DEEP TRANSPORTED GAUSSIAN MODELS FOR THE CORRELATION MATRICES

Anonymous authors
Paper under double-blind review

ABSTRACT

Covariance matrices are widely used in representation learning, supported by geometric methods and statistical modeling on the symmetric positive-definite (SPD) manifold. Correlation matrices provide a scale-invariant, standardized representation of statistical dependence, yet probabilistic modeling on the correlation manifold remains comparatively underexplored. We introduce CorTG, a deep transported Gaussian model for full-rank correlation matrices. CorTG combines the off-log map with a learnable invertible transformation and models the transformed data using anisotropic Gaussian distributions. This construction yields an exactly normalized probability density, enables tractable likelihood evaluation through the Jacobian determinant, and supports sampling through the inverse transformation. The induced Riemannian metric also enables efficient computation of distances, geodesics, and Fréchet means. We further derive spectral bounds for the Jacobian of the off-log map and its inverse, analyze the expressive capacity of structured Gaussian models, and establish conditions for geometric identifiability. Experiments on synthetic data and five public benchmarks across multiple domains demonstrate the effectiveness of CorTG.

1 INTRODUCTION

Covariance matrices have been extensively studied through Riemannian learning on the symmetric positive-definite (SPD) manifold, from the foundational affine-invariant metric (AIM) (Pennec et al., 2006) and log-Euclidean geometry (Arsigny et al., 2007) to neural architectures for applications in vision and EEG decoding (Huang & Van Gool, 2017; Brooks et al., 2019; Pan et al., 2022; Kobler et al., 2022; Chen et al., 2023; Wang et al., 2023; 2025; Li et al., 2025; Collas et al., 2025; Hu et al., 2026). Correlation matrices provide a standardized alternative that removes scales while preserving pairwise dependence (Archakov & Hansen, 2021; Archakov et al., 2024), yet remain comparatively underexplored. Recent work has developed dedicated geometries for correlation matrices, notably the Off-Log Metric (OLM) and Log-Scaled Metric (LSM), which are permutation invariant and admit closed-form geometric operations (Thanwerdas, 2024). These developments have begun to enable geometric learning directly on correlation matrices (Chen et al., 2026).

Gaussian distributions provide a natural template for probabilistic modeling, with means and covariances characterizing location and dispersion while supporting likelihood evaluation and sampling (Murphy, 2022). Extending this framework to correlation-valued observations requires respecting the geometry of the correlation manifold. Existing geometric Gaussian models have primarily been developed on SPD manifolds, where Riemannian Gaussians define radial densities through geodesic distance (Said et al., 2017), while anisotropic wrapped Gaussians transport Euclidean Gaussian distributions from tangent spaces to the manifold through exponential maps (De Surrel et al., 2025; Mallasto & Feragen, 2018). Related constructions have also been developed across other non-Euclidean spaces (Nagano et al., 2019; Mallasto et al., 2019). However, probabilistic Gaussian modeling tailored to correlation matrices remains comparatively underexplored.

The OLM (Archakov & Hansen, 2021; Thanwerdas, 2024) defines a permutation-invariant Riemannian metric on full-rank correlation matrices through the off-log diffeomorphism, with fundamental Riemannian operations. This structure also provides a natural setting for constructing Gaussian models on correlation matrices. However, this diffeomorphism is fixed, so the induced geometry is prescribed a priori rather than adapted to the observed data. We therefore compose the off-log map

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.

with a learnable invertible transformation, allowing the induced distribution to adapt jointly to the data while preserving probabilistic modeling.

We introduce CorTG, a deep transported Gaussian model for full-rank correlation matrices. CorTG applies the off-log map followed by a shared learnable invertible transformation. In the transformed space, each class is modeled by an anisotropic Gaussian with a diagonal-plus-low-rank (DPLR) covariance. We parameterize the transformation using a sequence of invertible blocks with tractable Jacobian determinants (Dinh et al., 2017; Papamakarios et al., 2021). This construction yields an exactly normalized probability density, enables tractable likelihood evaluation through the Jacobian determinant, and supports sampling through the inverse transformation. The same mapping induces a complete Riemannian metric on the correlation manifold. With respect to the corresponding Riemannian volume, each class-conditional density takes an anisotropic Gaussian form in the transformed coordinates. This connects flexible probabilistic modeling with efficient computation of distances, geodesics, Fréchet means, and probability regions. We evaluate CorTG on synthetic data and five public benchmarks. Our main contributions are summarized as follows:

- **Deep transported Gaussian modeling.** We introduce CorTG, a probabilistic model for full-rank correlation matrices that combines anisotropic Gaussian distributions with a learned Riemannian metric. We further extend it to strictly causal temporal modeling and joint modeling of marginal scales and correlations.
- **Closed-form expressions and efficient computation.** We derive explicit expressions for probability densities, distances, geodesics, Fréchet means, and probability regions under the proposed model. We also develop numerical procedures for inversion and differentiation, supported by spectral bounds on the relevant Jacobians.
- **Expressivity and identifiability.** We characterize when a shared affine transformation maps multiclass covariance families into diagonal or DPLR form, and derive a rank lower bound based on matrix commutators. With the class-specific Gaussian parameters fixed, we characterize transformations that preserve the likelihood and establish conditions under which they also preserve the induced Riemannian metric.

2 PRELIMINARIES

In this section, we briefly review correlation matrices and the OLM. Further in-depth discussions on correlation geometry can be found in David & Gu (2019); Thanwerdas (2024).

Correlation matrices. Let $X = (X_1, \dots, X_d)^\top$ be a random vector with finite second moments and a positive-definite covariance matrix $S = (\mathrm{Cov}(X_i, X_j))_{1 \le i,j \le d}$. The corresponding correlation matrix is obtained by normalizing the covariance matrix using the standard deviations

$$C = \mathrm{Corr}(S) = D^{-1/2} S D^{-1/2}, \quad D = \mathrm{Diag}(S_{11}, \dots, S_{dd}). \tag{1}$$

Here, $\mathrm{Diag}(\cdot)$ forms a diagonal matrix from the entries. This normalization is scale-invariant, *i.e.*, $\mathrm{Corr}(ASA) = \mathrm{Corr}(S)$ for any positive diagonal matrix A. The resulting matrix is positive definite with unit diagonal and therefore belongs to the manifold of full-rank correlation matrices,

$$\mathcal{C}_d = \{C \in \mathbb{R}^{d \times d} : C = C^\top,\ C \succ 0,\ \mathrm{diag}(C) = \mathbf{1}\}. \tag{2}$$

Here, diag extracts the diagonal entries as a vector, and $\mathcal{C}_d$ has dimension $m = d(d-1)/2$.

OLM. Several Riemannian metrics have been developed on $\mathcal{C}_d$ (David & Gu, 2019; Thanwerdas, 2024). Among these, we adopt the OLM for its permutation invariance and tractable geometric operations. The OLM is obtained by the global diffeomorphism $\phi : \mathcal{C}_d \to \mathbb{R}^m$, defined by

$$h = \phi(C) = \mathrm{hvec}(\mathrm{off}(\log C)). \tag{3}$$

Here, $\log(\cdot)$ denotes the matrix logarithm, and $\mathrm{off}(\cdot)$ sets the diagonal to zero. The operator $\mathrm{hvec}(\cdot)$ stacks the strictly lower-triangular entries and scales them by $\sqrt{2}$, preserving the Frobenius norm on symmetric matrices with zero diagonal. Its inverse $\mathrm{hvec}^{-1}(\cdot)$ reconstructs a symmetric matrix with zero diagonal. For any $h \in \mathbb{R}^m$, the inverse map is given by

$$\phi^{-1}(h) = \exp\big(\mathrm{Diag}(a^\star(h)) + \mathrm{hvec}^{-1}(h)\big), \tag{4}$$

where $\exp(\cdot)$ denotes the matrix exponential and $a^\star(h)$ is the unique solution to the strictly convex minimization problem

$$a^\star(h) = \arg\min_{a \in \mathbb{R}^d} \left\{ \operatorname{tr}\left[\exp\big(\operatorname{Diag}(a) + \operatorname{hvec}^{-1}(h)\big)\right] - \mathbf{1}^\top a \right\}. \tag{5}$$

The matrix exponential guarantees positive definiteness, and the first-order optimality condition ensures that the diagonal correction $\operatorname{Diag}(a^\star(h))$ yields a unit diagonal. Thus, $\phi^{-1}(h) \in \mathcal{C}_d$ for every $h \in \mathbb{R}^m$. Numerical inversion and implicit differentiation are detailed in Appendix C.2.

3 Methodology

CorTG combines the off-log map ϕ with a learnable invertible transformation and class-conditional Gaussian distributions to model probability distributions on $\mathcal{C}_d$. The same transformation supports likelihood evaluation and sampling and induces a Riemannian metric for geometric computations.

3.1 Invertible Transformation on the Correlation Manifold

Learnable transformation. The off-log map provides an unconstrained representation of correlation matrices, but it does not guarantee that the transformed class distributions are Gaussian. We therefore learn a smooth global diffeomorphism $T_\theta : \mathbb{R}^m \to \mathbb{R}^m$, shared across classes, and compose it with ϕ from Section 2. The resulting transformation is defined by

$$h = \phi(C), \quad z = \Psi_\theta(C) = T_\theta(h). \tag{6}$$

Both maps are global diffeomorphisms, so the inverse is given by $\Psi_\theta^{-1} = \phi^{-1} \circ T_\theta^{-1}$.

Affine coupling. We parameterize T_θ using invertible linear mixing and affine coupling blocks (Dinh et al., 2017). In block ℓ, an invertible matrix L_ℓ first mixes the input x_ℓ as $u_\ell = L_\ell x_\ell$. For an index partition (A_ℓ, B_ℓ) that alternates across blocks, the coupling update is defined by

$$v_{\ell,A_\ell} = u_{\ell,A_\ell}, \quad v_{\ell,B_\ell} = u_{\ell,B_\ell} \odot \exp\big(s_\ell(u_{\ell,A_\ell})\big) + t_\ell(u_{\ell,A_\ell}) \tag{7}$$

We bound the log-scale function by setting $s_\ell = s_{\max} \tanh \widetilde{s}_\ell$, with $s_{\max} > 0$, and use smooth sigmoid linear unit (SiLU) activations in both conditioners (Elfwing et al., 2018). Each L_ℓ is parameterized by a lower–upper (LU) factorization with signed exponential diagonal entries (Kingma & Dhariwal, 2018). A product of Householder reflections provides a less expensive alternative in high dimensions (Tomczak & Welling, 2016). The coupling Jacobians are triangular, yielding the log-determinant of the complete transformation as

$$\log|\det JT_\theta(h)| = \sum_\ell \left[\log|\det L_\ell| + \mathbf{1}^\top s_\ell(u_{\ell,A_\ell})\right] + \kappa_\theta, \tag{8}$$

where u_ℓ is evaluated along the transformation of $h = \phi(C)$. The term κ_θ accounts for an optional final affine normalization and is zero when that layer is omitted.

Induced geometry. The map Ψ_θ also induces a Riemannian metric on $\mathcal{C}_d$ as the pullback Euclidean metric g_{Rm} through Ψ_θ. For tangent vectors $U, V \in T_C\mathcal{C}_d$, this metric is given by $g_C^\theta(U,V) = \langle D\Psi_{\theta,C}[U], D\Psi_{\theta,C}[V]\rangle_2$. By construction, Ψ_θ is a global isometry to $\mathbb{R}^m$, so $(\mathcal{C}_d, g^\theta)$ is complete and flat, with distance $d_\theta(C,Q) = \|\Psi_\theta(C) - \Psi_\theta(Q)\|_2$. Geodesics and weighted Fréchet means are obtained by applying Ψ_θ^{-1} to their Euclidean counterparts. Appendix C.3 gives these expressions, together with the exponential and logarithm maps and parallel transport.

