1}{2} \mathrm{KL}(\Pi^* \| P_\psi^C)}. \tag{98}$$

Substitution proves Eq. (95), identical to Theorem 2; Eq. (81) yields the normalized-NLL form. □

These are population risks conditional on the selected maps; replacing them by finite-sample estimates requires separate generalization control.

C.6 Noncompact extension with bounded reconstruction range

Define the closed reconstruction range and its intrinsic diameter by

$$\mathcal{S}_{\mathcal{I}^*} = \overline{\bigcup_{c \in \mathcal{C}} \mathrm{supp}\, K^*(c, \cdot)}, \qquad D_{\mathcal{I}^*} = \mathrm{diam}_g(\mathcal{S}_{\mathcal{I}^*}). \tag{99}$$

Corollary C.15 (Generation bound on a bounded reconstruction range)**.** *For connected, possibly noncompact $(\mathcal{M}, g)$, assume $\mathcal{L}_{\mathrm{qrec}} < \infty$, $D_{\mathcal{I}^*} < \infty$, and $P_\psi^C(c) > 0$ whenever $\Pi^*(c) > 0$. Then*

$$W_1^g(\rho^*, \rho_{\theta^*,\psi}) \le a_{\mathcal{M}} \sqrt{\mathcal{L}_{\mathrm{qrec}}} + D_{\mathcal{I}^*} \sqrt{\tfrac{1}{2} \mathrm{KL}(\Pi^* \| P_\psi^C)} \tag{100}$$

$$= a_{\mathcal{M}} \sqrt{\mathcal{L}_{\mathrm{qrec}}} + D_{\mathcal{I}^*} \sqrt{\frac{m_0 \log K_0}{2} (\mathcal{L}_{\mathrm{P2}}(\psi) - \mathcal{L}_{\mathrm{P2}}^{\mathrm{oracle}})}. \tag{101}$$

Proof. First, fix $x_0 \in \mathcal{S}_{\mathcal{I}^*}$ and use the true-code coupling (X, Y) from Eq. (89). Then

$$\mathbb{E}d_g(X, x_0) \le \mathbb{E}d_g(X, Y) + \mathbb{E}d_g(Y, x_0) \le a_{\mathcal{M}}\sqrt{\mathcal{L}_{\mathrm{qrec}}} + D_{\mathcal{I}^*} < \infty. \tag{102}$$

The reconstruction and generated laws are supported in $\mathcal{S}_{\mathcal{I}^*}$, so they also have finite first moments.

Applying Lemma C.13 on $\mathcal{S}_{\mathcal{I}^*}$ gives

$$W_1^g(\Pi^* K^*, P_\psi^C K^*) \le D_{\mathcal{I}^*}\,\mathrm{TV}(\Pi^*, P_\psi^C). \tag{103}$$

The triangle inequality and Pinsker's inequality then yield Eq. (100) by the proof of Theorem C.14, replacing $\mathrm{diam}_g(\mathcal{M})$ with $D_{\mathcal{I}^*}$. Equation (81) gives the NLL form. □

A continuous extension of Rec to the finite code set times the closed noise cube has compact image and therefore finite intrinsic diameter. This sufficient condition is stronger than having a finite alphabet alone. For compact $\mathcal{M}$, $D_{\mathcal{I}^*} \le \mathrm{diam}_g(\mathcal{M})$.

D Algorithms and Training Details

Algorithms 1–3 give training and generation for the geometric benchmarks. Calibration uses training data only (Appendix B); validation selects checkpoints and hyperparameters (Appendix F.2).

D.1 Phase 1 training and selection

After W epochs of clean reconstruction, Phase 1 alternates H-epoch representation updates with interface calibration until epoch E. Here $E > W$, $(E - W) \bmod H = 0$, and $B = (E - W)/H$. Each update block holds the interface fixed and uses one within-cell draw per observation with the straight-through rule in Eq. (30). Full and Identity Compander optimize Eq. (12); Clean-Only omits qrec from training but uses the same validation criterion. For each block-end candidate t, recalibrate on its training latents and evaluate with common validation noise:

$$\mathcal{V}_{\mathrm{P1}}(t) = \mathcal{L}_{\mathrm{clean}}^{\mathrm{val}}(t) + \lambda_{\mathrm{qrec}}\mathcal{L}_{\mathrm{qrec}}^{\mathrm{val}}(t), \qquad t^* = \arg\min_t \mathcal{V}_{\mathrm{P1}}(t). \tag{104}$$

Exact ties select the earlier candidate. Validation uses the final target λ_{qrec} even when its training weight is ramped. Reusing each block-end interface gives $B + 1$ calibrations over the fixed epoch budget. A Cholesky failure during calibration terminates the run.

Algorithm 1 Phase 1 learning and validation selection

Require: train/validation splits, $E_\phi, D_\theta, K_0, \lambda_{\mathrm{qrec}}, W, H, E$
Check the epoch constraints; set $B \leftarrow (E - W)/H$ and candidate set $\mathcal{S} \leftarrow \varnothing$
Train on $\mathcal{L}_{\mathrm{clean}}$ for W epochs; calibrate $\mathcal{I}^{(0)}$ on training latents
for $t = 0, \ldots, B - 1$ **do**
    for each minibatch over H full epochs, including final partial batches **do**
        Draw one within-cell sample per observation
        Update (ϕ, θ) using the variant loss with $\mathcal{I}^{(t)}$ fixed
    Calibrate $\mathcal{I}^{(t+1)}$ on training latents, aligned to $\mathcal{I}^{(t)}$
    Add $(\phi, \theta, \mathcal{I}^{(t+1)})$ and its score from Eq. (104) to $\mathcal{S}$
    Reuse $\mathcal{I}^{(t+1)}$ in the next block
return the minimum-validation candidate in $\mathcal{S}$, with earlier-epoch tie-breaking

D.2 Phase 2 training and selection

The representation and interface remain frozen while a fresh recurrent prior models their hard codes. Previous-token and coordinate embeddings produce residual logits with the fixed $\log \Delta_{j,k}^*$ bias in Eq. (13); a zero-initialized output head starts from the cell-width law. Inactive factors are deterministic and excluded from the NLL sum, but not its normalization. Algorithm 2 uses the epoch budget E_2 in Table 9 and predeclared validation epochs $\mathcal{V}_2 \subseteq \{1, \ldots, E_2\}$, including E_2.

Algorithm 2 Phase 2 fitting of the frozen code law

Require: frozen $(\phi^*, \theta^*, \mathcal{I}^*)$; train/validation observations; E_2 and $\mathcal{V}_2$
1: Encode train/validation observations with Q^*; initialize a fresh zero-output-head prior
2: **for** $e = 1, \dots, E_2$ **do**
3: Train one epoch by teacher forcing with the normalized NLL in Eq. (13)
4: **if** $e \in \mathcal{V}_2$ **then**
5: Score the checkpoint by validation normalized NLL
6: **return** the lowest-scoring checkpoint, breaking exact ties by earlier epoch

D.3 Generation and computational cost

Generation uses the frozen interface and prior without recalibration, temperature changes, truncation, rejection, or trajectory integration. The prior's sequential updates precede one manifold-valued decoder evaluation.

Algorithm 3 One-decode generation

Require: frozen $(\psi^*, \theta^*, \mathcal{I}^*)$
1: **for** $j = 1, \dots, m_0$ **do**
2: Draw $C_j^{\mathrm{gen}} \sim p_{\psi^*,j}(\cdot \mid C_{<j}^{\mathrm{gen}})$ if active; otherwise set zero
3: Draw $\varepsilon \sim \mathrm{Unif}([0,1)^{m_0})$
4: For every coordinate j, set $U_j^{\mathrm{gen}} \leftarrow b^*_{j,C_j^{\mathrm{gen}}} + \varepsilon_j \Delta^*_{j,C_j^{\mathrm{gen}}}$ if active, else $1/2$
5: $\Xi^{\mathrm{gen}} \leftarrow (\widehat{F}^*)^{-1}(U^{\mathrm{gen}})$; $Z^{\mathrm{gen}} \leftarrow (\widehat{T}^*)^{-1}(\Xi^{\mathrm{gen}})$
6: **return** $D_{\theta^*}(Z^{\mathrm{gen}})$

After clean warm-up, a Full-model minibatch uses one encoder and two decoder passes. Calibration forms decoder pullbacks on N_G training points and uses at most $2m_0K_0M_C$ response-decoder outputs, alongside latent encoding, covariance estimation, CDF sorting, and an $O(m_0^3)$ eigendecomposition. Its cost depends on the decoder architecture and differentiation method. No $K_0^{m_0}$ joint grid is formed. For batch size n_b, batched prior logits occupy $O(n_b m_0 K_0)$ storage; streaming retains current logits, recurrent state, and generated codes. Per sample, categorical normalization/sampling costs $O(m_0K_0)$, with additional recurrent-network work; coordinatewise interpolation is linear apart from lookup, and the aligned inverse gauge $\widehat{\mu} + LV_{\mathrm{al}}\xi$ costs $O(m_0^2)$. End-to-end timings include this work (Appendix F.6).

E Manifold-Specific Implementations

Each backend specifies the encoder input, manifold-valued decoder, intrinsic distance, and pullback G_θ. The final support map is included before differentiating or measuring response. Training, calibration, and evaluation share the same intrinsic distance, with the normalization in Appendix B applied to reconstruction losses.

E.1 Sphere

Inputs on $\mathbb{S}^d$ are ambient unit vectors. For raw decoder output $a_\theta(z) \in \mathbb{R}^{d+1}$, a fixed threshold $\epsilon_S > 0$, and unit vector e_1,

$$D_\theta(z) = \begin{cases} a_\theta(z)/\|a_\theta(z)\|_2, & \|a_\theta(z)\|_2 > \epsilon_S, \\ e_1, & \text{otherwise.} \end{cases} \tag{105}$$

The map enforces the sphere constraint by construction. For unit x, y,

$$d_{\mathbb{S}^d}(x,y) = \mathrm{atan2}\big(\|x - (x^\top y)y\|_2, \mathrm{clip}(x^\top y, -1, 1)\big)\,. \tag{106}$$

On $\mathbb{S}^2$ the numerator is equivalently $\|x \times y\|_2$. The pullback differentiates the normalized decoder. The guard is constant strictly inside its branch; its switching boundary is not assumed differentiable.

E.2 Flat torus

For $\mathbb{T}^p$, inputs are $(\sin\theta_1, \ldots, \sin\theta_p, \cos\theta_1, \ldots, \cos\theta_p)$. Normalize each two-vector raw output $a_j(z)$, using $(1, 0)$ when its norm is at most a fixed ϵ_T, to obtain $\bar{a}_j$. Then

$$D_\theta(z)_j = \operatorname{atan2}(\bar{a}_{j,2}, \bar{a}_{j,1}), \qquad d^2_{\mathbb{T}^p}(\theta, \varphi) = \sum_{j=1}^{p} \operatorname{wrap}(\theta_j - \varphi_j)^2, \tag{107}$$

where $\operatorname{wrap}(t) = \operatorname{atan2}(\sin t, \cos t)$. The intrinsic pullback is $G_\theta = J_{\bar{a}}^\top J_{\bar{a}}$: tangent speed in the unit-circle embedding equals angular speed. Finite reconstruction and response errors use the wrapped product distance, not unwrapped or ambient straight-line differences.

E.3 Rotation group

The encoder receives the nine entries of $R \in \mathrm{SO}(3)$. The decoder outputs $\omega_\theta(z) \in \mathbb{R}^3$ and applies the exponential map, with $\Omega = [\omega_\theta(z)]_\times$ and $\vartheta = \|\omega_\theta(z)\|_2$:

$$D_\theta(z) = I + a(\vartheta)\Omega + b(\vartheta)\Omega^2, \qquad a(\vartheta) = \frac{\sin\vartheta}{\vartheta}, \quad b(\vartheta) = \frac{1 - \cos\vartheta}{\vartheta^2}. \tag{108}$$

The continuous values are $a(0) = 1$ and $b(0) = 1/2$; both coefficients are evaluated stably near zero. With $R_{\mathrm{rel}} = R_1^\top R_2$, set

$$\begin{aligned} c &= \operatorname{clip}((\operatorname{tr} R_{\mathrm{rel}} - 1)/2, -1, 1), \\ s &= \operatorname{clip}\left(\tfrac{1}{2}\| \operatorname{vee}(R_{\mathrm{rel}} - R_{\mathrm{rel}}^\top)\|_2, 0, 1\right), \qquad d_{\mathrm{SO}(3)}(R_1, R_2) = \operatorname{atan2}(s, c). \end{aligned} \tag{109}$$

Here vee inverts the cross-product matrix. The bi-invariant metric is $g_R(R[u]_\times, R[v]_\times) = u^\top v$, making distance the relative rotation angle. For differential dR, use $\nu = \operatorname{vee}(\operatorname{skew}(R^\top dR))$; the pullback is the Gram matrix of $dz \mapsto \nu$. Reported pairwise distances use float64.

E.4 Hyperbolic plane

At curvature -1, Poincaré coordinates $p \in \mathbb{R}^2$, $\|p\|_2 < 1$, map to the upper Lorentz hyperboloid as

$$x_0 = \frac{1 + \|p\|_2^2}{1 - \|p\|_2^2}, \qquad x_{1:} = \frac{2p}{1 - \|p\|_2^2}.$$

The encoder uses $p = x_{1:}/(x_0 + 1)$. For raw output $h_\theta(z) \in \mathbb{R}^2$,

$$D_\theta(z) = \left(\sqrt{1 + \|h_\theta(z)\|_2^2}, h_\theta(z)\right), \qquad G_\theta = J_D^\top \operatorname{diag}(-1, 1, 1) J_D.$$

For $\langle x, y\rangle_L = -x_0 y_0 + x_{1:}^\top y_{1:}$,

$$u = \max\{-\langle x, y\rangle_L - 1, 0\}, \qquad d_{\mathbb{H}^2}(x, y) = \log\left(1 + u + \sqrt{u(u+2)}\right).$$

This is the arcosh distance without a positive self-distance floor. The noncompact generation bound is in Appendix C.6.

F Experimental Details and Reproducibility

F.1 Datasets and preprocessing

Earth datasets follow Mathieu & Nickel (2020) and the underlying earthquake, volcano, flood, and fire sources (NOAA National Centers for Environmental Information, 2020a;b; Brakenridge, 2017; NASA Earth Science Data and Information System, 2020). Protein and RNA torsions use the processed benchmarks of Huang et al. (2022), based on Lovell et al. (2003) and Murray et al. (2003). Cone, Fisher24, and Line originate from Liu et al. (2023); we use the data distributed with RCM (Cheng et al., 2025). Following the public RMF implementation (Zhong et al., 2026), we merge their released training/test pools and split 40,000 samples 80/10/10. Swiss follows RCM's $\mathrm{SO}(3)$ Swiss-roll construction with noise standard deviation 1.0, using the same count and split. The

GFM Poincaré benchmark (Davis et al., 2026) draws 20,000 wrapped-normal samples with mean $(-0.6, 0.6)$ and standard deviation 0.7. All methods share the splits in Table 7; quality evaluation generates exactly the test count. Figure-specific sample counts are stated in Appendix I, rather than treated as evaluation counts.