3.2 Gaussian Modeling

Density and sampling. We define densities with respect to the original OLM volume, so likelihoods remain comparable as the transformation changes during training. For $U, V \in T_C\mathcal{C}_d$, the reference metric is given by $g_C^0(U,V) = \langle D\phi_C[U], D\phi_C[V]\rangle_2$. Its volume corresponds to Lebesgue measure under $h = \phi(C)$, so $\mathrm{dvol}_{g^0} = \mathrm{d}h$. For class k, we model $z = \Psi_\theta(C)$ by an anisotropic Gaussian with mean $\mu_k \in \mathbb{R}^m$ and a diagonal-plus-low-rank (DPLR) covariance defined by

$$\Sigma_k = \operatorname{Diag}\big(\operatorname{softplus}(a_k) + \delta\mathbf{1}\big) + U_k U_k^\top, \tag{9}$$

where $a_k \in \mathbb{R}^m$, $U_k \in \mathbb{R}^{m\times r}$, and $\delta > 0$. This parameterization ensures positive definiteness and uses $O(m(r+1))$ parameters. The diagonal term represents component-specific variation, while the low-rank term models dependence among the components of z. Let Θ collect all learnable parameters. The change-of-variables formula gives the class-conditional density as

$$q_\Theta(C \mid k) = \mathcal{N}\big(\Psi_\theta(C) \mid \mu_k, \Sigma_k\big)\,|\det JT_\theta(\phi(C))|. \tag{10}$$

This density integrates to one with respect to vol_{g^0}. We sample a correlation matrix by drawing z from the Gaussian and applying the inverse transformation:

$$z \sim \mathcal{N}(\mu_k, \Sigma_k), \quad C = \phi^{-1}\big(T_\theta^{-1}(z)\big). \tag{11}$$

We evaluate ϕ^{-1} by solving for the diagonal correction from Section 2 using damped Newton iterations (Boyd & Vandenberghe, 2004; Higham, 2008). We compute gradients by implicit differentiation at the converged solution (Blondel et al., 2022), as detailed in Appendix C.2.

The Gaussian parameters also define class representatives and probability regions on $\mathcal{C}_d$ under the learned metric. Its volume satisfies $\mathrm{dvol}_{g^\theta} = |\det JT_\theta(\phi(C))|\,\mathrm{dvol}_{g^0}$. For $q_k^{(0)} = q_\Theta(\cdot|k)$, the corresponding density relative to vol_{g^θ} is given by $q_k^{(\theta)}(C) = \mathcal{N}(\Psi_\theta(C)|\mu_k, \Sigma_k)$. Under g^θ, the unique Fréchet mean is $M_k = \Psi_\theta^{-1}(\mu_k)$. The squared Mahalanobis radius is defined as

$$R_k^2(C) = (\Psi_\theta(C) - \mu_k)^\top \Sigma_k^{-1}(\Psi_\theta(C) - \mu_k). \tag{12}$$

The energy $R_k^2(C)/2$ is strongly geodesically convex and attains its minimum at M_k. Under the fitted class-conditional model, $R_k^2(C)$ follows χ_m^2, whose $1-\alpha$ quantile therefore defines a region with model probability $1-\alpha$. Appendix C.4 provides further details.

Temporal modeling. For a sequence $C_{1:T}$, the distribution of each correlation matrix may depend on earlier observations. We share T_θ across time and allow the Gaussian parameters to depend on the transformed history. With $h_t = \phi(C_t)$ and $z_t = T_\theta(h_t)$, a right-shifted causal conditioner computes $c_t = \mathcal{A}_\omega(\mathrm{BOS}, z_{<t})$, initialized by a beginning-of-sequence (BOS) token. Class-specific heads parameterize the conditional mean and covariance as

$$\begin{aligned} &\mu_{t,k} = f_{\mu,k}(c_t), \quad d_{t,k} = \mathrm{softplus}(f_{d,k}(c_t)) + \delta\mathbf{1}, \\ &U_{t,k} = \mathrm{mat}_{m\times r}(f_{U,k}(c_t)), \quad \Sigma_{t,k} = \mathrm{Diag}(d_{t,k}) + U_{t,k}U_{t,k}^\top. \end{aligned} \tag{13}$$

Here, $\mathrm{mat}_{m\times r}$ reshapes its input into the low-rank factor. We denote the resulting parameter map by $H_k(c_t) = (\mu_{t,k}, \Sigma_{t,k})$. Since these parameters depend only on the past and class index, the conditional densities define a normalized joint density. The sequence log-likelihood relative to the product reference volume is given by

$$\log q_\Theta(C_{1:T} \mid k) = \sum_{t=1}^{T} \left[\log \mathcal{N}(z_t \mid \mu_{t,k}, \Sigma_{t,k}) + \log|\det JT_\theta(h_t)|\right]. \tag{14}$$

The conditioner predicts distribution parameters without transforming z_t, so no additional Jacobian term is needed. We sample sequentially from the conditional Gaussians using Eq. (11).

3.3 Learning Objective

We jointly optimize classification and class-conditional likelihood to distinguish classes while fitting the distribution within each class. For either a static observation or a sequence, Bayes' rule gives a class posterior proportional to $\pi_k q_\Theta(\cdot \mid k)$, with $\pi_k \geq 0$ and $\sum_k \pi_k = 1$. Because the transformation is shared across classes, its Jacobian cancels in the posterior but remains in the generative likelihood. We denote the posterior cross-entropy by $\mathcal{L}_{\mathrm{disc}}$ and the class-conditional negative log-likelihood by $\mathcal{L}_{\mathrm{gen}}$. The overall objective combines these losses with two regularizers and is defined as

$$\mathcal{L} = \mathcal{L}_{\mathrm{disc}} + \lambda\mathcal{L}_{\mathrm{gen}} + \beta\mathcal{R}_{\mathrm{scale}} + \gamma\mathcal{R}_{\mathrm{metric}}. \tag{15}$$

The weights λ, β, γ are nonnegative and control the contributions of the generative loss and the two regularizers. The regularizers are evaluated over $\mathcal{D}_{\mathrm{tr}}$, the empirical distribution of $h = \phi(C)$ over training observations, including all time points for sequence data. The scale penalty discourages large coupling log-scales and is defined by

$$\mathcal{R}_{\mathrm{scale}} = \mathbb{E}_{h\sim\mathcal{D}_{\mathrm{tr}}} \sum_\ell \|s_\ell(u_{\ell,A_\ell}(h))\|_2^2. \tag{16}$$

Algorithm 1 The Overview of CorTG

Input: Correlation sequence $C_{1:T}$, class priors $\{\pi_k\}_{k=1}^K$, invertible transport T_θ

for $t = 1, \dots, T$ **do**
  $h_t \leftarrow \phi(C_t)$
  $z_t \leftarrow T_\theta(h_t)$
  $\ell_t \leftarrow \log|\det J_{T_\theta}(h_t)|$
end for
if $T > 1$ **then**
  $c_t \leftarrow \mathcal{A}_\omega(z_{<t})$
else
  $c_t \leftarrow \varnothing$
end if
for $k = 1, \dots, K$ **do**
  $(\mu_{t,k}, \Sigma_{t,k}) \leftarrow H_k(c_t)$
  $g_k \leftarrow \sum_{t=1}^T [\log \mathcal{N}(z_t; \mu_{t,k}, \Sigma_{t,k}) + \ell_t]$
end for
$p_k \leftarrow \mathrm{softmax}_k(\log \pi_k + g_k)$
Training objective: $\mathcal{L} = \mathcal{L}_{\mathrm{disc}} + \lambda \mathcal{L}_{\mathrm{gen}} + \beta \mathcal{R}_{\mathrm{scale}} + \gamma \mathcal{R}_{\mathrm{metric}}$
Inverse sampling: $C^\star = \phi^{-1}\big(T_\theta^{-1}(z^\star)\big)$
Induced metric: $g^\theta = (T_\theta \circ \phi)^* g_{\mathrm{Rm}}$

The optional metric penalty discourages local distortion relative to g^0 by penalizing deviations of $JT_\theta(h)^\top JT_\theta(h)$ from the identity:

$$\mathcal{R}_{\mathrm{metric}} = \mathbb{E}_{h \sim \mathcal{D}_{\mathrm{tr}}} \| JT_\theta(h)^\top JT_\theta(h) - I_m \|_F^2. \tag{17}$$

Algorithm 1 summarizes the framework. Additional manifold-valued layers and detailed training and inference procedures are given in Appendices C.5 and F, respectively.

3.4 Theoretical Analysis

We first compare CorTG with the Gaussian model induced directly by ϕ. The following proposition characterizes their distribution-family inclusion.

Proposition 3.1 (**Distribution-family inclusion**). *Let $\mathcal{Q}$ denote the distribution family in Eq. (10), and let $\mathcal{Q}_0$ denote the family obtained with $T_\theta = \mathrm{Id}$ under the same Gaussian covariance parameterization. If the family of complete transformations contains the identity, then $\mathcal{Q}_0 \subseteq \mathcal{Q}$. Consequently, for every probability measure P on $\mathcal{C}_d$, the population Kullback–Leibler (KL) divergence satisfies*

$$\inf_{Q \in \mathcal{Q}} D_{\mathrm{KL}}(P \| Q) \le \inf_{Q \in \mathcal{Q}_0} D_{\mathrm{KL}}(P \| Q). \tag{18}$$

The density in Eq. (10) is defined relative to the OLM volume. Its expression relative to the independent correlation entries requires the Jacobian of ϕ, whose determinant follows from the existing parameterization (Archakov & Hansen, 2021; Archakov et al., 2024).

Theorem 3.2 (**Jacobian determinant**). *For $C \in \mathcal{C}_d$, let $c = (C_{ij})_{i<j}$ collect its unscaled independent entries and define $J_\phi(C) = \partial \phi(C)/\partial c$. Let $\mathcal{L}_C = D \exp_{\log C}$ act on symmetric matrices and define $A_C \in \mathbb{R}^{d \times d}$ by $(A_C)_{ij} = [\mathcal{L}_C(E_{jj})]_{ii}$, with E_{jj} the matrix whose only nonzero entry is a one at (j, j). Under the $\sqrt{2}$-scaled vectorization in ϕ, the determinant identities are given by*

$$\log |\det J_\phi(C)| = \frac{m}{2} \log 2 + \log \det A_C - \log \det \mathcal{L}_C, \tag{19}$$

$$\det \mathcal{L}_C = \det C \prod_{p<q} \frac{\lambda_p(C) - \lambda_q(C)}{\log \lambda_p(C) - \log \lambda_q(C)}. \tag{20}$$

The determinant of $\mathcal{L}_C$ is evaluated in a Frobenius-orthonormal basis of the space of symmetric matrices, and quotients involving repeated eigenvalues are interpreted by continuity.

The density relative to $\mathrm{d}c$ is therefore $q_k^{(0)}(C)|\det J_\phi(C)|$. For $h = \phi(C)$, the inverse-map residual $r(a, h) = \mathrm{diag}(\exp(\mathrm{Diag}(a) + \mathrm{hvec}^{-1}(h))) - \mathbf{1}$ satisfies $D_a r(a^\star(h), h) = A_C$, so implicit differentiation gives $D_h a^\star(h) = -A_C^{-1} D_h r(a^\star(h), h)$. We further derive spectral bounds for A_C and the constrained Schur complement governing the Jacobian and its inverse in Appendix B.

The correlation model can also retain marginal variances through a separate scale distribution. The following result gives the resulting joint likelihood on covariance matrices.

Theorem 3.3 (Scale–correlation likelihood). *For $S \in \mathbb{S}^d_{++}$, define $v_i = \log S_{ii}$ and $C = \mathrm{Corr}(S)$. The map $S \mapsto (v, C)$ is a diffeomorphism from $\mathbb{S}^d_{++}$ onto $\mathbb{R}^d \times \mathcal{C}_d$. Let $c = (C_{ij})_{i<j}$ and $\mathrm{vech}(S)$ collect the unscaled independent entries of C and S, respectively. The Jacobian determinant satisfies*

$$\log \left| \det \frac{\partial(v, c)}{\partial \,\mathrm{vech}(S)} \right| = -\frac{d+1}{2} \mathbf{1}^\top v. \tag{21}$$

Suppose v and C are conditionally independent given class k, with densities $p_v(v \mid k)$ relative to $\mathrm{d}v$ and $q_k^{(0)}(C)$ relative to vol_{g^0}. The induced density relative to $\mathrm{d}\,\mathrm{vech}(S)$ is then given by

$$\log p_S(S \mid k) = \log p_v(v \mid k) + \log q_k^{(0)}(C) + \log |\det J_\phi(C)| - \frac{d+1}{2} \mathbf{1}^\top v. \tag{22}$$

The same Jacobian correction applies to a normalized conditional density $p_v(v \mid C, k)$ in place of $p_v(v \mid k)$, allowing dependence between scales and correlations. Appendix B gives the corresponding reconstruction, sampling, and product geometry.

We next examine the DPLR covariance family $\mathfrak{S}_r = \{\mathrm{Diag}(v) + UU^\top : v \in \mathbb{R}^m_{++}, \quad U \in \mathbb{R}^{m \times r}\}$ under a shared affine transformation. Any one or two positive-definite covariance matrices admit simultaneous diagonalization by a shared invertible linear transformation. For larger families, this is possible exactly when the covariances commute pairwise after whitening by one of them. Their commutators also yield a lower bound on the required DPLR rank.

Theorem 3.4 (Rank lower bound). *Let $\Sigma_1, \ldots, \Sigma_K \in \mathbb{S}^m_{++}$ and let r be a nonnegative integer. Define $R_{k|a} = \Sigma_k \Sigma_a^{-1}$ for $a, k \in \{1, \ldots, K\}$ and let $[X, Y] = XY - YX$ denote the matrix commutator. The maximum commutator rank is defined as $\kappa_{\mathrm{nc}} = \max_{a,i,j \in \{1,\ldots,K\}} \mathrm{rank}([R_{i|a}, R_{j|a}])$. If a shared $B \in \mathrm{GL}(m)$ satisfies $B\Sigma_k B^\top \in \mathfrak{S}_r$ for every k, then the common rank obeys*

$$r \geq \left\lceil \frac{\kappa_{\mathrm{nc}}}{8} \right\rceil. \tag{23}$$

The likelihood does not generally determine the induced metric uniquely. Further theoretical analysis and normalization details are provided in Appendix B, with proofs in Appendix D.

4 Experiments

We evaluate the effectiveness of CorTG from multiple perspectives. Appendix E reports additional analyses of scale invariance and scale-aware modeling, temporal generation, density and coverage, model capacity, metric-dependent projection, and interpretability.

4.1 Probabilistic Modeling Performance Evaluation

Class-conditional structure. We first examine how different representations reorganize class distributions under a known generative process. Three Gaussian class distributions are transformed by an analytic, globally invertible triangular nonlinearity and then mapped through the inverse off-log chart to 3×3 correlation matrices. The three classes have distinct latent Gaussian means and covariances, with $600/200/600$ independently drawn training/validation/test samples per class. Each method is displayed in its native three-dimensional coordinates. On the prespecified train/test split in Figure 1, CorTG preserves global invertibility while aligning the test point clouds more closely with class-conditional Gaussian probability ellipsoids fitted only on training data, with less inter-class overlap and fewer misclassifications. This example motivates the cross-domain evaluation.

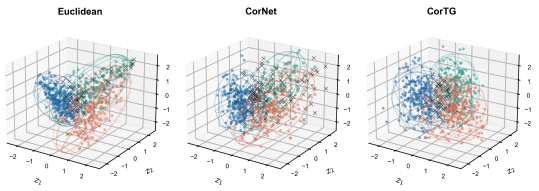


Figure 1: Native 3D class-conditional structure on synthetic datasets. Colored points represent test samples, wireframe ellipsoids denote 90% class-conditional Gaussian probability regions fitted using only training data, and black crosses indicate classifier misclassifications on the test set in each representation space.

Table 1: The results on five public matrix benchmarks (mean ± std over ten seeds). All methods use identical matrix samples and outer splits within each dataset. ECE denotes expected calibration error. Bold indicates the best value in each row, and underlining indicates the second-best.

Dataset	CorNet	AIRM-MDM	TS-LDA	TS-GaussianNB	Ho-WDA	He-WDA	CorTG-fixed	CorTG
(a) Accuracy (%) ↑								
Indian Pines	80.50 ± 0.37	58.01 ± 0.72	67.07 ± 0.53	73.38 ± 0.45	73.64 ± 0.08	74.39 ± 0.20	$\underline{83.74 \pm 0.52}$	$\mathbf{95.84 \pm 0.26}$
Pavia University	$\underline{92.03 \pm 0.11}$	72.32 ± 0.59	84.61 ± 0.06	87.16 ± 0.09	87.32 ± 0.12	85.59 ± 0.14	90.89 ± 0.10	$\mathbf{99.11 \pm 0.03}$
Salinas	79.30 ± 0.05	36.42 ± 0.12	46.30 ± 0.17	69.87 ± 0.21	69.30 ± 0.82	62.38 ± 0.22	$\underline{79.39 \pm 0.28}$	$\mathbf{97.28 \pm 0.15}$
Textile	89.53 ± 0.26	83.08 ± 0.62	83.12 ± 0.63	86.03 ± 0.66	84.76 ± 0.70	86.46 ± 0.64	$\underline{91.26 \pm 0.39}$	$\mathbf{93.88 \pm 0.22}$
BreizhCrops	67.17 ± 0.31	48.29 ± 0.27	50.12 ± 0.27	52.06 ± 0.09	52.73 ± 0.24	32.99 ± 0.98	$\underline{69.40 \pm 0.38}$	$\mathbf{73.25 \pm 0.39}$
(b) Posterior NLL ↓								
Indian Pines	$\underline{0.604 \pm 0.012}$	2.085 ± 0.049	2.471 ± 0.082	1.956 ± 0.058	1.705 ± 0.006	2.029 ± 0.032	0.757 ± 0.042	$\mathbf{0.114 \pm 0.007}$
Pavia University	$\underline{0.213 \pm 0.003}$	2.069 ± 0.014	1.657 ± 0.020	0.913 ± 0.012	0.925 ± 0.019	1.261 ± 0.023	0.323 ± 0.002	$\mathbf{0.028 \pm 0.003}$
Salinas	$\underline{0.547 \pm 0.004}$	8.853 ± 0.103	2.787 ± 0.048	1.415 ± 0.011	1.528 ± 0.013	1.951 ± 0.036	0.861 ± 0.022	$\mathbf{0.076 \pm 0.005}$
Textile	0.269 ± 0.007	0.451 ± 0.004	2.706 ± 0.298	0.721 ± 0.028	0.740 ± 0.031	0.639 ± 0.042	$\underline{0.221 \pm 0.009}$	$\mathbf{0.159 \pm 0.008}$
BreizhCrops	0.766 ± 0.005	1.662 ± 0.000	3.168 ± 0.047	4.406 ± 0.035	4.449 ± 0.067	6.757 ± 0.156	$\underline{0.712 \pm 0.006}$	$\mathbf{0.621 \pm 0.017}$
(c) ECE ↓								
Indian Pines	$\underline{0.058 \pm 0.003}$	0.127 ± 0.007	0.195 ± 0.005	0.201 ± 0.003	0.197 ± 0.001	0.211 ± 0.002	0.112 ± 0.003	$\mathbf{0.009 \pm 0.001}$
Pavia University	$\underline{0.014 \pm 0.001}$	0.171 ± 0.004	0.124 ± 0.001	0.106 ± 0.001	0.098 ± 0.001	0.117 ± 0.002	0.047 ± 0.001	$\mathbf{0.001 \pm 0.000}$
Salinas	$\underline{0.011 \pm 0.002}$	0.537 ± 0.002	0.274 ± 0.001	0.179 ± 0.003	0.156 ± 0.004	0.299 ± 0.002	0.128 ± 0.003	$\mathbf{0.005 \pm 0.001}$
Textile	$\underline{0.053 \pm 0.003}$	0.137 ± 0.006	0.150 ± 0.005	0.110 ± 0.005	0.117 ± 0.005	0.100 ± 0.004	$\mathbf{0.014 \pm 0.002}$	$\mathbf{0.014 \pm 0.002}$
BreizhCrops	0.045 ± 0.003	0.282 ± 0.003	0.349 ± 0.003	0.376 ± 0.001	0.377 ± 0.004	0.571 ± 0.011	$\mathbf{0.009 \pm 0.001}$	$\underline{0.036 \pm 0.005}$

Datasets and evaluation protocols. We evaluate CorTG on five public datasets spanning hyperspectral remote sensing (Indian Pines (Baumgardner et al., 2015), Pavia University, and Salinas (Grupo de Inteligencia Computacional, n.d.)), texture analysis (Textile (Bergmann et al., 2021)), and multispectral crop recognition (BreizhCrops (Rußwurm et al., 2020)). Matrix orders range from 5 to 13, and dataset sizes range from 14,641 to 185,176 samples. Appendix E.3 details the dataset construction and evaluation protocols. For hyperspectral datasets, we reproduce the source PCA-5, local-SCM, and random five-fold pipeline (De Surrel et al., 2025). All methods receive the same matrix samples and outer folds, while validation and checkpoint selection use only the outer training fold. BreizhCrops uses oracle approximating shrinkage (OAS) (Chen et al., 2010).

Baselines and training. Baselines include CorTG-fixed, a full-covariance Gaussian in fixed off-log coordinates, AIRM-MDM (Pennec et al., 2006; Barachant et al., 2010), TS-LDA and tangent-space Gaussian naive Bayes (TS-GaussianNB) (Barachant et al., 2012; Pedregosa et al., 2011), Ho-WDA and He-WDA (De Surrel et al., 2025), and CorNet under five fixed correlation geometries (Chen et al., 2026). Since each observation is a single correlation matrix, CorNet uses its correlation-manifold multinomial logistic regression classifier (CorMLR). Table 1 reports the best CorNet result among five evaluated fixed geometries for each dataset and metric. Appendix Table 7 reports all five variants under the same outer splits and inner early-stopping protocol. CorTG trains for at most 300 epochs, with patience 16 and checkpoint selection by posterior NLL on the inner validation split. Ho-WDA and He-WDA use diagonal latent covariances and Riemannian Conjugate Gradient, capped at 10,000 iterations or 1200 seconds.

Prediction and calibration. Table 1 shows consistent gains from adaptive coordinate learning across the evaluated domains. CorTG achieves the highest accuracy and lowest posterior NLL on all five datasets, averaging 91.87% and 0.200, respectively, and outperforms CorTG-fixed and CorNet on both metrics throughout. Because CorTG-fixed uses the same correlation projection and full-covariance Gaussian classifier but omits the learned transport, the comparison isolates the benefit of coordinate adaptation from that of the quotient representation. CorTG also has the lowest ECE on all three hyperspectral benchmarks. Together, these results under shared data construction and evaluation protocols show that the learned transport provides more accurate and better-calibrated class posteriors than the evaluated fixed-geometry alternatives.

4.2 Effectiveness of Invertible Transport

Controlled comparison. The benchmark suggests that coordinate adaptation can improve probabilistic fit. To isolate its statistical effect, we compare two strictly nested models that share the Gaussian head, initialization, data, loss, and optimization budget, differing only in whether the transport is trained. Data are generated from three rank-2 DPLR Gaussians followed by a shared four-layer in-

Table 2: Test-NLL improvement Δ_{NLL} in the nested transport experiment, reported as the paired mean and 95% bootstrap confidence interval over ten seeds. Positive values favor safe-flow. Here, r is the DPLR rank and γ controls the transport strength.

Samples n	$r = 2$			$r = m$		
	$\gamma = 0$	$\gamma = 1$	$\gamma = 2$	$\gamma = 0$	$\gamma = 1$	$\gamma = 2$
100	0.004 [0.000, 0.011]	0.078 [−0.016, 0.168]	0.923 [0.729, 1.132]	−0.009 [−0.019, 0.000]	0.006 [0.000, 0.014]	0.038 [−0.048, 0.118]
500	0.000 [0.000, 0.000]	0.250 [0.197, 0.309]	1.400 [1.246, 1.546]	−0.002 [−0.006, 0.000]	−0.003 [−0.020, 0.010]	0.101 [0.061, 0.149]
2000	0.000 [0.000, 0.000]	0.275 [0.223, 0.331]	1.482 [1.294, 1.665]	0.000 [0.000, 0.000]	0.011 [0.006, 0.015]	0.187 [0.158, 0.220]

vertible map in global off-log coordinates. We use $d = 5$, $m = 10$, transport strength $\gamma \in \{0, 1, 2\}$, and $n \in \{100, 500, 2000\}$ samples per class, with $60\%/20\%/20\%$ training/validation/test splits and ten paired seeds. We test both a rank-matched DPLR head ($r = 2$) and a full-covariance head ($r = m$). Evaluating $h = \phi(C)$ directly is equivalent to evaluating reconstructed matrices under $\mathrm{dvol}_{g^0} = \mathrm{d}h$, while avoiding numerical error from the inverse chart.

Optimization and model selection. For each condition, an identity-transport model is initialized by moments and cloned before training. The frozen branch continues training only the Gaussian head, whereas the learned-flow branch also trains the transport. The optimizer, learning rate, minibatch order, epoch budget, and early stopping are otherwise identical. Across all 180 seed–condition pairs, their coordinates, log-Jacobians, and validation NLL agree exactly at epoch 0, and both branches pass the same convergence check. For stable finite-sample selection, safe-identity chooses between the initialized and continued identity models by validation NLL, while safe-flow uses the learned transport when it improves validation NLL further. On the test set, we report $\Delta_{\mathrm{NLL}} = \mathrm{NLL}_{\text{safe-identity}} - \mathrm{NLL}_{\text{safe-flow}}$, where positive values favor the learned transport.