Table 7: Fixed train/validation/test sample counts. All methods use the same split memberships.

Family	Dataset/manifold	Train	Val.	Test
Earth	Earthquake / $\mathbb{S}^2$	4,896	612	612
Earth	Fire / $\mathbb{S}^2$	10,249	1,280	1,280
Earth	Flood / $\mathbb{S}^2$	3,901	487	487
Earth	Volcano / $\mathbb{S}^2$	663	82	82
Protein	General / $\mathbb{T}^2$	110,568	13,820	13,820
Protein	Glycine / $\mathbb{T}^2$	10,627	1,328	1,328
Protein	Proline / $\mathbb{T}^2$	6,108	763	763
Protein	Pre-Pro / $\mathbb{T}^2$	5,528	691	691
RNA	RNA / $\mathbb{T}^7$	7,579	947	947
SO(3)	Cone / $\mathrm{SO}(3)$	32,000	4,000	4,000
SO(3)	Fisher24 / $\mathrm{SO}(3)$	32,000	4,000	4,000
SO(3)	Line / $\mathrm{SO}(3)$	32,000	4,000	4,000
SO(3)	Swiss / $\mathrm{SO}(3)$	32,000	4,000	4,000
Hyperbolic	H2 / $\mathbb{H}^2$	16,000	2,000	2,000

F.2 Validation-based selection and controls

Both checkpoint selection and hyperparameter selection use validation data; test data is reserved for final evaluation. Interface fitting uses training data only. On the geometric benchmarks, Phase 1 selects the minimum validation clean-plus-qrec score after train-only recalibration, and Phase 2 selects the minimum validation normalized NLL (Appendix D). The controls in Table 8 share non-variant architecture, prior, optimization, selection, and sampling settings.

Table 8: Controls isolate cell placement and training through the finite reconstruction channel.

Variant	Cell placement	Phase 1 training objective
Full	Response-aware	Clean $+\lambda_{\mathrm{qrec}}$ qrec
Identity Compander	Uniform in probability space	Same as Full
Clean-Only	Response-aware cells calibrated from each clean-trained candidate	Clean only; validation includes clean $+\lambda_{\mathrm{qrec}}$ qrec

F.3 Architecture and optimization settings

Table 9 lists dataset-specific latent/prior widths and training schedules. The unconditional prior uses shared recurrent weights with token and coordinate embeddings (Appendix D.2). Both phases use AdamW. For ramp length $R > 0$ and power p, block $b \geq 1$ uses training weight $\lambda_{\mathrm{qrec}} \min(1, b/R)^p$; $R = 0$ uses the target weight immediately after warm-up. Validation always uses the target weight.

For each geometric dataset, network and optimization settings are fixed across the five training seeds. The accompanying code configurations specify dataset-specific architectures, training settings, gauge-bank sizes, and response-context budgets.

F.4 Baseline implementations

Baselines use the released manifold representations and samplers listed in Table 10 and are trained and evaluated on the common splits.

Table 9: RiCoTA settings. h/e are prior hidden/embedding widths; $W/H/E$ specifies Phase 1 warm-up, refresh interval, and epoch budget. E_2 is the Phase 2 epoch budget. R/p denotes qrec ramp blocks/power, with 0 meaning no ramp. Budgets are not selected epochs.

Dataset	m_0	K_0	λ_{qrec}	h/e	$W/H/E$	E_2	R/p
Earthquake	3	64	3	128/48	20/15/110	1200	4/1
Fire	3	64	1	96/48	20/15/125	800	3/1
Flood	3	96	0.5	64/64	100/20/300	750	5/1
Volcano	3	36	1	48/48	130/90/1030	1400	0
$\mathrm{SO}(3)$ Cone	4	96	1	64/64	30/10/100	200	6/1
$\mathrm{SO}(3)$ Fisher24	4	64	2	96/64	50/10/140	200	6/1
$\mathrm{SO}(3)$ Line	6	96	1.25	128/64	40/12/160	300	4/1
$\mathrm{SO}(3)$ Swiss	4	72	0.5	256/128	30/10/120	100	9/0.75
Protein General	2	128	0.5	32/16	100/20/300	800	5/1
Protein Glycine	4	60	1.75	64/32	20/10/120	400	4/1
Protein Proline	2	64	0.5	8/8	80/20/200	800	2/1
Protein Pre-Pro	3	64	1.5	32/32	20/10/100	400	4/1
RNA	4	24	2	64/32	20/10/150	200	0
Hyperbolic $\mathbb{H}^2$	2	48	3	64/32	25/30/175	75	0

Table 10: Baseline sources and reported sampling evaluation counts. RFM is the iterative reference.

Method	Implementation source	Evaluations
RFM	Chen & Lipman (2024); $\mathrm{SO}(3)$: GFM (Davis et al., 2026)	1000
G-LSD	Davis et al. (2026)	1
RCM/RCT	Cheng et al. (2025)	2
RMF-MT	Zhong et al. (2026)	1
M-FFF	Sorrenson et al. (2024)	1

$\mathrm{SO}(3)$ settings. RCM/RCT and RMF-MT use released configurations for these benchmarks. RFM uses the GFM repository's RFM implementation and $\mathrm{SO}(3)$ configuration (Davis et al., 2026); G-LSD and M-FFF apply their released $\mathrm{SO}(3)$ settings to the common datasets and splits.

M-FFF uses a differentiable polar projection of unconstrained decoder outputs during training. Its released eigendecomposition-based implementation produced non-finite gradients on our $\mathrm{SO}(3)$ tasks. We stabilized only the inverse-square-root computation with Newton–Schulz iteration (Higham, 1997), retaining the projection and training objective. Quality evaluation and timing use the same implementation.

F.5 Evaluation metrics and statistical reporting

Let $X = (x_i)_{i=1}^N$ be the fixed test set and $Y = (y_j)_{j=1}^N$ the generated bank (generation seed 104729). Distances d_g use radians on angular manifolds and native units on $\mathbb{H}^2$.

MMD. Following the geodesic exponential-kernel convention in prior manifold-generation work (Davis et al., 2026; Cheng et al., 2025; Zhong et al., 2026), use coefficient one and intrinsic distance:

$$k(x,y) = e^{-d_g(x,y)^2}, \qquad k_{\mathbb{T}^p}(\theta,\varphi) = \exp\left[-\sum_{j=1}^{p} \mathrm{wrap}(\theta_j - \varphi_j)^2\right]. \tag{110}$$

The torus kernel uses the full angle vector. All methods include diagonal terms and report

$$B_k(X,Y) = \frac{1}{N^2}\sum_{i,j=1}^{N}[k(x_i,x_j) + k(y_i,y_j) - 2k(x_i,y_j)],$$
$$\mathrm{MMD}(X,Y) = \sqrt{\max\{B_k(X,Y),0\}}.$$

Geodesic Gaussian kernels are not generally positive semidefinite (Da Costa et al., 2023); we therefore interpret this score as an empirical benchmark discrepancy, without claiming an RKHS-distance interpretation or distribution-identification guarantees (Gretton et al., 2012).

Local support and normalized tail. We adapt generative precision–recall (Kynkäänniemi et al., 2019) using intrinsic neighborhoods. Let $a_j^g = \min_i d_g(y_j, x_i)$, $b_i^g = \min_j d_g(x_i, y_j)$, and $i(j)$ be the nearest test point to y_j. The scale $\widetilde{s}_i$ is the distance from x_i to its fifth-nearest validation point, floored at 10^{-8}. Define

$$\begin{aligned} P_{\mathrm{loc}}(\lambda) &= \frac{1}{N}\sum_j \mathbf{1}[a_j^g \le \lambda \widetilde{s}_{i(j)}], \quad R_{\mathrm{loc}}(\lambda) = \frac{1}{N}\sum_i \mathbf{1}[b_i^g \le \lambda \widetilde{s}_i], \\ F_{1,\mathrm{loc}}(\lambda) &= \frac{2P_{\mathrm{loc}}(\lambda)R_{\mathrm{loc}}(\lambda)}{P_{\mathrm{loc}}(\lambda)+R_{\mathrm{loc}}(\lambda)}, \qquad \mathrm{sF1} = \frac{1}{4}\sum_{\lambda\in\{0.5,1,2,4\}} F_{1,\mathrm{loc}}(\lambda). \end{aligned} \tag{111}$$

Use zero for a vanishing F1 denominator. Local F_1 tables use $\lambda = 0.5$; the normalized tail is $\mathrm{t2g90} = Q_{0.90}(\{b_i^g/\widetilde{s}_i\}_{i=1}^N)$.

Geodesic shape and support violations. Write $\delta_{\mathcal{M}} = c_{\mathcal{M}} d_g$, where $c_{\mathcal{M}} = 180/\pi$ on angular manifolds and 1 on $\mathbb{H}^2$, and set $a_j = c_{\mathcal{M}} a_j^g$, $b_i = c_{\mathcal{M}} b_i^g$. We use an unsquared geodesic variant of Chamfer distance (Fan et al., 2017) and a quantile-based Hausdorff distance (Taha & Hanbury, 2015):

$$\text{G-Chamfer} = \tfrac{1}{2}\left(\frac{1}{N}\sum_j a_j + \frac{1}{N}\sum_i b_i\right), \qquad \mathrm{HD95} = \max\{Q_{0.95}(a), Q_{0.95}(b)\}.$$

HD95 takes the maximum of the two directed quantiles, not the quantile of their pooled union. For fixed train/validation splits $X_{\mathrm{tr}}, X_{\mathrm{va}}$,

$$\tau_{95} = Q_{0.95}\left(\left\{\min_{x\in X_{\mathrm{tr}}} \delta_{\mathcal{M}}(v,x) : v \in X_{\mathrm{va}}\right\}\right), \quad \mathrm{CVR95} = \frac{1}{N}\sum_j \mathbf{1}\left[\min_{x\in X_{\mathrm{tr}}} \delta_{\mathcal{M}}(y_j,x) > \tau_{95}\right].$$

This threshold is fixed before examining generated samples. CVR95 is a violation rate, not coverage recall. Smaller t2g90, Chamfer, HD95, and CVR95 are better. Chamfer/HD95 use degrees or curvature-radius units; the other scores are dimensionless.

All reported quantiles use linear interpolation between adjacent order statistics.

Statistical reporting. Training seed is the replication unit (Section 6.1); $\pm$ denotes sample standard deviation where shown. Comparisons use fixed splits, matched sample counts, and the prespecified generation seed. Family summaries weight dataset means equally; runtime uses wall-clock medians.

F.6 Generation-time measurement

Each geometric timing invocation generates 100,000 fresh samples in one native sampler call, without external chunking or microbatching, on an NVIDIA A100-SXM4-80GB in FP32 with TF32 disabled. Timing includes native sampling, required output processing, GPU-to-CPU transfer, and final synchronization; loading, compilation, metrics, and saving are excluded. All geometric methods use PyTorch eager execution without compilation or CUDA Graphs, with 20 warm-ups and 100 measured invocations. For method h and dataset d, complete-call medians and their geometry-family averages are

$$t_{h,d} = \mathrm{median}_r\big(t_{h,d,r}^{\mathrm{end}} - t_{h,d,r}^{\mathrm{start}}\big), \qquad \overline{\tau}_{\mathcal{G}}(h) = \frac{1}{|\mathcal{D}_{\mathcal{G}}|}\sum_{d\in\mathcal{D}_{\mathcal{G}}} t_{h,d}. \tag{112}$$

Per-dataset results are in Appendix G; none is rescaled from test-count timing. DNA uses the separate count and batch size in Appendix J.5.

G Additional Quantitative Results and Efficiency Analysis

These tables complement the main results with validation-normalized tail distance (t2g90), bidirectional geodesic error (G-Chamfer), validation-calibrated distance-threshold violations (CVR95), and per-dataset generation time. Definitions and timing follow Appendix F. RFM is the iterative reference above the rule; bold and underline rank RiCoTA and the one- or few-step baselines using unrounded values. In the tables, t2g abbreviates t2g90 and G-Ch. abbreviates G-Chamfer.

G.1 Distribution and support

Table 11: Per-dataset distribution and support results on $\mathbb{S}^2$.

Method	Eq.			Fire			Flood			Volc.		
	sF1↑	MMD↓	t2g↓	sF1↑	MMD↓	t2g↓	sF1↑	MMD↓	t2g↓	sF1↑	MMD↓	t2g↓
RFM	0.907	0.025	0.807	0.908	0.025	0.799	0.911	0.067	0.737	0.922	0.113	0.763
G-LSD	0.857	0.041	0.972	0.806	0.031	1.504	0.840	**0.038**	1.029	0.886	0.142	0.867
RCM/RCT	0.850	0.052	0.994	0.791	0.046	1.419	0.868	0.073	0.944	0.890	0.152	0.874
RMF-MT	0.794	0.051	1.358	0.617	0.052	3.950	0.808	0.043	1.178	0.729	0.156	3.315
M-FFF	0.856	0.042	0.904	0.825	0.030	1.038	0.878	0.069	0.879	0.890	0.147	0.818
RiCoTA	**0.898**	**0.027**	**0.795**	**0.866**	**0.023**	**0.919**	**0.900**	0.044	**0.826**	**0.927**	**0.084**	**0.696**

Table 12: Per-dataset distribution and support results on product tori ($\mathbb{T}^2$ proteins and $\mathbb{T}^7$ RNA).

Method	Gen.			Gly.			Pro.			Pre-Pro			RNA		
	sF1↑	MMD↓	t2g↓	sF1↑	MMD↓	t2g↓	sF1↑	MMD↓	t2g↓	sF1↑	MMD↓	t2g↓	sF1↑	MMD↓	t2g↓
RFM	0.909	0.007	0.761	0.900	0.040	0.752	0.904	0.041	0.767	0.910	0.037	0.733	0.757	0.033	0.910
G-LSD	0.890	0.015	0.942	0.885	0.038	0.915	0.855	0.041	1.175	0.858	0.053	1.026	0.615	0.217	1.414
RCM/RCT	0.876	0.124	0.942	0.857	0.099	0.945	0.762	0.223	1.147	0.761	0.280	1.165	0.545	0.547	2.015
RMF-MT	0.816	0.053	2.049	0.868	0.040	0.999	0.814	0.055	1.611	0.822	0.076	1.371	0.525	0.196	1.924
M-FFF	0.906	0.014	**0.775**	0.902	0.038	**0.771**	0.890	0.023	**0.801**	0.891	0.034	0.788	0.580	0.425	1.333
RiCoTA	**0.906**	**0.009**	0.782	**0.904**	**0.023**	0.780	**0.894**	**0.018**	0.801	**0.902**	**0.031**	**0.779**	**0.743**	**0.040**	**1.052**

Table 13: Per-dataset distribution and support results on $\mathrm{SO}(3)$.