Sample size and nonlinearity. Table 2 shows larger gains with more data and stronger nonlinearity. For the rank-matched head at $\gamma = 2$, Δ_{NLL} increases from 0.923 at $n = 100$ to 1.400 at $n = 500$ and 1.482 at $n = 2000$, with improvement in every paired seed. At $\gamma = 1$, the gain increases from 0.078 at $n = 100$ to 0.250 and 0.275 at $n = 500$ and $n = 2000$. One-sided Wilcoxon tests with Holm correction across the 12 positive-strength conditions give $p_{\mathrm{Holm}} = 0.0117$ for every rank-matched setting except $(n, \gamma) = (100, 1)$. In the prespecified primary condition $(r, n, \gamma) = (2, 2000, 2)$, $\Delta_{\mathrm{NLL}} = 1.482$ (95% CI $[1.294, 1.665]$). Without validation fallback, learned-flow improves on the equal-budget continued-identity branch by 1.483 (95% CI $[1.295, 1.666]$, $p_{\mathrm{Holm}} = 0.0117$), confirming that the primary result is stable to the selection rule.

The full-covariance head already captures second-order dependence, and the transport provides additional gains at $n = 2000$: 0.011 for $\gamma = 1$ ($p_{\mathrm{Holm}} = 0.0146$) and 0.187 for $\gamma = 2$ ($p_{\mathrm{Holm}} = 0.0117$). Under the identity-generating condition ($\gamma = 0$), all six comparisons meet the prespecified 0.02 NLL noninferiority margin, with a maximum upper confidence bound on degradation of 0.01945. Across all 180 cases, safe-flow selects learned-flow 112 times and retains the identity candidate 68 times. Thus, the transport improves test likelihood beyond the fixed chart, with the largest gains for the rank-matched head, while validation selection preserves the identity fit when preferable.

Real-data label efficiency. We evaluate label efficiency on Textile and BreizhCrops, the two non-hyperspectral benchmarks. Textile contains two classes of 10×10 image covariance matrices, and BreizhCrops contains six classes of 13×13 multispectral time-series covariance matrices. Statistical intervals use five stratified outer folds as independent units, with ten prespecified optimization seeds per fold. Within each training fold, every class is shuffled once and divided into disjoint optimization and validation pools. Budgets of 2%, 5%, 10%, 20%, 50%, and 100% use nested prefixes of these fixed orders, so smaller training sets are contained in larger ones and all budgets share the same outer test fold. CorNet selects its geometry by validation posterior NLL, whereas CorTG-fixed and TS-LDA fit all labels available at each budget. Figure 2 shows that CorTG maintains strong Macro-F1 and posterior NLL across all six budgets on both datasets. At full budget, CorTG reaches a Macro-F1 of 0.9358 on Textile and 0.8043 on BreizhCrops, with posterior NLL values of 0.1701 and 0.5219, respectively. CorTG also achieves lower NLL than CorTG-fixed on both datasets, indicating that the improvement is not explained solely by a more expressive Gaussian head.

With 50% of the training labels, CorTG reaches Macro-F1/NLL of $0.9221/0.2052$ on Textile and $0.7937/0.5520$ on BreizhCrops, outperforming every comparator at its full budget on the same dataset. On Textile, NLL gains become clear from the 20% budget onward and continue to increase

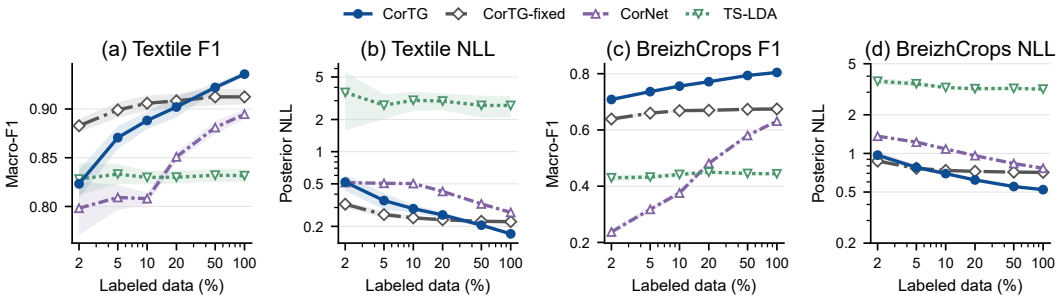


Figure 2: Label efficiency across training-label budgets. (a) Textile Macro-F1, (b) Textile posterior NLL, (c) BreizhCrops Macro-F1, and (d) BreizhCrops posterior NLL. Curves show means over ten seeds per fold; shaded bands denote Student-t 95% confidence intervals across five folds.

with label coverage, while BreizhCrops shows gains from the lowest budget. Thus, greater label coverage consistently strengthens estimation of the nonlinear coordinate map across both real tasks.

4.3 Computational Efficiency Analysis

Asymptotic complexity. Let N be the number of training matrices, d the matrix order, $m = d(d-1)/2$ the off-log dimension, $q = d(d+1)/2$ the SPD tangent dimension, and C the number of classes. CorTG caches the off-log coordinates in $\mathcal{O}(Nd^3)$. Each of its L transport blocks costs $\mathcal{O}(mh + h^2 + m^2)$, while each rank-$r$ DPLR head costs $\mathcal{O}(mr^2 + r^3)$ per sample. Its training and inference costs are therefore $\mathcal{O}\big(Nd^3 + n_{\text{ep}}N[L(mh + h^2 + m^2) + C(mr^2 + r^3)]\big)$ and $\mathcal{O}\big(d^3 + L(mh + h^2 + m^2) + C(mr^2 + r^3)\big)$. CorNet repeats coordinate construction and linear classification over n_{geo} geometries, giving $\mathcal{O}(n_{\text{geo}}Nd^3 + n_{\text{geo}}n_{\text{ep}}NmC)$ training cost (Chen et al., 2026). Ho-WDA and He-WDA repeat samplewise SPD operations and, respectively, one shared or C class-specific covariance factorizations per iteration (De Surrel et al., 2025); exact GP training instead factors an $N \times N$ kernel matrix in $\mathcal{O}(N^3 + mN^2)$ (Rasmussen & Williams, 2006). Complete expressions and their derivations are provided in Appendix Table 13.

Measured runtime. Table 3 compares the complete pipelines under the same four-thread CPU setting and optimization budget. CorTG requires 4.71 seconds end-to-end, compared with 15.06, 11.87, and 12.37 seconds for validation-selected CorNet, Ho-WDA, and He-WDA. It is therefore 3.20×, 2.52×, and 2.63× faster, respectively. The totals include preprocessing, model fitting, and one test-set prediction pass, thereby accounting for CorNet's repeated geometry search and WDA's iterative manifold optimization. CorTG provides the lowest measured end-to-end cost among the learned classifiers at this controlled operating point.

Table 3: End-to-end CPU time at $N = 1024$ and $d = 5$.

Method	E2E (s)	Ratio
CorTG	**4.71**	1.00×
CorNet	15.06	3.20×
Ho-WDA	11.87	2.52×
He-WDA	12.37	2.63×

5 Conclusion

We presented CorTG, a probabilistic framework for modeling full-rank correlation matrices. By combining scale-quotient preprocessing, global off-log coordinates, and a learnable invertible transport, CorTG provides an exactly normalized likelihood under a fixed reference measure while inducing a task-adaptive pullback geometry. The framework supports tractable sampling and intrinsic statistical computations, extends naturally to strictly causal sequence modeling, and supports structured Gaussian heads with explicit theoretical guarantees on invariance, geometry, and expressive capacity. Across multiple public benchmarks, CorTG achieves consistent improvements across evaluation metrics, with particularly strong performance in larger and more nonlinear regimes. Overall, these results suggest that learning coordinates on the correlation manifold provides a principled way to couple density estimation with task-adaptive geometry while retaining exact likelihood evaluation, efficient sampling, and scale invariance, enabling more flexible Gaussian modeling.

AI USE STATEMENT

Generative AI tools were used to polish the manuscript for clarity, search for and identify relevant prior work, and help prepare the paper's overview figure. They were not used to formulate mathematical claims, write or check mathematical proofs, develop theoretical models or conceptual frameworks, propose hypotheses, design the methodology or experiments, implement the proposed method, process datasets, or interpret empirical results. The authors reviewed and independently verified all AI-assisted text, literature-search results, and visual material, and take full responsibility for the final content of this work, including any text, claims, or artifacts produced with AI assistance.

ETHICS STATEMENT

This work only uses publicly available datasets and does not involve human subjects or sensitive information. We identify no specific ethical concerns.

REPRODUCIBILITY STATEMENT

All theoretical results are established under explicit assumptions, with complete proofs provided in the Appendix. The code will be released upon acceptance.

REFERENCES

Bijan Afsari. Sensitivity analysis for the problem of matrix joint diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1148–1171, 2008.

Ilya Archakov and Peter R. Hansen. A new parametrization of correlation matrices. *Econometrica*, 89(4):1699–1715, 2021.

Ilya Archakov, Peter Reinhard Hansen, and Yiyao Luo. A new method for generating random correlation matrices. *The Econometrics Journal*, 27(2):188–212, 2024.

Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 29(1):328–347, 2007.

Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Riemannian geometry applied to BCI classification. In *Latent Variable Analysis and Signal Separation*, volume 6365 of *Lecture Notes in Computer Science*, pp. 629–636. Springer, 2010.

Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Multiclass brain–computer interface classification by Riemannian geometry. *IEEE Transactions on Biomedical Engineering*, 59(4):920–928, 2012.

Marion F. Baumgardner, Larry L. Biehl, and David A. Landgrebe. 220 band AVIRIS hyperspectral image data set: June 12, 1992 Indian Pine test site 3. Purdue University Research Repository, 2015.

Paul Bergmann, Kilian Batzner, Michael Fauser, David Sattlegger, and Carsten Steger. The MVTec anomaly detection dataset: A comprehensive real-world dataset for unsupervised anomaly detection. *International Journal of Computer Vision*, 129(4):1038–1059, 2021.

Mathieu Blondel, Quentin Berthet, Marco Cuturi, Roy Frostig, Stephan Hoyer, Felipe Llinares-López, Fabian Pedregosa, and Jean-Philippe Vert. Efficient and modular implicit differentiation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 5230–5242, 2022.

V. I. Bogachev, A. V. Kolesnikov, and K. V. Medvedev. Triangular transformations of measures. *Sbornik: Mathematics*, 196(3):309–335, 2005.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44(4):375–417, 1991.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Daniel Brooks, Olivier Schwander, Frédéric Barbaresco, Jean-Yves Schneider, and Matthieu Cord. Riemannian batch normalization for SPD neural networks. In *Advances in Neural Information Processing Systems*, volume 32, pp. 15463–15474, 2019.

Hengchao Chen. Quotient geometry of bounded or fixed-rank correlation matrices. *SIAM Journal on Matrix Analysis and Applications*, 46(1):121–150, 2025.

Yilun Chen, Ami Wiesel, Yonina C. Eldar, and Alfred O. Hero. Shrinkage algorithms for MMSE covariance estimation. *IEEE Transactions on Signal Processing*, 58(10):5016–5029, 2010.

Ziheng Chen, Tianyang Xu, Xiao-Jun Wu, Rui Wang, Zhiwu Huang, and Josef Kittler. Riemannian local mechanism for SPD neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7104–7112, 2023.

Ziheng Chen, Xiaojun Wu, Bernhard Schölkopf, and Nicu Sebe. Riemannian networks over full-rank correlation matrices. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1734. Association for Computational Linguistics, 2014.

Antoine Collas, Ce Ju, Nicolas Salvy, and Bertrand Thirion. Riemannian flow matching for brain connectivity matrices via pullback geometry. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Paul David and Weiqing Gu. A Riemannian structure for correlation matrices. *Operators and Matrices*, 13(3):607–627, 2019.

Thibault De Surrel, Fabien Lotte, Sylvain Chevallier, and Florian Yger. Wrapped Gaussian on the manifold of symmetric positive definite matrices. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 12816–12839, 2025.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *International Conference on Learning Representations*, 2017.

Mathias Drton, Bernd Sturmfels, and Seth Sullivant. Algebraic factor analysis: Tetrads, pentads and beyond. *Probability Theory and Related Fields*, 138(3–4):463–493, 2007.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. In *Advances in Neural Information Processing Systems*, volume 32, pp. 7511–7522, 2019.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks*, 107:3–11, 2018.

Luca Falorsi, Pim de Haan, Tim R. Davidson, and Patrick Forré. Reparameterizing distributions on Lie groups. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3244–3253, 2019.

Wilfrid Gangbo and Robert J. McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2):113–161, 1996.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Grupo de Inteligencia Computacional. Hyperspectral remote sensing scenes. University of the Basque Country (UPV/EHU), n.d.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2008. ISBN 978-0-89871-646-7.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Chen Hu, Ziheng Chen, Rui Wang, Yefeng Zheng, and Nicu Sebe. Riemannian high-order pooling for brain foundation models. In *The Fourteenth International Conference on Learning Representations*, 2026.

Zhiwu Huang and Luc Van Gool. A Riemannian network for SPD matrix learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1):2036–2042, 2017.

Vinay Jayaram and Alexandre Barachant. MOABB: Trustworthy algorithm benchmarking for BCIs. *Journal of Neural Engineering*, 15(6):066011, 2018.

Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In *Advances in Neural Information Processing Systems*, volume 31, pp. 10236–10245, 2018.

Herbert Knothe. Contributions to the theory of convex bodies. *Michigan Mathematical Journal*, 4 (1):39–52, 1957.

Reinmar Kobler, Jun-ichiro Hirayama, Qibin Zhao, and Motoaki Kawanabe. SPD domain-specific batch normalization to crack interpretable unsupervised domain adaptation in EEG. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Shanglin Li, Motoaki Kawanabe, and Reinmar J. Kobler. SPDIM: Source-free unsupervised conditional and label shift adaptation in EEG. In *International Conference on Learning Representations*, 2025.

Jie Liang, Jun Zhou, Yuntao Qian, Lian Wen, Xiao Bai, and Yongsheng Gao. On the sampling strategy for evaluation of spectral-spatial methods in hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(2):862–880, 2017.

Anton Mallasto and Aasa Feragen. Wrapped Gaussian process regression on Riemannian manifolds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5580–5588, 2018.

Anton Mallasto, Søren Hauberg, and Aasa Feragen. Probabilistic Riemannian submanifold learning with wrapped Gaussian process latent variable models. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2368–2377. PMLR, 2019.

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

Kevin P. Murphy. *Probabilistic Machine Learning: An Introduction*. MIT Press, Cambridge, MA, 2022. ISBN 978-0-262-04682-4.

Yoshihiro Nagano, Shoichiro Yamaguchi, Yasuhiro Fujita, and Masanori Koyama. A wrapped normal distribution on hyperbolic space for gradient-based learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4693–4702, 2019.

Yue-Ting Pan, Jing-Lun Chou, and Chun-Shu Wei. MAtt: A manifold attention network for EEG decoding. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. In *Advances in Neural Information Processing Systems*, volume 30, pp. 2338–2347, 2017.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

K. R. Parthasarathy. Two remarks on normality preserving borel automorphisms of $\mathbb{R}^n$. *Proceedings of the Indian Academy of Sciences - Mathematical Sciences*, 123(1):75–84, 2013.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85): 2825–2830, 2011.

Xavier Pennec, Pierre Fillard, and Nicholas Ayache. A riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA, 2006. ISBN 978-0-262-18253-9.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1530–1538, 2015.

Murray Rosenblatt. Remarks on a multivariate transformation. *The Annals of Mathematical Statistics*, 23(3):470–472, 1952.

Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

Marc Rußwurm, Charlotte Pelletier, Maximilian Zollner, Sébastien Lefèvre, and Marco Körner. BreizhCrops: A time series dataset for crop type mapping. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B2-2020:1545–1551, 2020.

Salem Said, Lionel Bombrun, Yannick Berthoumieu, and Jonathan H. Manton. Riemannian Gaussian distributions on the space of symmetric positive definite matrices. *IEEE Transactions on Information Theory*, 63(4):2153–2170, 2017.

James Saunderson, Venkat Chandrasekaran, Pablo A. Parrilo, and Alan S. Willsky. Diagonal and low-rank matrix decompositions, correlation matrices, and ellipsoid fitting. *SIAM Journal on Matrix Analysis and Applications*, 33(4):1395–1416, 2012.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328, 2017.

Michael Tangermann, Klaus-Robert Müller, Ad Aertsen, Niels Birbaumer, Christoph Braun, Clemens Brunner, Robert Leeb, Carsten Mehring, Kai J. Miller, Gernot R. Müller-Putz, Guido Nolte, Gert Pfurtscheller, Hubert Preissl, Gerwin Schalk, Alois Schlögl, Carmen Vidaurre, Stephan Waldert, and Benjamin Blankertz. Review of the BCI competition IV. *Frontiers in Neuroscience*, 6:55, 2012.

Yann Thanwerdas. Permutation-invariant Log-Euclidean geometries on full-rank correlation matrices. *SIAM Journal on Matrix Analysis and Applications*, 45(2):930–953, 2024.

Yann Thanwerdas and Xavier Pennec. Theoretically and computationally convenient geometries on full-rank correlation matrices. *SIAM Journal on Matrix Analysis and Applications*, 43(4): 1851–1872, 2022.

Jakub M. Tomczak and Max Welling. Improving variational auto-encoders using Householder flow. arXiv preprint arXiv:1611.09630, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017.

Qilong Wang, Zhaolin Zhang, Mingze Gao, Jiangtao Xie, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Towards a deeper understanding of global covariance pooling in deep learning: An optimization perspective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15802–15819, 2023.

Rui Wang, Shaocheng Jin, Ziheng Chen, Xiaoqing Luo, and Xiao-Jun Wu. Learning to normalize on the SPD manifold under Bures-Wasserstein geometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8289–8298, 2025.

Kisung You. Exact likelihood and sampling for Riemannian Gaussian distributions on correlation matrices. *arXiv preprint arXiv:2608.02593*, 2026.

APPENDIX CONTENTS

A Related Work **17**

B Supplementary Methodological Development **17**

B.1 Scale-Quotient Projection 18
B.2 Base Correlation Manifold: Global Off-Log Coordinates 18
B.3 Learnable Transport Pullback Geometry 19
B.4 Deep Wrapped Flow and Exact Likelihood 19
B.5 Reusable Deep Components on the Correlation Manifold 23
B.6 Unified Generative–Discriminative Heads for Static and Temporal Data 23
B.7 Theoretical Inclusion Relations, Expressivity, and Complexity 23
B.8 Joint Scale–Correlation Extension 25

C Technical Constructions and Implementation Details **25**

C.1 Scale-Equivariant Regularization and the Quotient Guarantee 25
C.2 Off-Log Inversion, Differentiation, and Base Geometry 26
C.3 Closed-Form Operations under the Adaptive Geometry 26
C.4 Intrinsic Centers and Model Probability Regions 27
C.5 Reusable Coordinate-Based Network Layers 27
C.6 Distribution-Family Inclusion and Low-Rank Capacity 28

D Proofs of Formal Results **29**

D.1 Global off-log diffeomorphism and inverse characterization 29
D.2 Spectral bounds for the constrained off-log Schur complement 29
D.3 Off-log volume and shrinkage stability 29
D.4 Closed-form operations under the adaptive geometry 30
D.5 Normalization, sampling, and conditional identifiability under fixed transport . . . 30
D.6 Duality between fixed reference and learned volumes 30
D.7 Intrinsic center and model probability regions 30
D.8 Fréchet center and strong geodesic convexity 30
D.9 Affine absorption identity 30
D.10 Full-covariance absorption for one or two classes 30
D.11 Necessary and sufficient condition for shared multiclass diagonalization 31
D.12 Fixed-head likelihood stabilizer and metric equivalence class 31
D.13 Nonlinear radial twist preserving a standard Gaussian 31
D.14 Multiclass identifiability phase transition for shared-covariance Gaussians 31
D.15 Normalization conditions for the temporal joint density 32
D.16 Distribution-family inclusion and optimal risk 32
D.17 Noncommutative rank obstruction under a shared affine transport 32

D.18 Testable necessary lower bound for a common low-rank order 33
D.19 Robust version of the noncommutative rank obstruction 33
D.20 Global scale–correlation decomposition of SPD matrices and its Jacobian 33

E Supplementary Experiments **33**
E.1 Scale-Quotient Robustness to Positive Diagonal Gains 34
E.2 Task-Level Benefits of Scale-Quotient Modeling 34
E.3 Cross-Domain Benchmark Details 35
E.4 Strictly Causal Temporal Likelihood and Free Rollout 36
E.5 Joint Evaluation of Density, Sampling, and Probability Regions 37
E.6 Single-Factor Ablation 37
E.7 Noncommutative Rank Obstruction and DPLR Capacity Selection 38
E.8 Mechanisms of Invertible Transport and Visual Interpretation of the Learned Geometry 39
E.9 Neurophysiological Evidence in Cross-Session Brain–Computer Interfaces 40
E.10 Metric-Dependent Geodesic Projection on the Correlation Manifold 42
E.11 Computational Complexity Details 43

F Algorithm Pseudocode **44**

A Related Work

SPD statistics and geometric deep learning. Work on SPD learning follows two closely related lines: statistical methods under fixed metrics and neural architectures that preserve positive definiteness. The affine-invariant metric (AIM) gives a complete, nonpositively curved geometry with well-defined geodesics and Fréchet means(Pennec et al., 2006), while the log-Euclidean metric transfers the principal operations to a vector space through the matrix logarithm(Arsigny et al., 2007). These geometries underpin Riemannian means and tangent-space classifiers(Barachant et al., 2012). SPDNet introduced bilinear maps, eigenvalue rectification, and logarithmic layers(Huang & Van Gool, 2017), while Riemannian batch normalization subsequently used barycenters and parallel transport to normalize SPD features(Brooks et al., 2019). More recent work develops SPD-manifold normalization under Bures–Wasserstein geometry for vision and SPD-constrained adaptation for EEG(Wang et al., 2025; Li et al., 2025); flow matching has also been extended to brain-connectivity matrices on SPD and correlation manifolds(Collas et al., 2025). Because these methods operate on covariance matrices, they retain marginal scale, and the discriminative architectures generally do not specify normalized class-conditional densities.

Parameterizations, geometries, and neural networks for correlation matrices. Full-rank correlation matrices form the quotient of the SPD cone under positive diagonal congruence. David and Gu developed a quotient-affine Riemannian structure and corresponding optimization methods(David & Gu, 2019). Archakov and Hansen showed that the off-diagonal entries of the matrix logarithm give a global unconstrained parameterization of the open elliptope(Archakov & Hansen, 2021). Thanwerdas and Pennec used this representation to study tractable Hadamard and flat geometries(Thanwerdas & Pennec, 2022), including permutation-invariant off-log and log-scaled variants(Thanwerdas, 2024). Bounded- and fixed-rank correlation matrices instead lead to stratified quotient spaces(Chen, 2025). CorNet builds classification, fully connected, and convolutional layers under five prescribed correlation geometries(Chen et al., 2026). These methods establish valid representations and fixed geometric operations, while the present work uses the same manifold as the domain of a normalized density with a learned pullback geometry.

Manifold Gaussians and wrapped distributions. Riemannian Gaussians typically use squared geodesic distance as a radial energy. Said et al. derived normalizers, inference procedures, and mixture models for the affine-invariant SPD manifold(Said et al., 2017). Wrapped distributions instead map a tangent-space or Lie-algebra density through an exponential map with the appropriate volume correction, as in hyperbolic space(Nagano et al., 2019) and Lie groups(Falorsi et al., 2019). De Surrel et al. developed anisotropic wrapped Gaussians and maximum-likelihood classifiers for SPD matrices(De Surrel et al., 2025). On the correlation manifold, You (2026) studied isotropic Riemannian Gaussians under a fixed quotient-affine metric and showed that the normalizer can depend on the center from dimension three onward.

Normalizing flows and temporal densities. Normalizing flows obtain flexible densities by composing invertible maps and tracking their Jacobian determinants(Rezende & Mohamed, 2015; Papamakarios et al., 2021). Real Non-Volume Preserving (Real NVP) flows provide efficient finite-step likelihoods and inverse sampling through affine coupling(Dinh et al., 2017), masked autoregressive flows connect invertible density estimation with autoregressive factorization(Papamakarios et al., 2017), and neural spline flows increase coordinate-wise flexibility with monotone rational-quadratic splines(Durkan et al., 2019). Manifold extensions enforce constraints through charts, projections, or differential equations. Riemannian continuous normalizing flows apply on general manifolds but require numerical ordinary differential equation (ODE) solutions and divergence integration(Mathieu & Nickel, 2020). CorTG instead uses a global correlation-matrix chart and a finite-depth Euclidean transport, retaining exact finite-step likelihoods and inverse sampling while allowing strictly historical conditional parameters.

B Supplementary Methodological Development

This appendix develops the formal results and modeling consequences that support Section 3. It follows the CorTG computational path and records theorem and proposition statements, their assumptions, and the connections among geometry, likelihood, and model capacity. Definitions and equations already stated in the main paper are referenced rather than reproduced. Appendix C serves

a distinct role: it gives the corresponding numerical constructions, implementation formulas, and layer specifications. Complete proofs remain collected in Appendix D.