Method	Cone			Fisher24			Line			Swiss		
	sF1↑	MMD↓	t2g↓	sF1↑	MMD↓	t2g↓	sF1↑	MMD↓	t2g↓	sF1↑	MMD↓	t2g↓
RFM	0.090	0.105	7.287	0.150	0.129	10.26	0.657	0.080	2.007	0.257	0.201	9.668
G-LSD	0.055	0.152	6.146	0.161	0.087	6.547	0.667	0.076	1.756	0.127	0.523	13.11
RCM/RCT	0.520	0.098	1.370	0.329	0.087	2.613	0.688	0.063	1.373	0.856	0.131	0.951
RMF-MT	0.090	0.096	5.252	0.163	0.091	6.386	0.674	0.055	1.486	0.234	0.229	4.426
M-FFF	0.784	0.039	**0.883**	0.518	0.040	2.822	0.691	0.025	1.229	0.051	0.765	6.151
RiCoTA	**0.809**	**0.019**	1.116	**0.836**	**0.022**	**0.858**	**0.844**	**0.024**	**0.899**	**0.906**	**0.007**	**0.784**

G.2 GEODESIC SHAPE AND TAIL COVERAGE

Table 14: Per-dataset geodesic shape results on $\mathbb{S}^2$. G-Chamfer and HD95 are in degrees; CVR95 is dimensionless.

Method	Eq.			Fire			Flood			Volc.		
	G-Ch.↓	HD95↓	CVR95↓	G-Ch.↓	HD95↓	CVR95↓	G-Ch.↓	HD95↓	CVR95↓	G-Ch.↓	HD95↓	CVR95↓
RFM	1.479	4.881	0.030	0.596	2.016	0.011	1.991	6.360	0.025	2.449	19.42	0.000
G-LSD	1.825	5.556	0.258	1.032	4.219	0.291	2.964	9.132	0.280	3.293	20.58	**0.193**
RCM/RCT	2.230	9.841	0.303	1.469	8.606	0.364	2.674	9.377	0.202	3.774	21.10	0.473
RMF-MT	2.569	9.310	0.410	2.708	12.09	0.679	3.484	12.30	0.356	6.873	21.88	0.917
M-FFF	2.153	9.860	0.272	1.283	8.003	0.317	2.622	10.74	0.159	3.326	19.16	**0.193**
RiCoTA	**1.602**	**5.221**	**0.110**	**0.718**	**2.210**	**0.159**	**2.193**	**7.626**	**0.083**	**2.978**	**15.66**	0.220

Table 15: Per-dataset geodesic shape results on product tori ($\mathbb{T}^2$ proteins and $\mathbb{T}^7$ RNA). G-Chamfer and HD95 are in degrees; CVR95 is dimensionless.

Method	Gen.			Gly.			Pro.			Pre-Pro			RNA		
	G-Ch.↓	HD95↓	CVR95↓	G-Ch.↓	HD95↓	CVR95↓	G-Ch.↓	HD95↓	CVR95↓	G-Ch.↓	HD95↓	CVR95↓	G-Ch.↓	HD95↓	CVR95↓
RFM	0.636	1.791	0.053	2.849	9.161	0.052	1.485	4.654	0.065	2.323	5.480	0.045	16.02	60.41	0.059
G-LSD	0.847	2.753	0.076	3.205	11.58	0.083	2.175	7.309	0.107	3.576	19.80	0.130	27.98	98.88	0.340
RCM/RCT	1.772	16.92	0.248	4.572	24.66	0.228	6.480	45.56	0.475	9.694	60.07	0.469	56.65	162.16	0.829
RMF-MT	1.702	13.66	0.172	3.392	12.79	0.095	2.915	16.09	0.175	4.778	34.47	0.186	35.33	152.80	0.364
M-FFF	0.766	2.285	0.071	2.876	8.728	0.052	2.143	8.332	0.089	3.101	14.25	0.098	51.63	182.86	0.688
RiCoTA	**0.663**	**1.893**	**0.051**	**2.757**	**8.711**	**0.034**	**1.644**	**5.511**	**0.070**	**2.526**	**5.492**	**0.046**	**16.73**	**64.57**	**0.108**

Table 16: Per-dataset geodesic shape results on $\mathrm{SO}(3)$. G-Chamfer and HD95 are in degrees; CVR95 is dimensionless.

Method	Cone			Fisher24			Line			Swiss		
	G-Ch.↓	HD95↓	CVR95↓	G-Ch.↓	HD95↓	CVR95↓	G-Ch.↓	HD95↓	CVR95↓	G-Ch.↓	HD95↓	CVR95↓
RFM	8.472	34.42	0.988	22.40	50.89	0.988	12.08	27.88	0.702	12.45	96.58	0.976
G-LSD	11.90	64.47	0.995	20.31	46.24	0.986	11.58	26.28	0.697	30.25	129.40	0.991
RCM/RCT	5.446	49.40	0.644	13.62	43.73	0.909	10.66	22.54	0.683	0.750	2.461	0.300
RMF-MT	9.343	54.67	0.988	20.23	46.23	0.986	11.17	24.87	0.699	12.75	74.97	0.981
M-FFF	2.835	27.35	0.194	10.22	36.09	0.551	10.44	23.89	0.691	48.33	149.87	0.997
RiCoTA	**0.530**	**1.237**	**0.043**	**1.949**	**4.020**	**0.053**	**6.395**	**13.09**	**0.173**	**0.530**	**1.219**	**0.043**

G.3 PER-DATASET LATENCY AND HYPERBOLIC RESULTS

Table 17: Host-ready time (ms) for 100,000 samples on the 13 common benchmarks; entries are measured medians. RFM is the iterative reference; bold marks the fastest method excluding the iterative reference.

Method	Eq.	Fire	Flood	Volc.	Gen.	Gly.	Pro.	Pre-Pro	RNA	Cone	Fisher24	Line	Swiss
RFM	394350	393968	249488	490760	10108	34009	10104	10105	18023	13885	13902	13903	13880
G-LSD	483.23	483.20	308.60	600.58	21.59	65.90	21.64	21.71	36.35	76.28	78.00	76.29	77.93
RCM/RCT	130.20	130.06	130.07	130.31	129.71	129.59	129.56	129.52	130.07	138.20	137.74	138.12	136.67
RMF-MT	484.04	483.06	308.57	601.93	21.59	65.72	21.55	21.58	36.37	75.87	76.90	76.10	177.47
M-FFF	18.46	18.85	18.45	18.43	25.04	22.03	22.02	21.80	23.01	87.04	88.64	86.00	86.04
RiCoTA	**9.09**	**7.41**	**11.95**	**8.87**	**5.19**	**7.22**	**4.26**	**7.31**	**7.19**	**13.07**	**28.50**	**19.16**	**61.45**

Table 18: Hyperbolic $\mathbb{H}^2$ results and measured 100k time (ms). HD95 is in curvature-radius units; bold marks the better fast method.

Method	sF1↑	HD95↓	MMD↓	$F_{1,\mathrm{loc}}$ ↑	t2g↓	100k ms↓
RFM	0.898	0.183	0.091	0.642	0.823	51882
G-LSD	0.899	0.146	0.099	0.648	0.847	30.65
RiCoTA	**0.905**	**0.132**	**0.023**	**0.659**	**0.775**	**7.60**

Summary. The metrics distinguish typical geometric error, difficult-region coverage, and distance-threshold violations. On Cone, M-FFF has lower t2g90 than RiCoTA, whereas RiCoTA has higher sF1 and lower HD95. RFM retains strong torus support and t2g90 at much higher sampling cost.

H Ablations and Interface Diagnostics

H.1 Component comparisons across geometry families

Table 19 extends the component comparison in Table 5 with normalized tail distance, geodesic shape, and distance-threshold violations, reporting variability across five training seeds.

Identity uses the same canonicalization and marginal-CDF procedures, latent width m_0, alphabet size K_0, and prior architecture as Full, but sets $A_j(u) = u$. Both variants train through the finite channel, so their comparison tests response-adaptive allocation at a matched code budget.

Table 19: Geometry-family component comparisons (mean ± sample SD over five training seeds). Within each seed, datasets receive equal weight within families; product tori combine protein and RNA. HD95 and G-Chamfer use degrees except intrinsic units on $\mathbb{H}^2$; the other metrics are dimensionless. Bold marks the best mean.

Family	Metric	Full	Clean-Only	Identity
$\mathbb{S}^2$	$F_1(0.5)\uparrow$	$\mathbf{0.6999} \pm 0.0204$	0.6711 ± 0.0106	0.6747 ± 0.0125
	HD95↓	$\mathbf{7.678} \pm 1.293$	8.426 ± 0.899	9.495 ± 1.662
	MMD↓	$\mathbf{0.04452} \pm 0.00864$	0.04704 ± 0.00557	0.05157 ± 0.00625
	t2g90↓	$\mathbf{0.8090} \pm 0.0474$	0.8314 ± 0.0134	0.8670 ± 0.0262
	G-Chamfer↓	$\mathbf{1.872} \pm 0.176$	1.980 ± 0.154	2.099 ± 0.126
	CVR95↓	$\mathbf{0.1430} \pm 0.0228$	0.1649 ± 0.0128	0.1805 ± 0.0193
$\mathrm{SO}(3)$	$F_1(0.5)\uparrow$	$\mathbf{0.4780} \pm 0.0074$	0.4692 ± 0.0066	0.4654 ± 0.0068
	HD95↓	$\mathbf{4.893} \pm 0.077$	5.115 ± 0.096	5.180 ± 0.154
	MMD↓	$\mathbf{0.01799} \pm 0.00104$	0.02072 ± 0.00123	0.01982 ± 0.00198
	t2g90↓	$\mathbf{0.9142} \pm 0.0103$	0.9335 ± 0.0213	0.9375 ± 0.0308
	G-Chamfer↓	$\mathbf{2.351} \pm 0.018$	2.409 ± 0.039	2.407 ± 0.023
	CVR95↓	$\mathbf{0.0778} \pm 0.0084$	0.0927 ± 0.0098	0.0922 ± 0.0097
Product tori	$F_1(0.5)\uparrow$	$\mathbf{0.5501} \pm 0.0069$	0.5442 ± 0.0042	0.5399 ± 0.0139
	HD95↓	$\mathbf{17.236} \pm 0.772$	18.488 ± 0.443	19.379 ± 0.301
	MMD↓	$\mathbf{0.02382} \pm 0.00147$	0.02978 ± 0.00476	0.03133 ± 0.00210
	t2g90↓	$\mathbf{0.8387} \pm 0.0102$	0.8511 ± 0.0078	0.9325 ± 0.1145
	G-Chamfer↓	$\mathbf{4.863} \pm 0.074$	5.034 ± 0.029	4.929 ± 0.097
	CVR95↓	$\mathbf{0.0616} \pm 0.0012$	0.0636 ± 0.0031	0.0627 ± 0.0040
$\mathbb{H}^2$	$F_1(0.5)\uparrow$	$\mathbf{0.6593} \pm 0.0114$	0.6521 ± 0.0140	0.6503 ± 0.0111
	HD95↓	$\mathbf{0.13193} \pm 0.00319$	0.14540 ± 0.00665	0.14015 ± 0.01234
	MMD↓	$\mathbf{0.02283} \pm 0.00398$	0.02733 ± 0.00391	0.02936 ± 0.00278
	t2g90↓	$\mathbf{0.7751} \pm 0.0146$	0.7887 ± 0.0205	0.8291 ± 0.0256
	G-Chamfer↓	$\mathbf{0.04487} \pm 0.00051$	0.04718 ± 0.00058	0.04556 ± 0.00082
	CVR95↓	0.0581 ± 0.0031	0.0670 ± 0.0034	$\mathbf{0.0344} \pm 0.0046$

Full improves mean $F_1(0.5)$, HD95, MMD, t2g90, and G-Chamfer over Identity in every geometry family: MMD decreases by 9.2–24.0%, and local F_1 increases by 0.90–2.52 percentage points. These consistent gains support response-aware cell placement within the shared canonical-coordinate and code-prior design. The frozen-representation comparisons in Appendices H.2 and H.4 separately isolate cell allocation, complementing this end-to-end evidence. CVR95 favors Identity on $\mathbb{H}^2$.

Compared with Clean-Only, Full improves mean sF1 on 13 of the 14 manifold benchmarks (except Proline) and local $F_1(0.5)$ on 12 (except General and Proline). Mean t2g90 decreases on 11, with Fire, Flood, and Proline the exceptions.

H.2 HELD-OUT RESPONSE AND ALLOCATION DIAGNOSTICS

The four diagnostics test response prediction, axis ordering, and cell allocation on the same 70 frozen Phase 1 checkpoints across 14 datasets, with five model seeds per dataset and no prior. Interface statistics use training data only; held-out measurements do not select checkpoints or hyperparameters.

Response-prediction calibration. For a uniform probability-space cell (j, k), the checkpoint scale $a_{\mathcal{M}}$ in Eq. (19) and the floored train response in Eq. (27) give the high-rate prediction of mean squared intrinsic distortion conditional on assignment to that cell

$$\widehat{D}_{j,k} = \frac{\gamma_{j,k}^{\text{floor}} a_{\mathcal{M}}^2}{6K_0^2}.$$

Observed distortion averages squared displacement from the unperturbed decoder output over test observations assigned to each cell, using eight independent uniform draws that resample only coordinate j. Retained cells have at least five test assignments and finite, positive predicted and observed means. Pooled across seeds, these cell pairs define Spearman correlation ρ and a log–log ordinary-least-squares slope with a fitted intercept. Figure 5 summarizes them by predicted-distortion quantiles, with geometric-mean points and observed interquartile bands. Their agreement with the identity line supports train-response prediction of held-out coordinatewise risk.

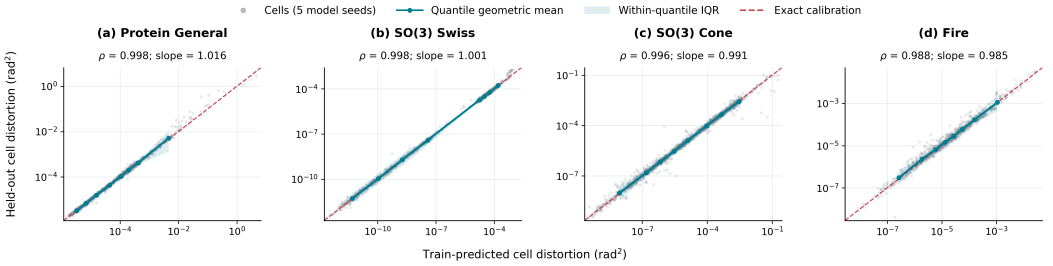


Figure 5: Train-predicted versus held-out mean coordinate–cell distortion. The diagonal represents equality between predicted and observed squared intrinsic distortion.

Response-ranked risk coverage. Within each active coordinate, rank the same uniform probability-space cells by train response and retain the highest-response $\lceil \alpha K_0 \rceil$ cells. Coverage is the sum of squared displacements for test assignments in these cells divided by the corresponding sum over all active coordinates and all test assignments. It weights cells by their observed occupancy, rather than summing their conditional means. The train-response ranking captures much of the held-out distortion in a small fraction of cells (Figure 6), supporting response-guided allocation. For this and the remaining curve diagnostics, lines are seed means and shaded bands denote one sample standard deviation.