B.1 Scale-Quotient Projection

Appendix C.1 records the standard regularization used to obtain a full-rank correlation matrix, as well as the scale-equivariant pre-projection ridge used when an SPD matrix must be formed first. For $C \in \mathcal{C}_d$, define

$$\mathcal{R}_\varepsilon(C) = (1-\varepsilon)C + \varepsilon I_d, \qquad 0 < \varepsilon < 1. \tag{24}$$

Thus $C_\varepsilon(S) = \mathcal{R}_\varepsilon(\mathrm{Corr}(S))$.

Proposition B.1 (Scale invariance)**.** *For every positive diagonal A, $C_\varepsilon(ASA) = C_\varepsilon(S)$, and $\mathrm{Corr}(S_1) = \mathrm{Corr}(S_2)$ holds exactly when $S_2 = AS_1A$ for some such A.*

Indeed, $D_{ASA} = AD_SA$ gives $\mathrm{Corr}(ASA) = \mathrm{Corr}(S)$. Conversely, if $\mathrm{Corr}(S_1) = \mathrm{Corr}(S_2) = C$, then $S_j = D_{S_j}^{1/2} C D_{S_j}^{1/2}$, and choosing $A = D_{S_2}^{1/2} D_{S_1}^{-1/2}$ yields $S_2 = AS_1A$.

B.2 Base Correlation Manifold: Global Off-Log Coordinates

The off-log chart and its inverse characterization are stated in Section 2. For later use, set $M(a,h) = \mathrm{Diag}(a) + \mathrm{hvec}^{-1}(h)$ and let $a^\star(h)$ denote the unique diagonal correction satisfying $r(a,h) = 0$, where r is defined in Section 3. The numerical inversion and differentiation details are collected in Appendix C.2.

Section 3 defines the parameter-independent metric g^0 and its reference volume. Appendix C.2 records the corresponding closed-form operators and explains why this construction avoids the base-point-dependent exponential-map Jacobian and redundant base point–mean parameterization of an AIRM-SPD wrapped Gaussian (De Surrel et al., 2025).

Differentiable inverse map. Appendix C.2 specifies the matrix-exponential derivative, stopping criterion, and numerical implementation used by the Newton solve and its implicit adjoint.

Stability of the constrained Schur complement. Prior work gives the Jacobian and determinant of the off-log map(Archakov & Hansen, 2021; Archakov et al., 2024). We further characterize the spectral stability after eliminating the unit-diagonal constraint. Let $M = \log C$ and $\mathcal{L}_C = D\exp_M$. Under the Frobenius inner product, use the orthogonal decomposition $\mathrm{Sym}_d = \mathcal{D}_d \oplus \mathbb{S}_d$, where $\mathcal{D}_d$ is the diagonal subspace. In the bases $\{E_{ii}\}_{i=1}^d$ and $\{(E_{ij}+E_{ji})/\sqrt{2}\}_{i<j}$, write $\mathcal{L}_C$ as

$$[\mathcal{L}_C]_{\mathcal{D}_d \oplus \mathbb{S}_d} = \begin{bmatrix} A_C & B_C \\ B_C^\top & D_C \end{bmatrix}, \qquad S_C = D_C - B_C^\top A_C^{-1} B_C. \tag{25}$$

Here $A_C = D_a r(a^\star, h)$ is the Newton Jacobian, and S_C is the constrained Schur complement obtained by eliminating the diagonal variables under $\mathrm{diag}(\exp M) = \mathbf{1}$.

Theorem B.2 (Spectral bounds for the constrained off-log Schur complement)**.** *For any $C \in \mathcal{C}_d$, let $\alpha = \lambda_{\min}(C)$ and $\beta = \lambda_{\max}(C)$. Then*

$$\alpha I_d \preceq A_C \preceq \beta I_d, \qquad \alpha I_m \preceq S_C \preceq \beta I_m. \tag{26}$$

Consequently,

$$\kappa_2(A_C) \le \kappa_2(C), \qquad \kappa_2(S_C) \le \kappa_2(C). \tag{27}$$

If $c = (C_{ij})_{i<j}$ contains the unscaled independent correlations and $h = \phi(C) = \mathrm{hvec}(\mathrm{off}(\log C))$, then the local Jacobians satisfy

$$D_h c = \frac{1}{\sqrt{2}} S_C, \qquad D_c h = \sqrt{2}\, S_C^{-1}, \tag{28}$$

Thus, the forward and inverse off-log coordinate transformations have the same condition number, and neither exceeds $\kappa_2(C)$.

Corollary B.3 (Off-log volume and shrinkage stability)**.** *The Jacobian identities in Eqs. (19)–(20) combine with the constrained Schur-complement bounds above to give the following additional stability consequence. For C_ε produced by Eq. (24), let $\bar{\lambda}_\varepsilon = (1-\varepsilon)d + \varepsilon$. Then*

$$\varepsilon I \preceq A_{C_\varepsilon}, S_{C_\varepsilon} \preceq \bar{\lambda}_\varepsilon I, \qquad \max\{\kappa_2(A_{C_\varepsilon}), \kappa_2(S_{C_\varepsilon})\} \leq \frac{\bar{\lambda}_\varepsilon}{\varepsilon}, \tag{29}$$

and $\|A_{C_\varepsilon}^{-1}\|_2, \|S_{C_\varepsilon}^{-1}\|_2 \leq \varepsilon^{-1}$.

Theorem B.2 simultaneously controls three local linear problems: the Newton step $A_C \Delta a = -r$, the adjoint system $A_C^\top v = u$ in implicit backpropagation, and the forward and inverse coordinate transformations in Eq. (28). Eq. (29) strictly bounds the linearization at the target solution C_ε and ensures continuous stability in a neighborhood of convergence. It does not by itself guarantee that every damped Newton iterate far from the solution obeys the same spectral bound. An iteration-wise guarantee additionally requires backtracking or a trust region that explicitly restricts the spectrum of $\exp M(a, h)$.

B.3 Learnable Transport Pullback Geometry

Section 3 specifies the learned pullback, smooth affine coupling blocks, and their exact log-determinant in Eq. (8). Those definitions are not repeated here. One optional extension is worth recording: Affine coupling can be replaced by monotone rational-quadratic splines to improve the representation of tails and local shape (Durkan et al., 2019). Piecewise C^1 splines suffice for a change-of-variables likelihood. If the transport is also interpreted through a classical smooth Riemannian metric, however, the spline joints must be smooth, or the smooth affine coupling above should be retained.

The corresponding geometric result is stated in Section 3, and Appendix C.3 records only its explicit operators and exact constraint guarantee.

CorTG therefore extends a fixed off-log geometry by learning a geometric representative jointly with a normalized probabilistic model. Its relation to fixed OLM and CorNet, together with the associated gauge qualification, is discussed in Appendix C.3.

B.4 Deep Wrapped Flow and Exact Likelihood

The DPLR Gaussian head, normalized fixed-volume density, and inverse-sampling rule are given once in Eqs. (10)–(11). The remaining results in this subsection clarify normalization, reference measures, and identifiability rather than redefining the model.

Theorem B.4 (Conditional identifiability under fixed transport)**.** *With T_θ fixed, if two parameter pairs (μ_1, Σ_1) and (μ_2, Σ_2) induce the same distribution, then $\mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2$.*

This construction removes the distributional parameter redundancy caused by jointly moving the base point and tangent-space mean in the SPD wrapped Gaussian. The result is conditional on a fixed T_θ and does not imply joint identifiability of the transport and Gaussian head.

Theorem B.5 (Duality between fixed reference and learned volumes)**.** *Let Q_k be the class-conditional probability measure defined by Eq. (10). The fixed- and learned-volume identities stated in Section 3 hold globally. Thus, CorTG is a normalizing flow relative to the fixed reference volume and an exact Gaussian in global isometric coordinates relative to its learned volume.*

This duality gives two representations of the same probability measure. It does not permit the training reference measure to change. As θ varies, dvol_{g^θ} varies as well. Directly comparing learned-volume density values across parameter settings would therefore compare densities defined with respect to different measures. Training, model selection, and cross-geometry NLL reporting must still use the parameter-independent dvol_{g^0} and Eq. (10). The isotropic case depends only on the learned distance and is therefore a Riemannian Gaussian in a purely metric sense. A general Σ_k defines an anisotropic Gaussian in the global parallel orthonormal frame induced by Ψ_θ.

The intrinsic-center and probability-region result is summarized in Section 3, and Appendix C.4 gives its explicit density and region formulas.

Proposition B.6 (Fréchet center and strong geodesic convexity). *For any positive-definite Σ_k, the point $M_k = \Psi_\theta^{-1}(\mu_k)$ is the unique g^θ-Fréchet mean:*

$$M_k = \operatorname*{arg\,min}_{M \in \mathcal{C}_d} \mathbb{E}_{C \sim Q_k}[d_\theta(C, M)^2]. \tag{30}$$

Let the C-dependent energy in the negative log-density relative to the learned volume be

$$E_k(C) = \frac{1}{2}(\Psi_\theta(C) - \mu_k)^\top \Sigma_k^{-1}(\Psi_\theta(C) - \mu_k). \tag{31}$$

Then, for any $C \in \mathcal{C}_d$ and $V \in T_C\mathcal{C}_d$,

$$\operatorname{Hess}^{g^\theta} E_k(C)[V, V] = \left\langle D\Psi_{\theta,C}[V], \Sigma_k^{-1} D\Psi_{\theta,C}[V] \right\rangle, \tag{32}$$

$$\lambda_{\min}(\Sigma_k^{-1}) g_C^\theta(V, V) \le \operatorname{Hess}^{g^\theta} E_k(C)[V, V] \le \lambda_{\max}(\Sigma_k^{-1}) g_C^\theta(V, V). \tag{33}$$

Therefore, E_k is strongly geodesically convex on all of $(\mathcal{C}_d, g^\theta)$ with parameter $\lambda_{\min}(\Sigma_k^{-1})$, and M_k is its unique minimizer.

M_k is also the unique mode of the learned-volume density $q_k^{(\theta)}$, but it need not be the mode of the fixed-volume density $q_k^{(0)}$. If the latter has a differentiable interior maximum at $h^\star \in \mathbb{R}^m$, its necessary first-order condition is

$$JT_\theta(h^\star)^\top \Sigma_k^{-1}\big(T_\theta(h^\star) - \mu_k\big) = \nabla_h \log |\det JT_\theta(h^\star)|. \tag{34}$$

This equation follows by differentiating $\log q_k^{(0)}(\phi^{-1}(h))$. At the preimage $h_k = T_\theta^{-1}(\mu_k)$ of the latent mean, the left-hand side is zero. Unless $\nabla_h \log |\det JT_\theta(h_k)| = 0$, the point $M_k = \phi^{-1}(h_k)$ is not even a stationary point of the fixed-volume density. This distinction prevents conflating the latent Gaussian mean, the Fréchet center under the learned metric, and the mode under the fixed reference measure.

Proposition B.7 (Affine absorption identity). *Let $B \in \mathrm{GL}(m)$ and $b \in \mathbb{R}^m$, and define*

$$T'_\theta = BT_\theta + b, \qquad \mu'_k = B\mu_k + b, \qquad \Sigma'_k = B\Sigma_k B^\top. \tag{35}$$

If the transformed Gaussian head permits covariance Σ'_k, then $q_{\Theta'}(C \mid k) = q_\Theta(C \mid k)$ for every class. However, unless B is orthogonal, the pullback metric generally changes.

This identity applies to arbitrary full covariances, but an arbitrary B need not preserve the DPLR parameterization stated in Section 3. An affine subgroup that preserves this family exactly consists of monomial matrices $B = P\,\mathrm{Diag}(s_1, \ldots, s_m)$ with $s_i \neq 0$, where P is a permutation matrix. The diagonal terms are only permuted and multiplied by s_i^2, while the low-rank term becomes $(BU_k)(BU_k)^\top$. Positive diagonal scaling is therefore not the full affine parameter symmetry of the restricted Gaussian head. Sign flips and coordinate permutations also belong to this gauge.

Corollary B.8 (Full-covariance absorption for one or two classes). *Assume that T_θ contains an arbitrary invertible linear layer. For a single full-covariance Gaussian $\mathcal{N}(h; \mu, \Sigma)$, choosing $B = \Sigma^{-1/2}$ and $b = -B\mu$ makes the transformed Gaussian head $\mathcal{N}(0, I)$. Thus, in the single-class case, CorTG with $r = 0$ already contains any full-covariance base wrapped Gaussian.*

For any two positive-definite class covariances Σ_1, Σ_2, further let

$$A = \Sigma_1^{-1/2}\Sigma_2\Sigma_1^{-1/2} = U\Lambda U^\top, \qquad B = U^\top \Sigma_1^{-1/2}. \tag{36}$$

Then $B\Sigma_1 B^\top = I$ and $B\Sigma_2 B^\top = \Lambda$. Hence, when $K \le 2$, a single shared linear transport and diagonal Gaussian heads ($r = 0$) suffice to represent any specified family of full-covariance Gaussian classes exactly.

Theorem B.9 (Necessary and sufficient condition for shared multiclass diagonalization). *For $K \ge 3$ positive-definite covariances, define the whitened matrices*

$$A_k = \Sigma_1^{-1/2}\Sigma_k\Sigma_1^{-1/2}, \qquad k = 2, \ldots, K. \tag{37}$$

There exists a single $B \in \mathrm{GL}(m)$ such that every $B\Sigma_k B^\top$ is diagonal if and only if

$$A_i A_j = A_j A_i, \qquad \forall i, j \in \{2, \ldots, K\}. \tag{38}$$

In principle, the likelihood identifies a family of equivalent parameters rather than a unique task geometry. To select a stable representative, we apply ZCA gauge normalization after the raw transport $\widetilde{z} = \widetilde{T}_\theta(h)$ using aggregate training-set statistics:

$$T_\theta(h) = \widehat{V}_\theta^{-1/2}\big(\widetilde{T}_\theta(h) - \widehat{m}_\theta\big), \qquad \widehat{m}_\theta = \mathbb{E}_{\mathrm{tr}}[\widetilde{z}], \quad \widehat{V}_\theta = \mathrm{Cov}_{\mathrm{tr}}(\widetilde{z}). \tag{39}$$

When $\widehat{V}_\theta \succ 0$, the normalized coordinates have zero mean and identity covariance under the training distribution. Any subsequent affine reparameterization that preserves both conditions must be orthogonal, and orthogonal transformations do not change the pullback metric. For restricted diagonal-plus-low-rank Gaussian heads, signed permutations remain at least at the parameter level. Eq. (39) is an invertible affine layer, so its $-\frac{1}{2}\log\det\widehat{V}_\theta$ contribution must be included in Eq. (8). After training, we freeze $\widehat{m}_\theta$ and $\widehat{V}_\theta$ so that the coordinate transformation of one test sample does not depend on other samples in the test batch. When the sample size is insufficient to obtain a full-rank $\widehat{V}_\theta$, we only standardize each coordinate and explicitly interpret g^θ as a numerical representative of an equivalence class.

To characterize exactly the ambiguity that remains after ZCA, we first fix the latent Gaussian heads and compare only the transports. For strictly positive smooth densities $p_1, \dots, p_K$, define their common measure-preserving automorphism group as

$$\mathrm{Aut}(p_1, \dots, p_K) = \{G \in \mathrm{Diff}^\infty(\mathbb{R}^m) : p_k(Gz)|\det JG(z)| = p_k(z),\ \forall z, k\}. \tag{40}$$

Theorem B.10 (Fixed-head likelihood stabilizer and metric equivalence class)**.** *Let $T_1, T_2 : \mathbb{R}^m \to \mathbb{R}^m$ be global smooth diffeomorphisms, and let*

$$q_{a,k}(h) = p_k(T_a(h))|\det JT_a(h)|, \qquad a \in \{1, 2\}. \tag{41}$$

Define $G = T_2 \circ T_1^{-1}$. Then

$$q_{1,k} = q_{2,k}\ \forall k \quad \Longleftrightarrow \quad G \in \mathrm{Aut}(p_1, \dots, p_K). \tag{42}$$

Furthermore, if $g^{(a)} = (T_a \circ \phi)^* g_{\mathrm{E}}$, then

$$g^{(1)} = g^{(2)} \quad \Longleftrightarrow \quad G(z) = Oz + b, \qquad O^\top O = I. \tag{43}$$

Consequently, the fixed-head likelihood uniquely determines the task metric if and only if

$$\mathrm{Aut}(p_1, \dots, p_K) \subseteq \mathrm{Isom}(\mathbb{R}^m). \tag{44}$$

More generally, the distinct metrics induced by likelihood-equivalent transports are characterized by the coset, or orbit, space of $\mathrm{Aut}(p_1, \dots, p_K)$ modulo its Euclidean-isometry subgroup. This subgroup need not be normal, so we do not call the resulting space a quotient group.

Theorem B.10 completely characterizes the fixed-head stabilizer, but not the full gauge when all parameters are learnable. If the head densities may also change, any diffeomorphism G for which the transformed densities remain in the chosen head family induces the joint reparameterization

$$T' = G \circ T, \qquad p'_k(z') = p_k(G^{-1}z')|\det JG^{-1}(z')|, \tag{45}$$

while leaving the observation-space likelihood unchanged. Eq. (40) is precisely the stabilizer of Eq. (45) under $p'_k = p_k$. For full Gaussian heads, affine maps give Eq. (35). For DPLR heads, monomial affine transformations preserve the parameter family. Heads with special symmetries may also admit the nonlinear stabilizers below. We therefore report the common automorphisms for fixed heads separately from the joint transport–head gauge.

ZCA fixes only the second-order affine gauge and cannot make a general nonlinear transport unique. The following construction gives the most basic non-isometric stabilizer.

Proposition B.11 (Nonlinear radial twist preserving a standard Gaussian)**.** *Let $m \geq 2$ and write $z = (x, w)$ with $x = (z_1, z_2) \in \mathbb{R}^2$ and $w \in \mathbb{R}^{m-2}$. For any smooth function $a : [0, \infty) \to \mathbb{R}$, define*

$$G_a(z) = \Big(R_{a(\|x\|_2^2)}x, w\Big), \tag{46}$$

where R_α is the two-dimensional rotation matrix. Then G_a is a global smooth diffeomorphism, $G_a^{-1} = G_{-a}$, and

$$\|G_a(z)\|_2 = \|z\|_2, \qquad \det JG_a(z) = 1, \qquad \gamma(G_a(z))|\det JG_a(z)| = \gamma(z), \tag{47}$$

where γ is the standard Gaussian density. Therefore, when the Gaussian head is $\mathcal{N}(0, I)$, T_θ and $G_a \circ T_\theta$ have exactly the same likelihood and sampling distribution but generally induce different pullback metrics.

The counterexample above extends to a complete classification for Gaussian heads with shared covariance. Let

$$p_k = \mathcal{N}(\mu_k, \Sigma), \qquad W = \operatorname{span}\left\{\Sigma^{-1/2}(\mu_k - \mu_1) : k = 2, \dots, K\right\}, \qquad q = m - \dim W. \tag{48}$$

Here q is the codimension, in latent space, of the affine hull of the whitened class means.

Theorem B.12 (Multiclass identifiability phase transition for shared-covariance Gaussians)**.** *Let G be a global smooth diffeomorphism on $\mathbb{R}^m$, and define its whitened conjugate map as*

$$H(x) = \Sigma^{-1/2}\left[G\left(\mu_1 + \Sigma^{1/2}x\right) - \mu_1\right]. \tag{49}$$

Under the orthogonal decomposition $\mathbb{R}^m = W \oplus W^\perp$, write $x = (w, y)$. Then $G \in \mathrm{Aut}(p_1, \dots, p_K)$ if and only if

$$H(w, y) = \big(w, F_w(y)\big), \qquad \gamma_q(F_w(y))\,|\det D_y F_w(y)| = \gamma_q(y), \tag{50}$$

where F_w varies smoothly with w and is, for every w, a standard-Gaussian-measure-preserving diffeomorphism on $\mathbb{R}^q$. Consequently:

1. *If $q = 0$, then $G = \mathrm{Id}$, and both the transport and its induced metric are identifiable.*
2. *If $q = 1$, then in whitened coordinates H can only be the identity or the global reflection about W. The transport has at most two candidates. Whether the nontrivial candidate preserves the metric in the original latent coordinates also depends on the compatibility between this whitened reflection and Σ.*
3. *If $q \geq 2$, the common automorphism group contains an infinite-dimensional nonlinear, non-isometric subfamily. The likelihood therefore cannot uniquely determine the task metric.*

More specifically, let $S_W = I - 2nn^\top$ be the reflection about W in whitened coordinates, where n is the unit basis vector of $W^\perp$. After conjugation back to the original coordinates, its linear part is $A = \Sigma^{1/2} S_W \Sigma^{-1/2}$. By Theorem B.10, when $q = 1$ the metric is identifiable if and only if

$$A^\top A = I \quad \Longleftrightarrow \quad S_W \Sigma S_W = \Sigma \quad \Longleftrightarrow \quad S_W \Sigma = \Sigma S_W. \tag{51}$$

This condition holds automatically when $\Sigma = \sigma^2 I$, but not for a general shared full covariance. The precise phase transition is therefore as follows. The transport is unique when $q = 0$. When $q = 1$, the ambiguity collapses from an infinite-dimensional family to one discrete reflection, and Eq. (51) determines whether the metric is unique. When $q \geq 2$, an infinite-dimensional nonlinear metric ambiguity remains.

Theorem B.12 clarifies what ZCA can and cannot remove. ZCA selects one representative of the joint affine gauge that changes first- and second-order moments. The remaining orthogonal transformations do not change the metric. In contrast, Eq. (97) preserves each class-conditional distribution itself and therefore also preserves the mean and covariance of any class mixture. ZCA can neither detect nor remove this ambiguity. For Borel automorphisms, Parthasarathy (2013) proved that mapping $m+1$ Gaussians with common identity covariance and affinely independent means to Gaussian laws forces the map to be affine. If an application also requires a unique transport gauge, one may fix the coordinate order and use the almost-everywhere unique Knothe–Rosenblatt monotone triangular rearrangement when source and target have positive densities(Knothe, 1957; Rosenblatt, 1952; Bogachev et al., 2005). Another option is to fix the transport cost in advance. For an absolutely continuous source, the squared Euclidean cost gives an almost-everywhere unique Brenier map(Brenier, 1991). More generally, for costs $c(x, y) = h(x-y)$ with strictly convex h, the optimal transport map is also almost everywhere unique under the corresponding integrability and regularity conditions(Gangbo & McCann, 1996). These uniqueness properties arise from additional structural constraints, not from the likelihood itself.

To report density relative to Lebesgue measure over the independent correlations $c = (C_{ij})_{i<j}$, add

$$\log q_{\mathrm{Leb}}(C \mid k) = \log q_\Theta(C \mid k) + \log|\det J_\phi(C)|. \tag{52}$$

J_ϕ can be constructed from the Loewner form of $D\log_C$. This term is independent of Θ, so omitting it during training does not change the optimum. It may be removed from the objective to reduce computation when optimizing Θ on a fixed dataset.

B.5 Reusable Deep Components on the Correlation Manifold

Global isometric coordinates support constraint-preserving fully connected, activation, residual, normalization, pooling, and attention layers. Complete definitions are collected in Appendix C.5. For the temporal likelihood, the required interface is a deterministic conditioner $c_t = \mathcal{A}_\omega(\mathrm{BOS}, z_{<t})$ that accesses neither the current variable nor the future and therefore contributes no additional change-of-variables Jacobian.

B.6 Unified Generative–Discriminative Heads for Static and Temporal Data

For static samples, we use Eq. (10) directly. For a sequence $C_{1:T}$, the right-shifted causal conditioner detailed in Appendix C.5 produces the temporal DPLR parameters defined in Section 3, with every output a deterministic function of $c_t = \mathcal{A}_\omega(\mathrm{BOS}, z_{<t})$ and the class index. Substitution into Eq. (14) gives the exact class-conditional sequence likelihood.

Proposition B.13 (Normalization conditions for the temporal joint density)**.** *Suppose that (i) T_θ is a global diffeomorphism shared across time points, (ii) c_t is a deterministic measurable function of $z_{<t}$ and no computational path accesses z_t or $z_{>t}$, and (iii) the Jacobian in Eq. (14) includes every invertible transformation from h_t to the evaluated variable z_t. Then Eq. (14) defines a valid class-conditional joint density.*

Restricting only the keys and values to $s < t$ is insufficient. If the query or distance term still contains z_t, the conditional parameters depend on the variable being integrated and are generally no longer normalized. Dropout is disabled in the generative conditioner during both training and evaluation. If a random mask is retained, it must be incorporated into the model as an explicit random variable. Bidirectional context and Eq. (85) may be used only in the discriminative encoder, and no generative likelihood is reported from them.

By Bayes' rule, the predictive probability is

$$p_\Theta(y = k \mid C_{1:T}) = \frac{\pi_k q_\Theta(C_{1:T} \mid k)}{\sum_{j=1}^{K} \pi_j q_\Theta(C_{1:T} \mid j)}. \tag{53}$$

The log-determinant of the shared transport T_θ cancels in the class posterior but is retained in the generative loss. The complete objective and its two regularizers are defined in Eq. (15). $\mathcal{L}_{\mathrm{disc}}$ directly optimizes classification, while $\mathcal{L}_{\mathrm{gen}}$ constrains the class-conditional probabilistic structure of the latent representation to support density estimation and calibration. $\mathcal{R}_{\mathrm{scale}}$ regularizes the partial log-determinant of the affine coupling, and $\mathcal{R}_{\mathrm{metric}}$ controls the full Jacobian, including shear from the translation network and linear mixing. We implement the latter with a stochastic trace estimator based on Jacobian–vector products. Setting $\lambda = 0$ gives a purely discriminative model. Exact NLL uses the complete invertible data path from the observation to z_t with all Jacobians included. The historical parameter-conditioning network itself need not be invertible.