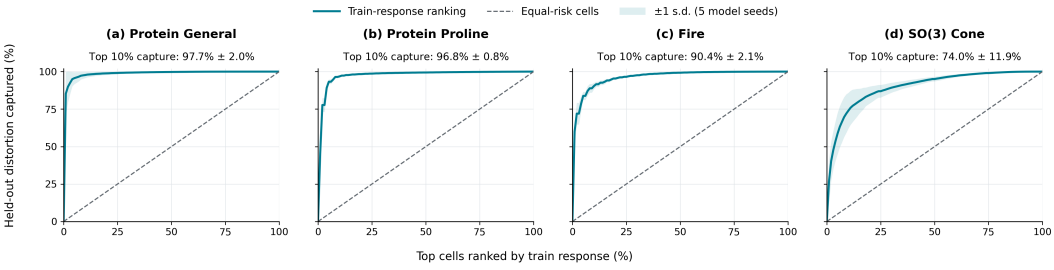


Figure 6: Held-out distortion captured by train-response-ranked cells. Annotations give the top-10% capture; the diagonal is the equal-risk reference.

Canonical-axis fidelity. For held-out canonical latent ξ and retained-axis set S, keep ξ_j for $j \in S$ and replace other coordinates by their training medians. Let $E_{\text{axis}}(S)$ be the mean squared intrinsic distance to the same checkpoint's full-axis output. Define

$$F_{\text{axis}}(S) = 1 - \frac{E_{\text{axis}}(S)}{E_{\text{axis}}(\varnothing)}.$$

The reference is the full decoder output, not the original observation; $E_{\text{axis}}(\varnothing)$ is the center-only error. At each cardinality q, the baseline is the median over all q-axis subsets, not one random ordering. Half-width values are linearly interpolated where needed. Canonical prefixes outperform the all-subset median at half width on 13 of 14 datasets, with Volcano the exception (Figure 7; Table 20).

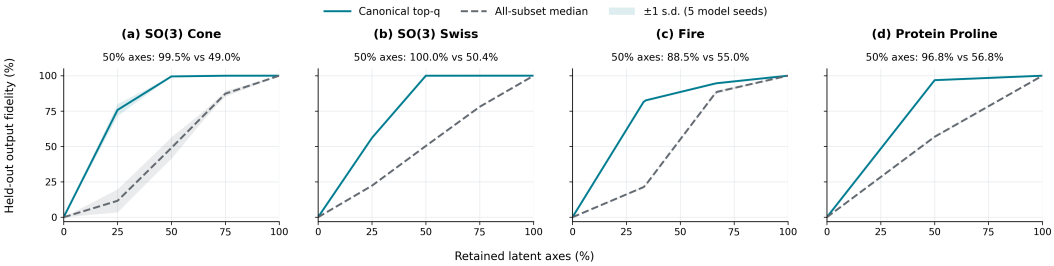


Figure 7: Full-output fidelity retained by canonical top-q axes versus the median over all equally sized axis subsets.

Cell-distortion concentration. Uniform cells and the compander share the checkpoint, gauge, CDFs, token budget, active mask, and held-out examples. For each test observation and perturbed coordinate, the same eight uniform-noise draws are mapped through each partition's own cell boundaries. Let $n_{j,k}^{(v)}$ be the number of test observations assigned to cell (j,k) under variant v, and $M_{j,k}^{(v)}$ their mean squared displacement over those observations and draws. Its contribution to total distortion is

$$D_{j,k}^{(v)} = \frac{n_{j,k}^{(v)}}{N} M_{j,k}^{(v)}, \qquad N = \text{test-set size},$$

with zero contribution for an empty cell. Thus $M_{j,k}^{(v)}$ is a conditional cell mean, whereas $D_{j,k}^{(v)}$ is its occupancy-weighted contribution. Sort the contributions in ascending order and divide their cumulative sum by $\sum_{j,k} D_{j,k}^{(v)}$. Unlike risk coverage, this ranking uses each variant's observed distortion, not train response. Response-aware cells distribute error more evenly in Figure 8; Appendix H.3 checks whether its total also decreases.

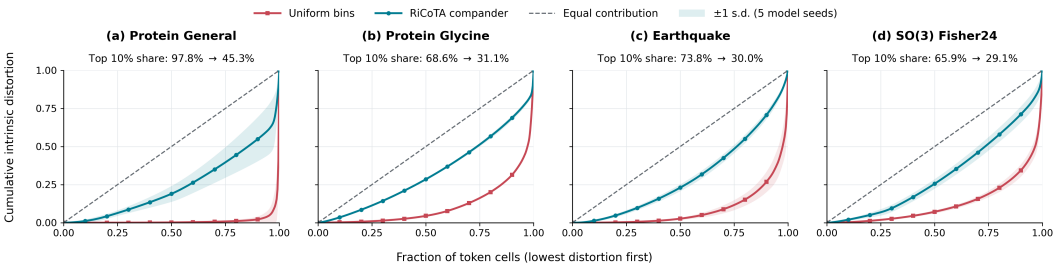


Figure 8: Cumulative coordinatewise distortion for uniform cells and the compander, sorted from lowest to highest contribution within each variant.

Table 20: Per-dataset held-out mechanism diagnostics with five model seeds per dataset. ρ and log–log slope are computed from pooled retained cell pairs as defined in Appendix H.2; other entries are seed means. Risk top-10 ranks cells by train response. Axis advantage is the canonical-prefix fidelity minus the all-subset median at half width, in percentage points. Cell top-10 ranks each variant by observed distortion contribution and gives Uniform / RiCoTA shares.

Dataset	$\rho \uparrow$	Slope$\rightarrow 1$	Risk top-10 (%)$\uparrow$	Axis advantage (pp)$\uparrow$	Cell top-10 U / R (%)$\downarrow$
Earthquake	0.981	0.971	72.2	+16.3	73.8 / 30.0
Flood	0.959	0.956	59.1	+20.9	62.1 / 28.6
Fire	0.988	0.985	90.4	+33.5	90.6 / 48.8
Volcano	0.949	0.875	83.8	-10.4	93.7 / 67.5
SO(3) Cone	0.996	0.991	74.0	+50.4	86.6 / 43.6
SO(3) Fisher24	0.989	0.993	64.4	+7.0	65.9 / 29.1
SO(3) Line	0.978	0.995	72.0	+28.9	76.7 / 40.6
SO(3) Swiss	0.998	1.001	56.8	+49.6	67.8 / 37.6
Protein General	0.998	1.016	97.7	+35.2	97.8 / 45.3
Protein Glycine	0.985	1.018	65.5	+16.7	68.6 / 31.1
Protein Proline	0.961	1.023	96.8	+40.0	97.1 / 60.5
Protein Pre-Pro	0.983	1.031	95.2	+21.7	95.7 / 68.4
RNA	0.989	1.009	91.7	+15.2	91.0 / 46.8
Hyperbolic $\mathbb{H}^2$	0.956	1.025	81.0	+10.5	82.0 / 44.6

For Volcano, each run retains fewer than ten cells with at least five held-out assignments, giving 33 eligible cells in total.

H.3 Total coordinatewise diagnostic distortion

To distinguish error reduction from redistribution, sum the occupancy-weighted cell contributions without converting them into cumulative shares: $T_v = \sum_{j,k} D^{(v)}_{j,k}$. This is a sum of one-coordinate-at-a-time errors, not the joint loss from simultaneous quantization. Table 21 reports the totals and Gini coefficients, which measure inequality among cell contributions. Mean total distortion and Gini both decrease on all 14 datasets. Relative reduction is $100(1 - \overline{T}_{\mathrm{RiCoTA}}/\overline{T}_{\mathrm{Uniform}})$; absolute totals are not pooled across geometries.

Table 21: Magnitude and inequality of matched-budget coordinatewise distortion. Totals and Gini are seed means; reductions use unrounded mean totals. Totals are in squared radians on angular manifolds and squared curvature-radius units on $\mathbb{H}^2$. Gini is shown as Uniform / RiCoTA.

Dataset	$\overline{T}_{\mathrm{Uniform}}$	$\overline{T}_{\mathrm{RiCoTA}}$	Reduction (%)	Gini U / R
Earthquake	7.758×10^{-4}	2.469×10^{-4}	68.2	0.822 / 0.400
Flood	2.314×10^{-4}	1.129×10^{-4}	51.2	0.748 / 0.424
Fire	1.405×10^{-3}	2.361×10^{-4}	83.2	0.926 / 0.565
Volcano	1.069×10^{-2}	6.571×10^{-4}	93.9	0.939 / 0.797
SO(3) Cone	4.292×10^{-3}	1.034×10^{-3}	75.9	0.919 / 0.700
SO(3) Fisher24	2.263×10^{-3}	1.150×10^{-3}	49.2	0.739 / 0.374
SO(3) Line	3.528×10^{-3}	1.600×10^{-3}	54.7	0.823 / 0.587
SO(3) Swiss	1.926×10^{-4}	1.299×10^{-4}	32.6	0.811 / 0.627
Protein General	2.023×10^{-2}	6.881×10^{-4}	96.6	0.975 / 0.514
Protein Glycine	1.182×10^{-3}	5.296×10^{-4}	55.2	0.777 / 0.347
Protein Proline	4.860×10^{-3}	6.905×10^{-4}	85.8	0.947 / 0.678
Protein Pre-Pro	7.773×10^{-3}	1.231×10^{-3}	84.2	0.954 / 0.746
RNA	1.988×10^{-1}	2.257×10^{-2}	88.6	0.911 / 0.528
Hyperbolic $\mathbb{H}^2$	9.667×10^{-3}	4.101×10^{-3}	57.6	0.838 / 0.510

H.4 LATENT-RESOLUTION SCALING

Frozen-representation protocol. For each dataset and evaluated model, we keep the Phase 1 encoder, decoder, gauge, and marginal transform fixed. We recalibrate the response-aware finite interface at $K_0 \in \{8, 16, 32, 64, 128, 256\}$ from training data and compare with uniform probability-space cells at the same K_0. Each test observation and active coordinate uses eight exact-cell reconstructions, with the same uniform noise mapped through the two partitions. Neither neural retraining nor Phase 2 is involved. The reported distortion sums the one-coordinate-at-a-time squared intrinsic decoder displacements, not the joint reconstruction loss or generation error. Tables 23–24 give means and sample standard deviations over model seeds.

Training-context support. A dataset–K_0 pair is support-valid only when every active response interval has training contexts for every evaluated model. Volcano has 663 training observations and tied latent support; its $K_0 = 64, 128, 256$ recalibrations contain empty active intervals. They are marked *N/A: insufficient calibration support* and excluded from the paired comparisons and aggregate curves; valid uniform controls remain listed. This restriction does not affect the primary Volcano model at $K_0 = 36$.

Aggregation and decay. Let $D_{d,v}(K_0)$ be the mean coordinatewise distortion for dataset d and cell rule v. Both curves in Figure 4(a) use the common denominator $D_{d,\mathrm{Uniform}}(8)$. We plot the median and interquartile range across the 13 datasets supported at all six resolutions, excluding Volcano and including $\mathbb{H}^2$; raw distortions with different geometric units are not pooled. The K_0^{-2} reference is anchored to the response-aware median at $K_0 = 32$. Fitting $\log D$ against $\log K_0$ at $32, 64, 128, 256$ gives a mean decay exponent of 1.92 over these 13 datasets. Across all support-valid pairs, response-aware seed means are lower in 81/81 settings, including Volcano at $8, 16, 32$. This is a comparison of means, not a claim of statistical significance in every setting.

Table 22: Training-context support and distortion decay. Coverage counts dataset–K_0 settings supported for every evaluated model. The exponent β fits $D \propto K_0^{-\beta}$ to the response-aware means at $K_0 = 32, 64, 128, 256$.

Dataset	Supported K_0 settings	β	Fit R^2
Earthquake	6/6	1.932	0.9991
Fire	6/6	1.919	0.9957
Flood	6/6	1.705	0.9802
Volcano	3/6	N/A	N/A
SO(3)-Cone	6/6	2.029	0.9999
SO(3)-Fisher24	6/6	2.020	0.9998
SO(3)-Line	6/6	1.984	0.9998
SO(3)-Swiss	6/6	1.927	0.9999
Protein General	6/6	2.440	0.9923
Protein Glycine	6/6	1.947	1.0000
Protein Proline	6/6	1.423	0.9812
Protein Pre-Pro	6/6	1.705	0.9970
RNA	6/6	1.889	0.9985
$\mathbb{H}^2$	6/6	1.994	0.9999

Table 23: Frozen-representation coordinatewise distortion (1/2), following Appendix H.4. Values are means ± sample standard deviations across model seeds, in rad^2. N/A: insufficient calibration support.

Dataset	K_0	Uniform cells	RiCoTA compander	Uniform/RiCoTA
Earthquake	8	$(3.02 \pm 0.50) \times 10^{-2}$	$(1.63 \pm 0.11) \times 10^{-2}$	1.85
	16	$(8.67 \pm 1.92) \times 10^{-3}$	$(4.17 \pm 0.13) \times 10^{-3}$	2.08
	32	$(2.47 \pm 0.16) \times 10^{-3}$	$(1.04 \pm 0.10) \times 10^{-3}$	2.39
	64	$(7.76 \pm 1.94) \times 10^{-4}$	$(2.47 \pm 0.11) \times 10^{-4}$	3.14
	128	$(2.04 \pm 0.34) \times 10^{-4}$	$(6.52 \pm 0.26) \times 10^{-5}$	3.13
	256	$(6.33 \pm 2.18) \times 10^{-5}$	$(1.86 \pm 0.18) \times 10^{-5}$	3.40
Fire	8	$(3.59 \pm 0.24) \times 10^{-2}$	$(1.87 \pm 0.15) \times 10^{-2}$	1.91
	16	$(1.44 \pm 0.11) \times 10^{-2}$	$(4.27 \pm 0.44) \times 10^{-3}$	3.38
	32	$(3.78 \pm 0.97) \times 10^{-3}$	$(1.07 \pm 0.10) \times 10^{-3}$	3.54
	64	$(1.41 \pm 0.31) \times 10^{-3}$	$(2.36 \pm 0.43) \times 10^{-4}$	5.95
	128	$(4.36 \pm 1.13) \times 10^{-4}$	$(6.11 \pm 1.65) \times 10^{-5}$	7.14
	256	$(1.33 \pm 0.45) \times 10^{-4}$	$(1.99 \pm 1.12) \times 10^{-5}$	6.71
Flood	8	$(2.46 \pm 0.13) \times 10^{-2}$	$(1.64 \pm 0.11) \times 10^{-2}$	1.51
	16	$(6.47 \pm 0.38) \times 10^{-3}$	$(4.00 \pm 0.31) \times 10^{-3}$	1.62
	32	$(1.84 \pm 0.14) \times 10^{-3}$	$(1.01 \pm 0.07) \times 10^{-3}$	1.81
	64	$(4.75 \pm 0.36) \times 10^{-4}$	$(2.78 \pm 0.26) \times 10^{-4}$	1.71
	128	$(1.47 \pm 0.17) \times 10^{-4}$	$(6.38 \pm 0.16) \times 10^{-5}$	2.31
	256	$(3.87 \pm 0.85) \times 10^{-5}$	$(3.23 \pm 0.76) \times 10^{-5}$	1.20
Volcano	8	$(5.45 \pm 3.08) \times 10^{-2}$	$(2.36 \pm 0.81) \times 10^{-2}$	2.31
	16	$(3.23 \pm 2.78) \times 10^{-2}$	$(1.46 \pm 1.13) \times 10^{-2}$	2.21
	32	$(1.54 \pm 1.83) \times 10^{-2}$	$(1.34 \pm 0.35) \times 10^{-3}$	11.48
	64	$(2.86 \pm 3.64) \times 10^{-3}$	N/A	N/A
	128	$(4.06 \pm 2.60) \times 10^{-4}$	N/A	N/A
	256	$(1.13 \pm 1.13) \times 10^{-4}$	N/A	N/A
SO(3)-Cone	8	$(3.51 \pm 2.12) \times 10^{-1}$	$(1.45 \pm 0.85) \times 10^{-1}$	2.43
	16	$(1.12 \pm 0.75) \times 10^{-1}$	$(4.03 \pm 2.03) \times 10^{-2}$	2.77
	32	$(3.49 \pm 2.54) \times 10^{-2}$	$(1.02 \pm 0.57) \times 10^{-2}$	3.43
	64	$(1.02 \pm 0.80) \times 10^{-2}$	$(2.41 \pm 1.35) \times 10^{-3}$	4.25
	128	$(2.47 \pm 1.94) \times 10^{-3}$	$(5.96 \pm 3.26) \times 10^{-4}$	4.15
	256	$(6.10 \pm 4.23) \times 10^{-4}$	$(1.49 \pm 0.80) \times 10^{-4}$	4.09
SO(3)-Fisher24	8	$(1.03 \pm 0.13) \times 10^{-1}$	$(9.75 \pm 2.49) \times 10^{-2}$	1.06
	16	$(2.84 \pm 0.39) \times 10^{-2}$	$(2.06 \pm 0.51) \times 10^{-2}$	1.38
	32	$(7.95 \pm 1.14) \times 10^{-3}$	$(4.82 \pm 0.85) \times 10^{-3}$	1.65
	64	$(2.26 \pm 0.37) \times 10^{-3}$	$(1.15 \pm 0.22) \times 10^{-3}$	1.97
	128	$(6.77 \pm 1.70) \times 10^{-4}$	$(2.80 \pm 0.56) \times 10^{-4}$	2.42
	256	$(1.80 \pm 0.56) \times 10^{-4}$	$(7.25 \pm 1.40) \times 10^{-5}$	2.48
SO(3)-Line	8	$(2.24 \pm 0.09) \times 10^{-1}$	$(1.99 \pm 0.08) \times 10^{-1}$	1.12
	16	$(7.78 \pm 0.81) \times 10^{-2}$	$(5.85 \pm 0.30) \times 10^{-2}$	1.33
	32	$(2.68 \pm 0.57) \times 10^{-2}$	$(1.37 \pm 0.08) \times 10^{-2}$	1.96
	64	$(7.80 \pm 1.81) \times 10^{-3}$	$(3.69 \pm 0.32) \times 10^{-3}$	2.11
	128	$(2.04 \pm 0.38) \times 10^{-3}$	$(8.87 \pm 0.55) \times 10^{-4}$	2.30
	256	$(5.33 \pm 1.20) \times 10^{-4}$	$(2.25 \pm 0.19) \times 10^{-4}$	2.37

Table 24: Frozen-representation coordinatewise distortion (2/2), following Appendix H.4. Values are means $\pm$ sample standard deviations across model seeds, in rad^2 for angular manifolds and squared native geodesic units for $\mathbb{H}^2$.

Dataset	K_0	Uniform cells	RiCoTA compander	Uniform/RiCoTA
SO(3)-Swiss	8	$(1.00 \pm 0.02) \times 10^{-2}$	$(8.89 \pm 0.18) \times 10^{-3}$	1.13
	16	$(2.82 \pm 0.07) \times 10^{-3}$	$(2.31 \pm 0.04) \times 10^{-3}$	1.22
	32	$(8.05 \pm 0.18) \times 10^{-4}$	$(6.17 \pm 0.13) \times 10^{-4}$	1.30
	64	$(2.38 \pm 0.09) \times 10^{-4}$	$(1.63 \pm 0.04) \times 10^{-4}$	1.46
	128	$(7.38 \pm 0.26) \times 10^{-5}$	$(4.17 \pm 0.09) \times 10^{-5}$	1.77
	256	$(2.48 \pm 0.12) \times 10^{-5}$	$(1.13 \pm 0.07) \times 10^{-5}$	2.19
Protein General	8	$(4.76 \pm 1.59) \times 10^{-1}$	$(2.32 \pm 0.83) \times 10^{-1}$	2.05
	16	$(2.24 \pm 0.82) \times 10^{-1}$	$(5.08 \pm 0.53) \times 10^{-2}$	4.41
	32	$(7.00 \pm 1.84) \times 10^{-2}$	$(2.15 \pm 0.84) \times 10^{-2}$	3.25
	64	$(4.18 \pm 1.67) \times 10^{-2}$	$(5.76 \pm 2.83) \times 10^{-3}$	7.27
	128	$(2.02 \pm 1.32) \times 10^{-2}$	$(6.88 \pm 2.01) \times 10^{-4}$	29.39
	256	$(7.30 \pm 4.61) \times 10^{-3}$	$(1.56 \pm 0.46) \times 10^{-4}$	46.90
Protein Glycine	8	$(4.17 \pm 0.03) \times 10^{-2}$	$(2.61 \pm 0.02) \times 10^{-2}$	1.60
	16	$(1.17 \pm 0.01) \times 10^{-2}$	$(6.82 \pm 0.13) \times 10^{-3}$	1.71
	32	$(3.53 \pm 0.04) \times 10^{-3}$	$(1.79 \pm 0.01) \times 10^{-3}$	1.97
	64	$(1.07 \pm 0.02) \times 10^{-3}$	$(4.64 \pm 0.07) \times 10^{-4}$	2.30
	128	$(3.39 \pm 0.11) \times 10^{-4}$	$(1.19 \pm 0.03) \times 10^{-4}$	2.86
	256	$(1.04 \pm 0.05) \times 10^{-4}$	$(3.14 \pm 0.20) \times 10^{-5}$	3.32
Protein Proline	8	$(2.84 \pm 0.35) \times 10^{-1}$	$(3.35 \pm 0.59) \times 10^{-2}$	8.48
	16	$(3.40 \pm 1.09) \times 10^{-2}$	$(1.13 \pm 0.26) \times 10^{-2}$	3.00
	32	$(1.56 \pm 0.42) \times 10^{-2}$	$(2.41 \pm 0.17) \times 10^{-3}$	6.48
	64	$(4.86 \pm 0.91) \times 10^{-3}$	$(6.90 \pm 1.26) \times 10^{-4}$	7.04
	128	$(1.47 \pm 0.22) \times 10^{-3}$	$(2.43 \pm 1.15) \times 10^{-4}$	6.04
	256	$(5.53 \pm 1.12) \times 10^{-4}$	$(1.27 \pm 0.89) \times 10^{-4}$	4.34
Protein Pre-Pro	8	$(1.52 \pm 0.71) \times 10^{-1}$	$(8.25 \pm 1.01) \times 10^{-2}$	1.84
	16	$(6.17 \pm 2.43) \times 10^{-2}$	$(1.19 \pm 0.11) \times 10^{-2}$	5.18
	32	$(2.59 \pm 1.00) \times 10^{-2}$	$(3.37 \pm 0.41) \times 10^{-3}$	7.69
	64	$(7.77 \pm 4.08) \times 10^{-3}$	$(1.23 \pm 0.15) \times 10^{-3}$	6.31
	128	$(2.69 \pm 1.11) \times 10^{-3}$	$(3.59 \pm 0.69) \times 10^{-4}$	7.48
	256	$(9.70 \pm 1.54) \times 10^{-4}$	$(9.90 \pm 3.66) \times 10^{-5}$	9.81
RNA	8	$(9.13 \pm 1.04) \times 10^{-1}$	$(2.69 \pm 0.33) \times 10^{-1}$	3.40
	16	$(3.65 \pm 0.30) \times 10^{-1}$	$(5.26 \pm 0.93) \times 10^{-2}$	6.93
	32	$(1.16 \pm 0.14) \times 10^{-1}$	$(1.36 \pm 0.38) \times 10^{-2}$	8.54
	64	$(3.20 \pm 0.39) \times 10^{-2}$	$(3.40 \pm 0.50) \times 10^{-3}$	9.39
	128	$(1.08 \pm 0.34) \times 10^{-2}$	$(8.79 \pm 1.76) \times 10^{-4}$	12.23
	256	$(3.87 \pm 2.59) \times 10^{-3}$	$(2.72 \pm 0.88) \times 10^{-4}$	14.20
$\mathbb{H}^2$	8	$(1.50 \pm 0.04) \times 10^{-1}$	$(1.12 \pm 0.02) \times 10^{-1}$	1.34
	16	$(5.21 \pm 0.17) \times 10^{-2}$	$(3.22 \pm 0.09) \times 10^{-2}$	1.62
	32	$(1.84 \pm 0.04) \times 10^{-2}$	$(8.78 \pm 0.13) \times 10^{-3}$	2.10
	64	$(6.43 \pm 0.14) \times 10^{-3}$	$(2.28 \pm 0.05) \times 10^{-3}$	2.82
	128	$(2.45 \pm 0.10) \times 10^{-3}$	$(5.48 \pm 0.34) \times 10^{-4}$	4.47
	256	$(8.17 \pm 0.73) \times 10^{-4}$	$(1.41 \pm 0.06) \times 10^{-4}$	5.80

End-to-end protocol. We separately train the full model at each of the six K_0 values on Fire, Protein General, $\mathrm{SO}(3)$-Swiss, and $\mathbb{H}^2$. All non-K_0 hyperparameters are held fixed within each dataset. Figure 4(b) reports mean sF1; Table 25 gives mean and sample standard deviation, together with MMD and HD95 measurements. This sweep evaluates generation quality, not latency as a function of K_0.

Table 25: End-to-end resolution sweep. Metrics are means $\pm$ sample standard deviations at each K_0. HD95 uses degrees for Fire, General, and Swiss, and native hyperbolic units for $\mathbb{H}^2$. General is Protein General; Swiss is $\mathrm{SO}(3)$-Swiss.

Dataset	K_0	sF1 $\uparrow$	HD95 $\downarrow$	MMD $\downarrow$
Fire	8	0.722 ± 0.007	9.025 ± 0.509	0.0220 ± 0.0041
	16	0.792 ± 0.010	5.141 ± 0.634	0.0256 ± 0.0055
	32	0.837 ± 0.008	3.113 ± 0.207	0.0238 ± 0.0032
	64	0.866 ± 0.003	2.210 ± 0.102	0.0233 ± 0.0047
	128	0.886 ± 0.006	2.454 ± 0.113	0.0314 ± 0.0056
	256	0.885 ± 0.007	3.664 ± 0.812	0.0369 ± 0.0031
Protein General	8	0.761 ± 0.093	14.044 ± 8.486	0.0403 ± 0.0190
	16	0.868 ± 0.031	5.439 ± 2.732	0.0208 ± 0.0075
	32	0.891 ± 0.010	3.552 ± 1.450	0.0176 ± 0.0082
	64	0.907 ± 0.001	1.843 ± 0.031	0.0111 ± 0.0032
	128	0.906 ± 0.001	1.893 ± 0.089	0.0086 ± 0.0024
	256	0.906 ± 0.002	1.958 ± 0.103	0.0120 ± 0.0062
SO(3)-Swiss	8	0.855 ± 0.013	1.539 ± 0.093	0.0089 ± 0.0037
	16	0.892 ± 0.003	1.343 ± 0.037	0.0123 ± 0.0021
	32	0.903 ± 0.004	1.270 ± 0.014	0.0126 ± 0.0034
	64	0.904 ± 0.001	1.277 ± 0.017	0.0136 ± 0.0022
	128	0.905 ± 0.001	1.239 ± 0.016	0.0108 ± 0.0019
	256	0.905 ± 0.002	1.242 ± 0.014	0.0114 ± 0.0035
$\mathbb{H}^2$	8	0.854 ± 0.058	0.432 ± 0.210	0.0856 ± 0.0417
	16	0.894 ± 0.004	0.200 ± 0.026	0.0389 ± 0.0051
	32	0.906 ± 0.002	0.128 ± 0.004	0.0251 ± 0.0025
	64	0.900 ± 0.002	0.147 ± 0.007	0.0283 ± 0.0030
	128	0.907 ± 0.003	0.128 ± 0.004	0.0434 ± 0.0025
	256	0.903 ± 0.001	0.127 ± 0.005	0.0399 ± 0.0030

Interpretation. Retrained support fidelity improves from small alphabets and largely saturates, while MMD and HD95 need not improve monotonically. Lower frozen-representation distortion therefore does not guarantee improvement in every end-to-end metric. The sweep is descriptive and does not select a test-optimal K_0.

H.5 SENSITIVITY TO THE COMPANDING EXPONENT

We vary the companding exponent α in the allocation density

$$\lambda_{j,\alpha}(s) = \frac{[\widehat{\gamma}_j^{\mathrm{floor}}(s)]^\alpha}{\int_0^1 [\widehat{\gamma}_j^{\mathrm{floor}}(t)]^\alpha \, dt},$$

where $\alpha = 0$ gives uniform probability-space cells and $\alpha = 1/3$ is the cube-root rule.

Frozen-representation sensitivity. For all 14 datasets and five model seeds, we fix the encoder, decoder, gauge, empirical CDFs, and active coordinates learned at $\alpha = 1/3$. We evaluate $\alpha \in \{0, 1/6, 1/4, 1/3, 5/12, 1/2, 1\}$ at the native alphabet size and at $K_0 \in \{16, 64, 256\}$. Native-K_0 measurements reuse the saved responses; at each additional K_0, one train-only response estimate is shared across exponents. Empty training intervals use the nearest supported response interval only in this additional-K_0 sensitivity diagnostic; deployment calibration retains the support requirement in Appendix B.4.

Distortion $D(\alpha)$ sums the one-coordinate-at-a-time squared intrinsic displacements from the unperturbed decoder output, using eight common within-cell draws per test observation. We compute $D(\alpha)/D(0)$ within each model seed and K_0, then report the five-seed mean and sample standard deviation. At native K_0, $\alpha = 1/3$ lowers mean distortion on all 14 datasets, with a 67.8% equal-dataset mean reduction, and lies within 4.91% of the observed grid minimum on each dataset. These measurements assess allocation conditional on the learned representation.

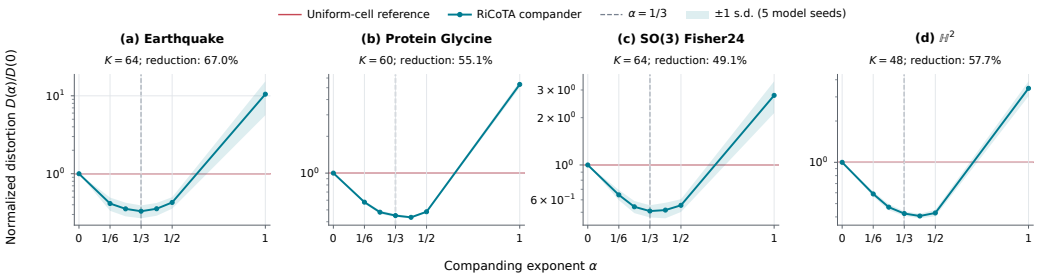


Figure 9: Frozen distortion at native alphabet sizes on four representative datasets. Annotations give the reduction at $\alpha = 1/3$ versus uniform cells; bands show one sample standard deviation across five model seeds. Vertical axes are logarithmic.

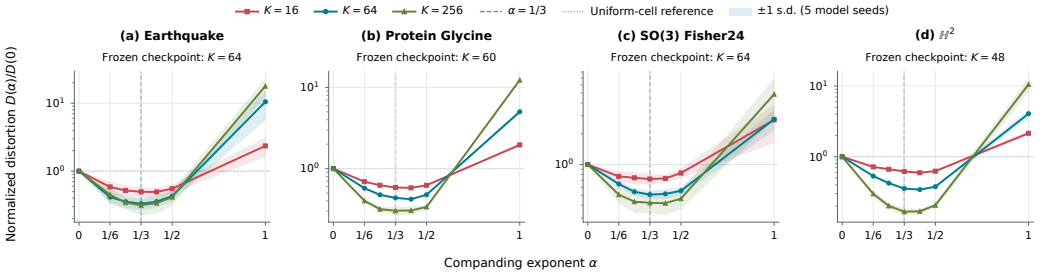


Figure 10: Frozen exponent sensitivity at $K_0 = 16, 64, 256$ on the same checkpoints. Each curve uses its own uniform-cell denominator, comparing relative allocation benefits across exponents. Means and bands use five model seeds; vertical axes are logarithmic.

Retrained generation. We evaluate $\alpha \in \{0, 1/6, 1/3, 1/2, 1\}$ on all 14 datasets, holding all non-α settings, budgets, splits, and validation-based checkpoint rules fixed within each dataset. The $\alpha = 0$ and $1/3$ entries reuse the matched Identity and Full models; the other exponents train both phases afresh, with five model seeds per setting. Each model is evaluated using one generated bank at the test-set sample count. Table 26 reports local F_1 at $\lambda = 0.5$.

Table 26: Generation sensitivity to the companding exponent. Local F_1 at $\lambda = 0.5$ (higher is better), shown as mean $\pm$ sample standard deviation across five model seeds. Each model uses one generated bank at the test-set sample count. Bold marks the highest unrounded mean in each row.

Dataset	$\alpha = 0$	$1/6$	$1/3$	$1/2$	1
Earthquake	0.669 ± 0.014	0.664 ± 0.010	$\mathbf{0.685 \pm 0.008}$	0.674 ± 0.022	0.473 ± 0.027
Fire	0.588 ± 0.006	0.610 ± 0.014	$\mathbf{0.614 \pm 0.006}$	0.582 ± 0.018	0.418 ± 0.050
Flood	0.664 ± 0.025	0.670 ± 0.020	$\mathbf{0.686 \pm 0.013}$	0.680 ± 0.017	0.630 ± 0.018
Volcano	0.777 ± 0.031	0.777 ± 0.024	$\mathbf{0.815 \pm 0.068}$	0.783 ± 0.047	0.689 ± 0.034
General	0.654 ± 0.009	0.661 ± 0.003	0.663 ± 0.003	$\mathbf{0.667 \pm 0.006}$	0.549 ± 0.085
Glycine	0.652 ± 0.004	0.655 ± 0.016	$\mathbf{0.660 \pm 0.014}$	0.651 ± 0.007	0.635 ± 0.017
Proline	0.610 ± 0.040	0.631 ± 0.008	0.632 ± 0.023	$\mathbf{0.644 \pm 0.006}$	0.574 ± 0.053
Pre-Pro	0.651 ± 0.017	0.646 ± 0.020	$\mathbf{0.665 \pm 0.011}$	0.651 ± 0.008	0.618 ± 0.045
RNA	0.132 ± 0.015	$\mathbf{0.139 \pm 0.007}$	0.131 ± 0.016	0.124 ± 0.012	0.082 ± 0.016
SO(3)-Cone	0.343 ± 0.021	0.360 ± 0.027	$\mathbf{0.369 \pm 0.010}$	0.365 ± 0.011	0.272 ± 0.040
SO(3)-Fisher24	0.406 ± 0.004	0.416 ± 0.003	$\mathbf{0.420 \pm 0.006}$	0.417 ± 0.016	0.329 ± 0.032
SO(3)-Line	0.453 ± 0.010	0.459 ± 0.008	$\mathbf{0.461 \pm 0.008}$	0.455 ± 0.013	0.427 ± 0.012
SO(3)-Swiss	0.659 ± 0.004	0.656 ± 0.007	$\mathbf{0.663 \pm 0.007}$	0.659 ± 0.005	0.655 ± 0.007
$\mathbb{H}^2$	0.650 ± 0.011	$\mathbf{0.664 \pm 0.010}$	0.659 ± 0.011	0.664 ± 0.008	0.659 ± 0.012

Interpretation. The cube-root rule improves mean local support fidelity over uniform cells on 13 of 14 datasets. Nearby exponents remain competitive, and the highest observed mean varies across datasets. The frozen allocation benefit therefore supports the cube-root default without requiring it to maximize every finite-resolution generation metric.

I QUALITATIVE COMPARISONS

The galleries compare Ground truth, RiCoTA, G-LSD, RCM/RCT, RMF-MT, and M-FFF in that order. Methods share the sample count and display convention within each dataset. These visualizations complement the test-count quantitative evaluation in Appendix G; the iterative RFM reference appears only in the quantitative comparisons.

I.1 SPHERICAL EARTH DATA

The four Earth figures compare training observations with train-count-matched generations, using one shared zoom region per dataset.

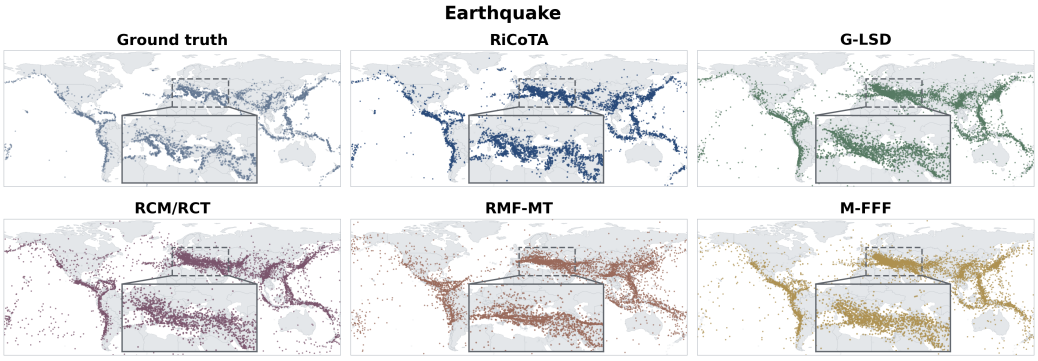


Figure 11: Earthquake on $\mathbb{S}^2$. The Mediterranean/West Asia inset resolves the seismic belt and generated mass away from it.

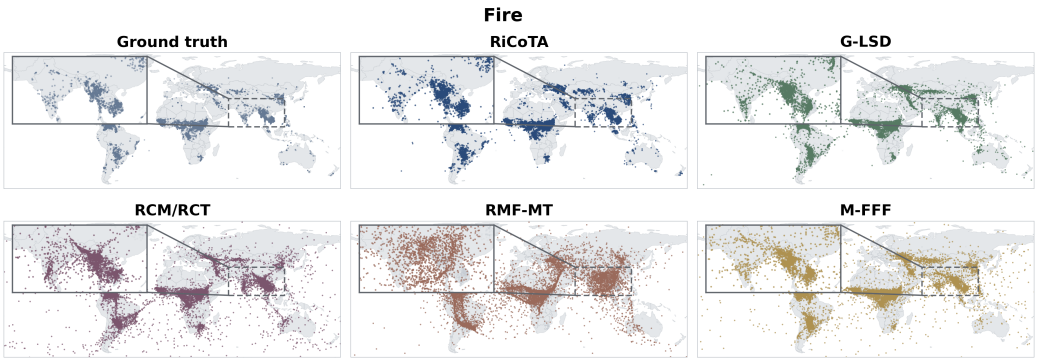


Figure 12: Fire on $\mathbb{S}^2$. The South/Southeast Asia inset resolves concentrated fire-location support and surrounding diffuse samples.

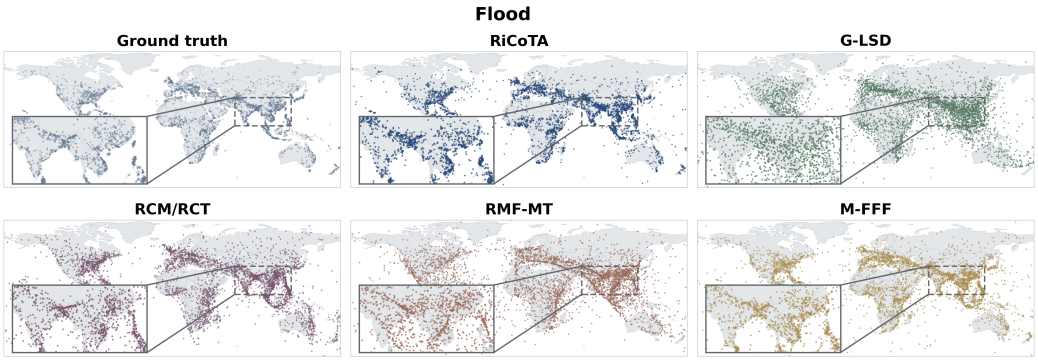


Figure 13: Flood on $\mathbb{S}^2$. The South/Southeast Asia inset resolves local concentration within the observed flood-support region.

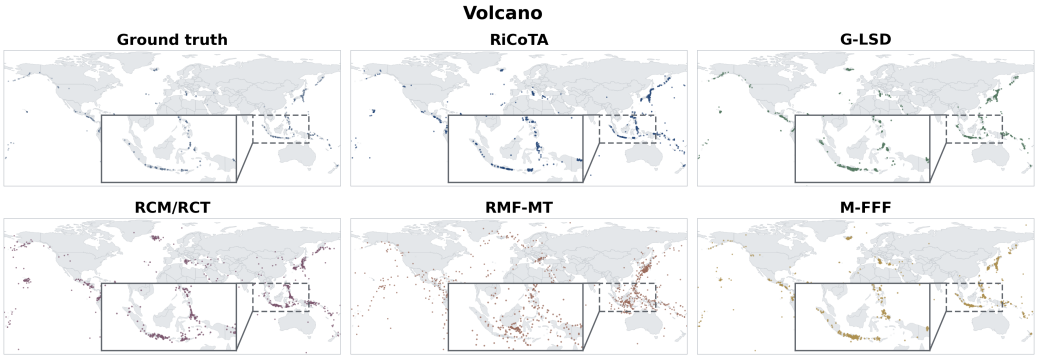


Figure 14: Volcano on $\mathbb{S}^2$. The Indonesia/Philippines inset resolves the volcanic arc and off-arc generated samples.

I.2 PROTEIN TORSION DISTRIBUTIONS

Each Ramachandran figure shows the test set followed by test-count-matched generated samples, without a training-data overlay. The four protein benchmarks complement Table 3.

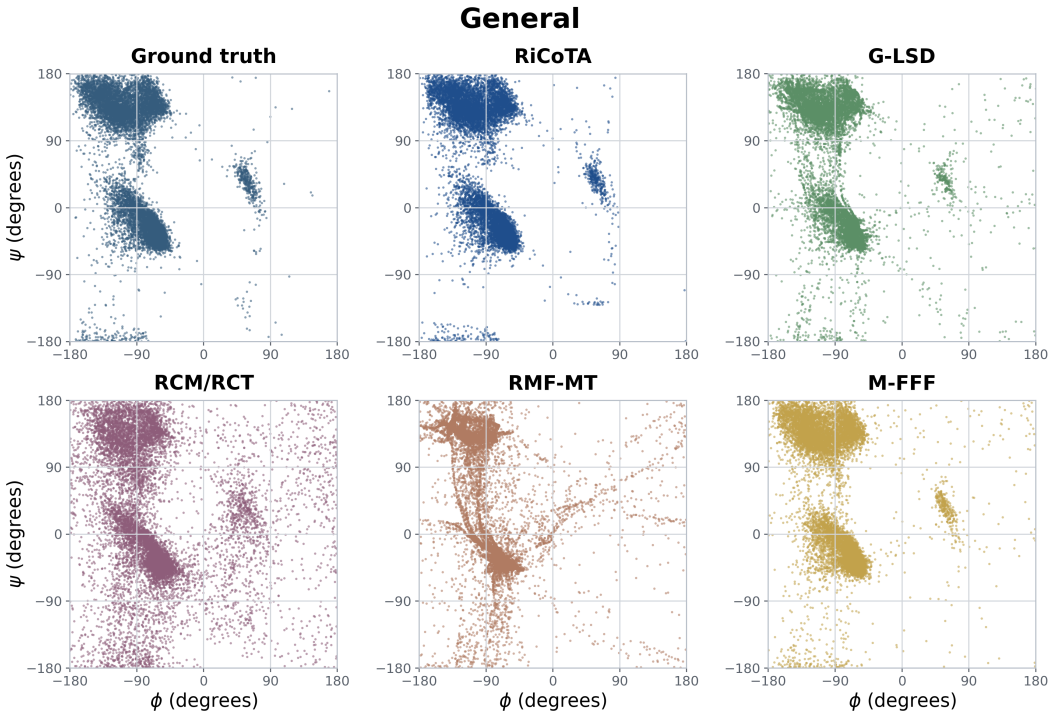


Figure 15: Protein General torsions on $\mathbb{T}^2$. Dominant helical/beta regions and lower-density surrounding structure.

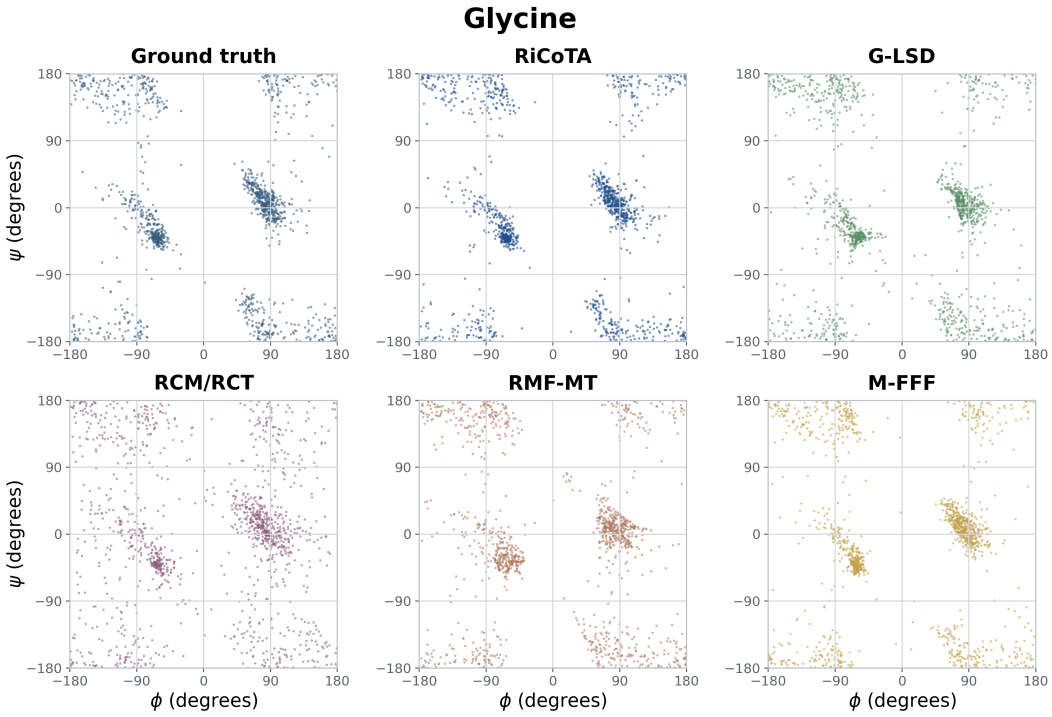


Figure 16: Protein Glycine torsions on $\mathbb{T}^2$. Broad, multimodal structure across the periodic domain.

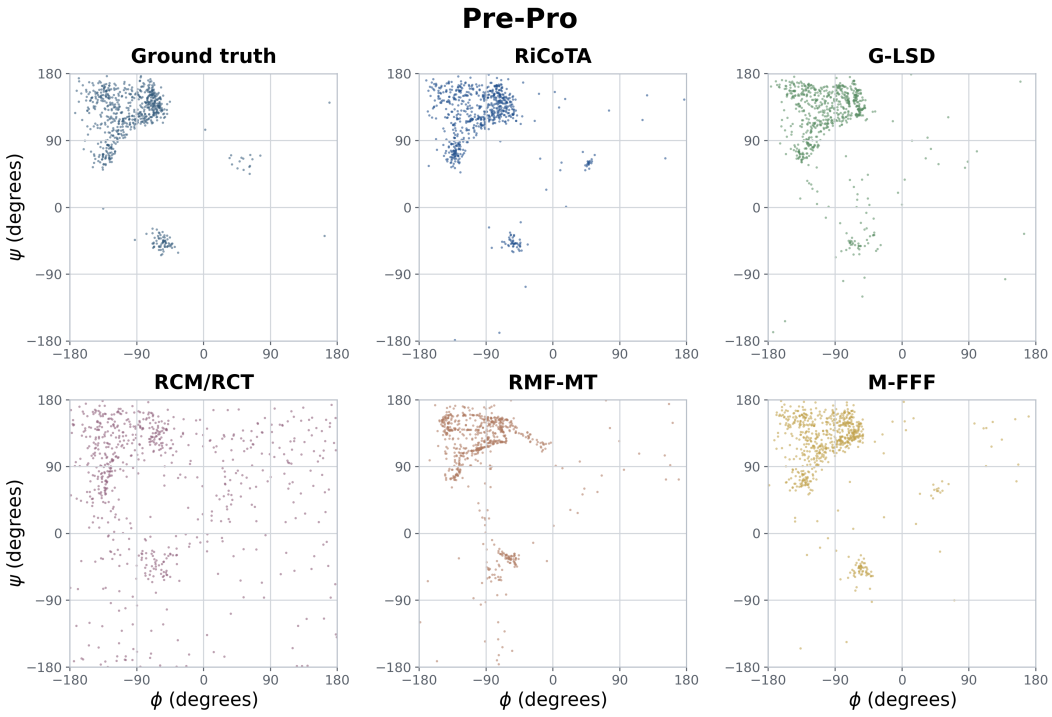


Figure 17: Protein Pre-Pro torsions on $\mathbb{T}^2$. Concentrated target regions compared with the spread of generated samples.

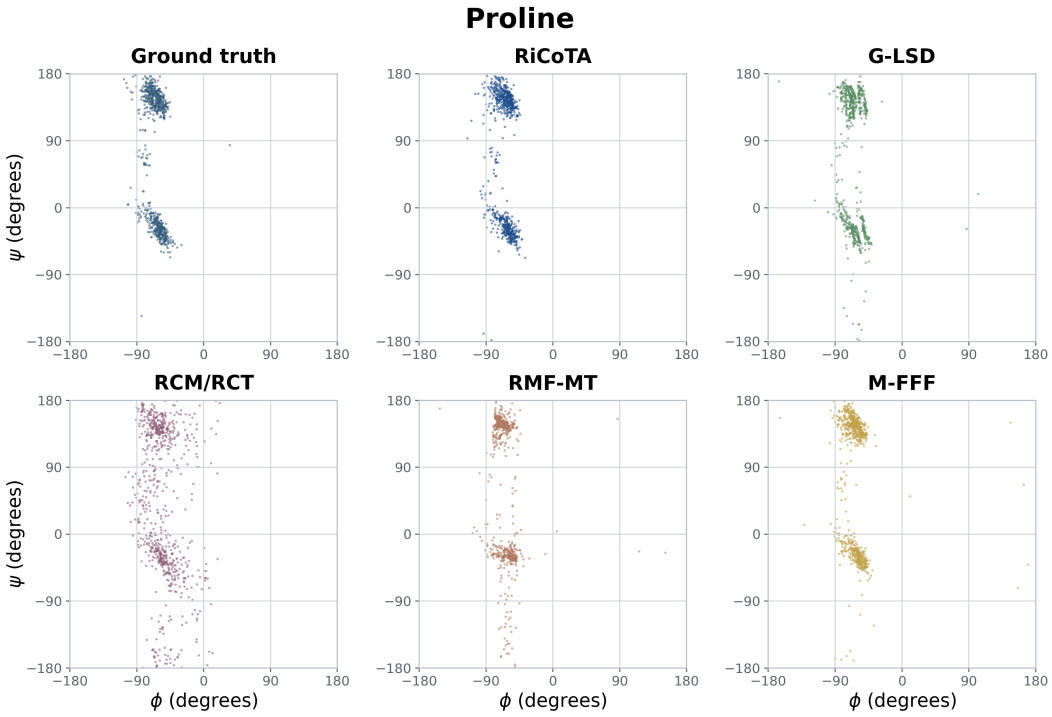


Figure 18: Protein Proline torsions on $\mathbb{T}^2$. Principal ridges and isolated points in the periodic domain.

I.3 ROTATION DISTRIBUTIONS

Each rotation contributes endpoints Re_x, Re_y, and Re_z, colored red, green, and blue. Training observations and train-count-matched generations share the same viewing angle and point styling. These pooled endpoints reveal modal structure but do not identify the full joint rotation law; Table 2 reports five-training-seed quality in the full rotation space.

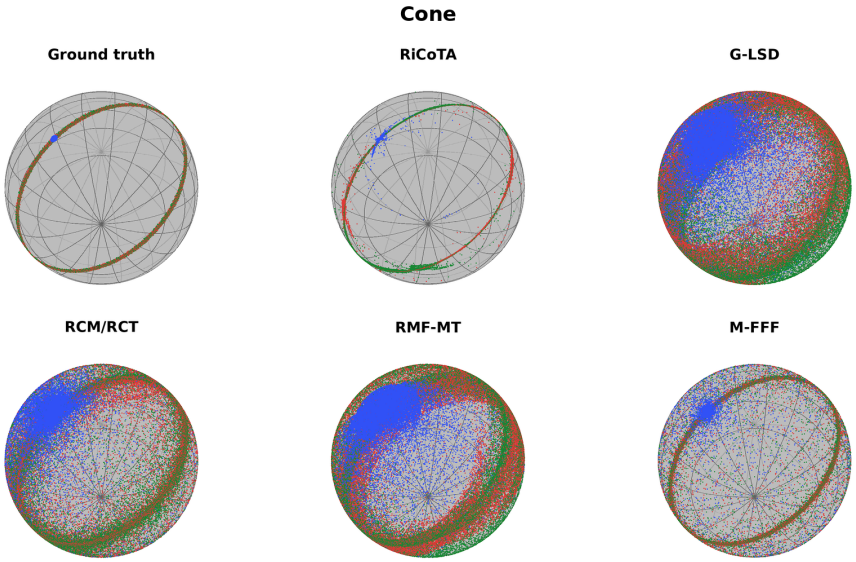


Figure 19: Cone on $\mathrm{SO}(3)$. The narrow endpoint traces and diffuse spherical mass distinguish concentrated from dispersed rotational support.

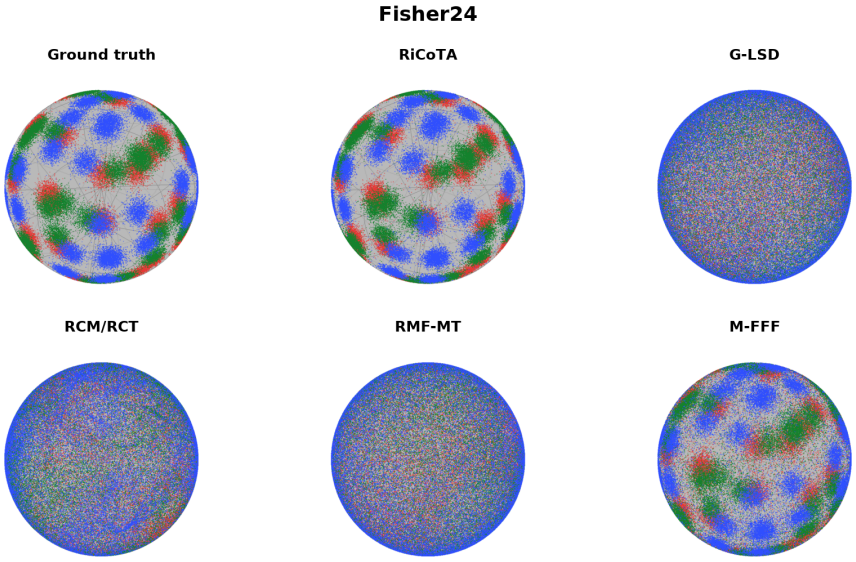


Figure 20: Fisher24 on $\mathrm{SO}(3)$. Separated endpoint clusters visualize the multimodal target structure.

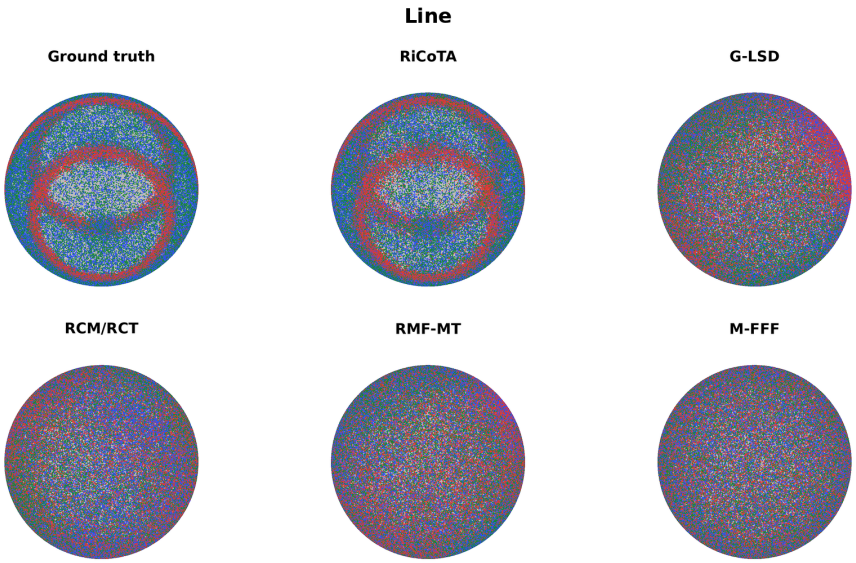


Figure 21: Line on $\mathrm{SO}(3)$. Endpoint bands visualize the elongated rotational support.

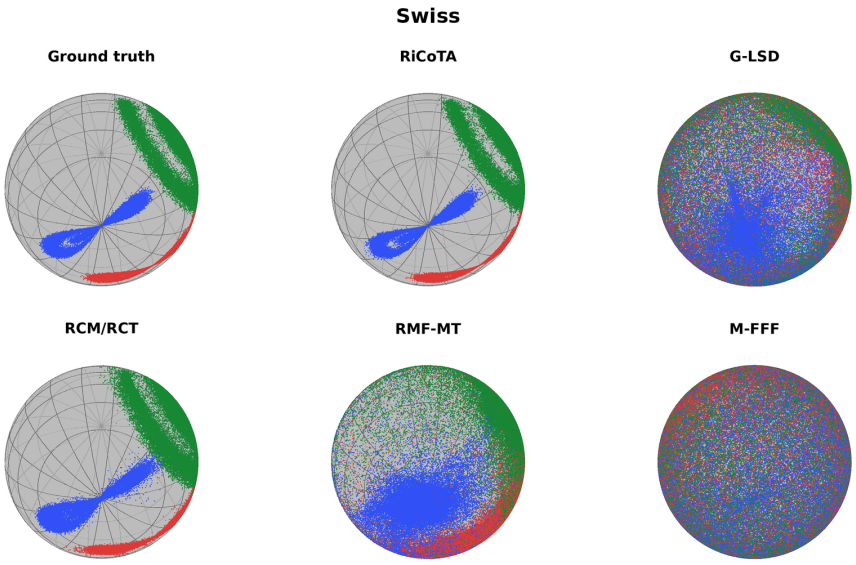


Figure 22: Swiss on $\mathrm{SO}(3)$. Curved endpoint traces visualize the target support and its dispersion across methods.

J CASE STUDY: HIGH-DIMENSIONAL CONDITIONAL PROMOTER GENERATION

This case study explores high-dimensional conditional generation by adapting the encoder–decoder, gauge, and prior while retaining RiCoTA's hard cells and within-cell reconstruction. As in Appendix F.3, detailed architectures, interface settings, training and validation settings, loss coefficients, and calibration budgets are specified in the accompanying DNA configuration.

J.1 TASK, DATA, AND REPRESENTATION

The promoter benchmark (Avdeyev et al., 2023; Stark et al., 2024; Davis et al., 2024; Woo et al., 2026) pairs a length-$L = 1024$ DNA sequence with an aligned cap analysis of gene expression (CAGE) profile $Y^{\mathrm{raw}} \in \mathbb{R}_{\geq 0}^{L\times 1}$. We reserve C for the finite latent code. For nucleotide probabilities $p_\ell \in \Delta^3$,

$$x_\ell = \sqrt{p_\ell} \in \mathbb{S}_+^3 := \{u \in \mathbb{R}^4 : u_a \geq 0,\ \|u\|_2 = 1\}, \qquad X \in \mathcal{M} = (\mathbb{S}_+^3)^{1024}. \tag{113}$$

One-hot nucleotides are boundary points. The relaxation has 3072 intrinsic and 4096 ambient coordinates.

Genomic preparation. We use the first 100,000 FANTOM transcription-start-site rows and CAGE profiles (The FANTOM Consortium and the RIKEN PMI and CLST (DGT), 2014), following the public benchmark's extraction and strand alignment and using its first raw condition channel. The chromosome split contains 88,570 training examples (excluding chr8/chr9/chr10), 3,933 validation examples (chr10), and 7,497 test examples (chr8/chr9).[1] Minus-strand extraction uses the $+1$ coordinate offset, sequence reverse-complement, and reversed CAGE profile.

J.2 CONDITIONING AND SPATIAL-GRID AUTOENCODER

The backend uses two condition representations. With $y_\ell = Y_\ell^{\mathrm{raw}}$,

$$Y_\ell^{\mathrm{AE}} = \frac{\log(1+y_\ell) - \mu_{\log}}{\sigma_{\log}}, \qquad Y_\ell^{\mathrm{P2}} = \left[Y_\ell^{\mathrm{AE}}, \frac{y_\ell - \mu_{\mathrm{raw}}}{\sigma_{\mathrm{raw}}}\right].$$

Train-fitted normalization statistics are frozen: the autoencoder uses Y^{AE} and the prior uses both channels in Y^{P2}, unlike the raw-condition RMF preprocessing. At fixed Y^{AE}, the code map $Q_{\phi,\theta}$ and reconstruction map $\mathrm{Rec}_{\theta,\mathcal{I}}$ follow Appendix B, with the condition also passed to the encoder and decoder:

$$\begin{aligned} Z &= E_\phi(X, Y^{\mathrm{AE}}) \in \mathbb{R}^{m_0}, \qquad C = Q_{\phi,\theta}(X; Y^{\mathrm{AE}}), \\ C^{\mathrm{gen}} &\sim P_\psi^C(\cdot \mid Y^{\mathrm{P2}}), \qquad X^{\mathrm{gen}} = \mathrm{Rec}_{\theta,\mathcal{I}}(C^{\mathrm{gen}}, \varepsilon; Y^{\mathrm{AE}}). \end{aligned} \tag{114}$$

The convolutional encoder combines sequence and condition positionwise and pools features into a $c_z \times G$ spatial latent grid, with $m_0 = c_z G$. The decoder interpolates this grid to sequence length and combines it with aligned condition features and learned position embeddings; residual dilated convolutions produce nucleotide logits $h_\ell(z, Y^{\mathrm{AE}})$. The final support map is

$$D_\theta(z, Y^{\mathrm{AE}})_{\ell,a} = \exp\left[\tfrac{1}{2}\log \mathrm{softmax}_a\, h_\ell(z, Y^{\mathrm{AE}})\right]. \tag{115}$$

Nucleotide readout is $\arg\max_a D_\theta(z, Y^{\mathrm{AE}})_{\ell,a}^2$. No trajectory integration or decoder refinement is used.

J.3 CONDITIONAL GEOMETRY AND INTERFACE CALIBRATION

Under the unit-sphere metric convention, define

$$\begin{aligned} c_\ell &= \mathrm{clip}(x_\ell^\top x'_\ell, -1, 1), \qquad s_\ell = \|x'_\ell - c_\ell x_\ell\|_2, \\ d_{\mathcal{M}}(X, X') &= \left[\sum_{\ell=1}^{L} \mathrm{atan2}(s_\ell, c_\ell)^2\right]^{1/2}, \\ G(z; Y^{\mathrm{AE}}) &= J_z D_\theta(z, Y^{\mathrm{AE}})^\top J_z D_\theta(z, Y^{\mathrm{AE}}). \end{aligned} \tag{116}$$

[1] The text of Woo et al. (2026) reports 88,470 training examples. We retain the executable split count; validation and test counts agree.

Pullback and paired-secant calibration vary the latent coordinates while holding the associated CAGE profile Y_i^{AE} fixed. The latent mean and covariance are computed from training latents. For DNA, we estimate the gauge-bank mean of G using seeded Rademacher vector–Jacobian products, averaging outer products of the resulting latent gradients over probes and paired training examples. Response probes use the same paired latent–condition information. The intrinsic scale in Eq. (19) is estimated from seeded, training-only query and candidate subsamples: we take each query's fifth-neighbor distance within the candidate bank, excluding self-matches, and then the median over queries. Sample and probe budgets are specified in the DNA configuration.

Spectral-pivot gauge completion. For the configured spectral-prefix size r, gauge completion retains all m_0 latent coordinates by preserving the leading r spectral directions and completing the remaining coordinates (the tail) with deterministic reference directions. Let F_{spec} and F_{ref} be the spectral and coordinate-standardized forward matrices fitted on the same training latents. A deterministic rank-revealing row-pivot set $\mathcal{P} = \{p_1, \dots, p_r\}$ of

$$A = F_{\mathrm{ref}}^{-1}[F_{\mathrm{spec}}]_{:,1:r}$$

defines the completion

$$\widetilde{F}_{:,p_i} = [F_{\mathrm{spec}}]_{:,i} \quad (i = 1, \dots, r), \qquad \widetilde{F}_{:,\mathcal{P}^c} = [F_{\mathrm{ref}}]_{:,\mathcal{P}^c}.$$

The completed matrix is used to construct the calibrated coordinate transformation and its inverse. All tail coordinates remain active. The interface combines cube-root response-adaptive spectral-prefix companding with uniform tail cells in probability coordinates. Empirical CDFs, hard tokenization, and exact-cell reconstruction follow Appendix B.

Phase 1 retains the objective, one-draw reconstruction, and train-only recalibration and validation-selection procedures in Appendix D.1.

J.4 Spectral-prefix conditional prior and training objective

In this subsection, Y abbreviates the prior input Y^{P2} and $W_{Y,\mathrm{P2}}$ is the prior's condition-to-logit map. The prior combines a global CAGE summary $g(Y)$ and position-aligned dilated features $S(Y)$ into contexts

$$h_j(Y) = \mathcal{F}\big([g(Y), \overline{S}_j(Y)]\big)\,.$$

The alignment $\overline{S}_j$ follows the flattened latent-grid layout and the frozen coordinate order. Let $C_i^{\mathrm{pre}} = C_{p_i}$ denote the spectral tokens in descending eigenvalue order. Partition the spatial grid into disjoint contiguous blocks and remove the pivot coordinates from each to obtain the ordered tails $\mathcal{T}_b = (j_{b,1}, \dots, j_{b,n_b})$. The conditional code law is

$$\begin{aligned} P_\psi^C(C \mid Y^{\mathrm{P2}}) &= \prod_{i=1}^{r} p_\psi(C_i^{\mathrm{pre}} \mid C_{<i}^{\mathrm{pre}}, Y^{\mathrm{P2}}) \\ &\quad \times \prod_{b} \prod_{t=1}^{n_b} p_\psi(C_{j_{b,t}} \mid C_{j_{b,<t}}, C_{1:r}^{\mathrm{pre}}, Y^{\mathrm{P2}}), \end{aligned} \tag{117}$$

$$p_\psi(C_{j_{b,t}} = k \mid \cdot) = \mathrm{softmax}_k\Big([W_o s_{b,t}]_k + [W_{Y,\mathrm{P2}} h_{j_{b,t}}(Y)]_k + \log \Delta^*_{j_{b,t},k}\Big)\,.$$

Here $s_{b,t}$ is the recurrent state, initialized in each tail block by the spectral prefix. Prefix and tails share the recurrent transition and frozen cell-width bias; tail blocks are generated in parallel conditional on the prefix, with autoregressive sampling within each block. Since $\sum_b n_b = m_0 - r$, each latent token is generated exactly once. Exact-cell dequantization and the inverse interface are followed by one conditional decoder evaluation.

Training augmentation keeps sequence and condition crops aligned: Phase 1 uses uniform integer offsets in $\{-10, \dots, 10\}$, and Phase 2 samples crops at offsets $\{-10, 0, 10\}$ encoded by the frozen Phase 1 model. Validation and test use center crops only. Each model seed independently trains Phase 1 from random initialization through clean warm-up and finite-channel training, then freezes its own encoder, decoder, and interface and trains a freshly initialized prior.

Both prefix and tails use teacher forcing. For a training minibatch $\mathcal{B}$, let $p_{i,j,k}$ be the probability of symbol k under the observed token history specified in Eq. (117) and prior input Y_i^{P2}. Write $Z_{\mathcal{B}} = |\mathcal{B}| m_0 \log K_0$ and

$$\mathcal{L}_{\mathrm{NLL}}(\mathcal{B}) = -Z_{\mathcal{B}}^{-1} \sum_{i \in \mathcal{B}, j} \log p_{i,j,C_{i,j}},$$

$$\mathcal{H}(\mathcal{B}) = -Z_{\mathcal{B}}^{-1} \sum_{i \in \mathcal{B}, j, k} p_{i,j,k} \log p_{i,j,k}.$$

The objective supplements NLL with entropy, conditional marginal-matching, and condition-contrastive terms. Partition the frozen training crop bank by quantiles of a train-only CAGE activity statistic into strata $s_i \in \{1, \dots, S\}$. Let $\mathcal{I}_s$ index all crop-bank entries in stratum s, $\mathcal{B}_s = \{i \in \mathcal{B} : s_i = s\}$, and $\mathcal{S}_{\mathcal{B}} = \{s : |\mathcal{B}_s| > 0\}$. Define the fixed target and minibatch prediction means by

$$\rho^{\mathrm{train}}_{s,j,k} = \frac{1}{|\mathcal{I}_s|} \sum_{i \in \mathcal{I}_s} \mathbf{1}[C_{i,j} = k], \qquad \overline{p}^{\mathcal{B}}_{s,j,k} = \frac{1}{|\mathcal{B}_s|} \sum_{i \in \mathcal{B}_s} p_{i,j,k},$$

$$\mathcal{L}_{\mathrm{marg}}(\mathcal{B}) = \frac{1}{|\mathcal{S}_{\mathcal{B}}| m_0 \log K_0} \sum_{s \in \mathcal{S}_{\mathcal{B}}} \sum_{j} \mathrm{KL}(\rho^{\mathrm{train}}_{s,j} \| \overline{p}^{\mathcal{B}}_{s,j}).$$

Thus the marginal penalty equally weights the strata present in the current minibatch. Let $\ell_i(Y)$ be normalized NLL for example i, $\varrho_{\mathcal{B}}$ a deterministic non-self condition permutation within $\mathcal{B}$, and η_{ctr} the contrastive margin. Then

$$\mathcal{L}_{\mathrm{ctr}}(\mathcal{B}) = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left[\eta_{\mathrm{ctr}} - \left(\ell_i(Y^{\mathrm{P2}}_{\varrho_{\mathcal{B}}(i)}) - \ell_i(Y_i^{\mathrm{P2}}) \right) \right]_+,$$

$$\begin{aligned} \mathcal{L}_{\mathrm{DNA,P2}}(\mathcal{B}) = \mathcal{L}_{\mathrm{NLL}}(\mathcal{B}) &+ \lambda_H \mathcal{H}(\mathcal{B}) \\ &+ \lambda_M \mathcal{L}_{\mathrm{marg}}(\mathcal{B}) + \lambda_C \mathcal{L}_{\mathrm{ctr}}(\mathcal{B}). \end{aligned}$$

A positive entropy penalty favors concentrated teacher-forced predictions. The condition path is first warmed up with normalized cross-entropy on direct logits $W_{Y,\mathrm{P2}} h_j(Y_i)$ alone, without recurrent token history, then jointly optimized with the prefix and tail prior. For each training seed q, DNA Phase 2 selects a checkpoint using validation generation metrics:

$$e^*_q = \arg\min_e \max \left\{ \frac{\mathrm{MSE}_{\mathrm{val},q,e}}{\tau_M}, \frac{1 - \rho_{6,\mathrm{val},q,e}}{\tau_6} \right\}.$$

Here ρ_6 is 6-mer correlation, and the positive normalizers τ_M, τ_6 are fixed across seeds. Test evaluation uses the selected checkpoint.

J.5 Sequence evaluation and generation time

Sei MSE (Stark et al., 2024; Davis et al., 2024) and 6-mer correlation (Woo et al., 2026) measure complementary aspects of sequence quality. The non-strand-specific Sei predictor (Chen et al., 2022) evaluates generated/reference pairs under the same condition. Each sequence is padded by 1,536 uniform-probability bases on each side to length 4,096, and forward/reverse-complement predictions are averaged. For each generated/reference pair, we first average the 2,350 H3K4me3 outputs separately for each sequence, then square the difference between the two means. Sei MSE averages these squared differences over test examples. It measures discrepancy in the mean predicted H3K4me3 response, not nucleotide reconstruction error or experimentally measured activity. For 6-mer correlation, counts from all overlapping length-six windows are pooled across the complete generated bank and, separately, across the complete reference bank. Each count vector is normalized by its own total. Pearson correlation is computed over the union of 6-mers observed in either bank, not averaged over sequencewise correlations.

Matched-batch timing. We measured the wall-clock time to generate all 7,497 test DNA sequences in 58 batches of 128 and one batch of 73. Timing included per-batch condition transfer to the GPU, each method's native sampling and output conversion, transfer of final nucleotide labels to CPU,

and concatenation of the complete output. Model loading, compilation, and metric evaluation were excluded. We used the hardware in Appendix F.6, FP32 with TF32 disabled, and performed 100 full-test repetitions after 20 warmups. RiCoTA ran in PyTorch eager mode; the RMF baselines used their native JAX implementation with the inner network JIT-compiled. Table 27 reports quality and generation time as means and standard deviations.

Table 27: Conditional DNA case study. Quality is reported as mean $\pm$ standard deviation across three training seeds for each method and variant. Timing summarizes 100 full-test repetitions, in seconds for 7,497 sequences. RMF uses v-prediction. RiCoTA Full timing is reported from a previously completed sampler measurement; Identity timing is not reported.

Method	Sei MSE↓	6-mer corr.↑	Time mean $\pm$ SD
Eulerian RMF	0.03345 ± 0.00191	0.95874 ± 0.02043	23.505 ± 0.058
Lagrangian RMF	0.03200 ± 0.00434	0.84367 ± 0.05222	23.438 ± 0.073
Semigroup RMF	0.03364 ± 0.00217	0.82735 ± 0.11958	23.412 ± 0.175
RiCoTA Full	$\mathbf{0.031485} \pm 0.000349$	$\mathbf{0.960048} \pm 0.001894$	6.44939 ± 0.01255
RiCoTA Identity	0.032823 ± 0.000311	0.949802 ± 0.002175	—

Full–Identity comparison. Full uses the response-adaptive interface described in Appendix J.3; Identity uses the identity-compander variant defined in Appendix B, yielding uniform cells in probability coordinates. In Table 27, Full reduces mean Sei MSE from 0.032823 to 0.031485 (4.08%) and increases mean 6-mer correlation from 0.949802 to 0.960048. Both mean quality metrics favor response-adaptive allocation within this conditional model, complementing the geometric-benchmark comparisons.