Training and inference procedure. For each mini-batch, we (1) obtain C_t using Eq. (24); (2) compute (h_t, z_t) using Eq. (6), invoking the inverse map only when explicit manifold-valued outputs, data augmentation, or sampling are required; (3) pass through the complete invertible data path and accumulate its log-determinant; (4) evaluate the low-rank Gaussian head directly for static tasks, run Eq. (87) on the right-shifted sequence for temporal generation, or use Eq. (85) for discrimination; and (5) backpropagate Eq. (15). At inference, classification uses Eq. (53). To generate a new correlation matrix, we apply Eq. (11) and, when needed, combine the result with external marginal standard deviations to form a covariance matrix.

B.7 Theoretical Inclusion Relations, Expressivity, and Complexity

Proposition 3.1 is stated in Section 3. Appendix C.6 retains only the class-count distinction, initialization objective, and rank-selection interpretation.

The required low-rank order depends on the number and joint structure of the class covariances, rather than on m alone. This motivates a shared approximate joint diagonalization followed by class-specific low-rank residuals. The full class-count distinction, initialization objective, and rank-selection interpretation are deferred to Appendix C.6. We next derive a necessary cross-class rank bound for this shared congruence model.

Theorem B.14 (Noncommutative rank obstruction under a shared affine transport)**.** Suppose that there exists a common $B \in \mathrm{GL}(m)$ such that every class-conditional covariance satisfies

$$\widetilde{\Sigma}_k = B\Sigma_k B^\top = D_k + U_k U_k^\top, \qquad D_k \succ 0 \text{ is diagonal}, \quad \mathrm{rank}(U_k) \le r_k. \tag{54}$$

For any anchor class a, define the relative covariance operator $R_{k|a} = \Sigma_k \Sigma_a^{-1}$. Then, for any classes i, j,

$$\mathrm{rank}\big([R_{i|a}, R_{j|a}]\big) \le 2(r_i + r_j + 2r_a), \qquad [X, Y] = XY - YX. \tag{55}$$

The common-rank specialization is stated once as the rank lower-bound theorem in Section 3. See the definition of κ_{nc} and Eq. (23). The proof is retained in Appendix D.

When $K \le 2$, each anchor class has at most one nonidentity relative operator, and hence $\kappa_{\mathrm{nc}} = 0$. The bound imposes no rank obstruction, while Corollary B.8 further establishes the sufficiency of $r = 0$. When $K \ge 3$, a nonzero κ_{nc} provides a necessary rank that no shared affine absorption scheme can evade. The constant 8 arises from a general column-space rank bound. We establish a valid necessary condition without claiming that this constant is tight for every covariance family.

With finite samples, numerical rank is sensitive to the threshold, so the tail singular values of the commutator provide a more robust quantity. For a matrix K, define

$$\tau_s(K) = \left(\sum_{\ell > s} \sigma_\ell(K)^2 \right)^{1/2}, \tag{56}$$

where the singular values are ordered nonincreasingly.

Proposition B.15 (Robust version of the noncommutative rank obstruction)**.** *Fix an anchor class a and a coordinate representation. Let $\widehat{R}_{k|a}$ be the relative operator produced by a shared affine DPLR model with common rank r, and suppose that*

$$\|R_{k|a} - \widehat{R}_{k|a}\|_F \le \delta, \qquad k \in \{i, j\}. \tag{57}$$

Let $K_{ij} = [R_{i|a}, R_{j|a}]$. Then

$$\tau_{8r}(K_{ij}) \le 2\delta \left(\|R_{i|a}\|_2 + \|R_{j|a}\|_2 + \delta \right). \tag{58}$$

Consequently, any common-rank-r model attaining Eq. (57) must satisfy

$$\delta \ge \frac{\sqrt{S_{ij}^2 + 2\tau_{8r}(K_{ij})} - S_{ij}}{2}, \qquad S_{ij} = \|R_{i|a}\|_2 + \|R_{j|a}\|_2. \tag{59}$$

Eq. (58) requires δ to be a Frobenius error. If only a spectral-norm error bounded by δ is known, the same proof yields the exact statement

$$\sigma_{8r+1}(K_{ij}) \le 2\delta(S_{ij} + \delta), \tag{60}$$

The corresponding Frobenius tail bound requires an additional factor of $\sqrt{(m - 8r)_+}$. Moreover, commutator rank is invariant under similarity transformations, whereas tail singular values are not. A representation must therefore be fixed in numerical calculations. To reduce the influence of scale conditioning, we use the symmetrically whitened operators $A_{k|a} = \Sigma_a^{-1/2} \Sigma_k \Sigma_a^{-1/2}$ and their commutators, and estimate Eq. (56) using only training-set covariances and a shrinkage parameter selected within the training data. Since $R_{k|a} = \Sigma_a^{1/2} A_{k|a} \Sigma_a^{-1/2}$, the exact rank lower bound is unchanged. The robust tail values are always reported in this fixed whitening gauge.

This result turns noncommutativity into a practical rank diagnostic for a shared affine DPLR head. Eqs. (23) and (56) suggest candidate values of r, which can then be compared with the validation-optimal rank. The exact statement concerns necessary affine-head capacity, while the robust version quantifies relative-covariance approximation error. Together with the inclusion proposition and affine absorption results, it links population expressivity, empirical NLL, and the rank sweep evaluated in Section E.7.

In terms of computational efficiency, correlation projection costs $O(d^2)$ per window, while the matrix logarithm in the forward off-log coordinates costs $O(d^3)$. Explicit inverse off-log reconstruction

requires I_{N} iterations of the diagonal solver, with a dominant matrix-function cost of approximately $O(I_{\mathrm{N}} d^3)$. Our implementation uses Jacobian–vector products to avoid explicitly constructing a fourth-order Fréchet tensor, whose naive construction can be more expensive. Let the coupling depth be L and the conditioner width be r_f. Applying dense LU mixing costs $O(Lm^2)$, while using R_H Householder reflections per layer costs $O(LR_H m)$. The conditioner adds approximately $O(Lmr_f)$. The low-rank class-conditional covariance uses the Woodbury identity to compute its inverse and log-determinant at a per-class cost of $O(mr^2 + r^3)$, avoiding the $O(d^4)$ parameter growth of the original full-Σ estimate. Full attention costs $O(T^2 p)$. For long sequences, restricting the history to a local window of length w reduces this cost to $O(Twp)$.

B.8 Joint Scale–Correlation Extension

The forward scale–correlation map, its global-diffeomorphism property, its Jacobian, and the resulting covariance density are stated once in the scale–correlation theorem of Section 3. See Eqs. (21)–(22). This appendix supplies the inverse transformation and the extensions needed for conditional scale modeling and product geometry. The inverse transformation is

$$S = \mathrm{Diag}(e^{v/2}) C \,\mathrm{Diag}(e^{v/2}). \tag{61}$$

In the independent-element coordinates used in the main theorem, its inverse Jacobian is

$$\left| \det \frac{\partial \,\mathrm{vech}(S)}{\partial (v, c)} \right| = \exp\left[\frac{d+1}{2} \mathbf{1}^\top v \right]. \tag{62}$$

The independent-branch density in Eq. (22) is normalized with respect to Lebesgue measure on the independent SPD elements. Its off-log term converts the correlation branch from dvol_{g^0} to $\mathrm{d}c$, while the final term is the forward scale–correlation Jacobian. More generally, $p_v(v \mid k)$ may be replaced by a normalized conditional density $p_v(v \mid C, k)$ to model scale–correlation dependence without changing that Jacobian. To sample, we first draw (v, C) from the corresponding joint or conditional branches and then reconstruct S using Eq. (61).

This decomposition also provides a tractable geometry on the full SPD space. Let $T_s : \mathbb{R}^d \to \mathbb{R}^d$ and $T_c : \mathbb{R}^m \to \mathbb{R}^m$ both be smooth global diffeomorphisms, and define

$$\Xi(S) = \big(T_s(v), T_c(\phi(C))\big), \qquad g^{\mathrm{SPD}} = \Xi^* g_{\mathrm{E}}. \tag{63}$$

Because F, ϕ, T_s, and T_c are all global diffeomorphisms, Ξ is a global isometry from $(\mathbb{S}^d_{++}, g^{\mathrm{SPD}})$ to $\mathbb{R}^{d+m}$. This space is therefore also a complete, flat, and simply connected Hadamard manifold, and it inherits closed-form distances, geodesics, and Fréchet means analogous to the adaptive operations in Appendix C.3.

Three implementation choices align the model with the data. Variable identities are kept consistent between training and testing, or the conditioner in T_θ can use a permutation-equivariant structure. The shrinkage parameter ε in Eq. (24) is selected using training data only. Correlation-only tasks use the quotient representation, while tasks that use marginal variance adopt the joint scale–correlation model in Eq. (22).

C Technical Constructions and Implementation Details

This appendix is the implementation companion to Section 3 and Appendix B. Whereas Appendix B organizes the formal statements and their methodological consequences, the sections below specify the numerical reconstruction, regularization, closed-form operators, network layers, and capacity parameterizations needed to reproduce the model. The notation and assumptions are unchanged, proofs are kept separately in Appendix D.

C.1 Scale-Equivariant Regularization and the Quotient Guarantee

Post-projection shrinkage must be distinguished from a covariance ridge applied before projection. The common choice $S + \lambda I_d$ is not, in general, equivariant under positive diagonal congruence and

therefore breaks scale-quotient invariance numerically. When a downstream SPD procedure requires a positive-definite covariance before projection, we instead use

$$\mathcal{R}_\lambda(S) = S + \lambda D_S, \qquad \lambda > 0. \tag{64}$$

Because $D_{ASA} = AD_SA$, this choice satisfies $\mathcal{R}_\lambda(ASA) = A\mathcal{R}_\lambda(S)A$. It is therefore compatible with the correlation projection and does not reintroduce positive diagonal gains into the correlation branch. More explicitly, $\mathrm{Corr}(\mathcal{R}_\lambda(S)) = (1+\lambda)^{-1}\,\mathrm{Corr}(S) + \lambda(1+\lambda)^{-1}I_d$, which is precisely a scale-invariant correlation shrinkage. In the exact-invariance audit, we use double precision before applying gains and use Eq. (64) for every pre-projection ridge.

The maximal-invariant statement is given in Appendix B as Proposition B.1. The elementary spectral bounds follow directly from Eq. (24).

C.2 Off-Log Inversion, Differentiation, and Base Geometry

Inverse reconstruction. Section 2 gives the inverse reconstruction and its strictly convex characterization. Writing $M(a,h) = \mathrm{Diag}(a) + \mathrm{hvec}^{-1}(h)$ and $r(a,h) = \mathrm{diag}(\exp M(a,h)) - \mathbf{1}$, we solve $r(a,h) = 0$ using damped Newton iterations. Backpropagation does not unroll the solver. Instead, implicit differentiation of the converged equation gives $D_h a^\star = -(D_a r)^{-1} D_h r$ (Blondel et al., 2022). Below we record the matrix-exponential derivative and stopping rule used in the implementation.

Differentiation through the inverse. By the standard Daleckii–Krein formula (Higham, 2008), if $M = Q\Lambda Q^\top$, the Fréchet derivative of the matrix exponential is

$$D\exp_M[E] = Q\big(\Omega \odot (Q^\top E Q)\big)Q^\top, \quad \Omega_{ij} = \begin{cases} \dfrac{e^{\lambda_i} - e^{\lambda_j}}{\lambda_i - \lambda_j}, & \lambda_i \neq \lambda_j, \\ e^{\lambda_i}, & \lambda_i = \lambda_j. \end{cases} \tag{65}$$

Equation (65) supplies the directional derivatives used by both the Newton system and its implicit adjoint. We use $\|r\|_\infty \le 10^{-7}$ as the default stopping criterion. Positive definiteness is preserved by the matrix exponential, while the unit diagonal holds numerically only up to this residual tolerance.

The base metric and its reference volume are defined in Section 3. The space $(\mathcal{C}_d, g^0)$ is globally isometric to $\mathbb{R}^m$ and is a complete, simply connected, zero-curvature Hadamard manifold. In particular,

$$d_0(C,Q) = \|\phi(C) - \phi(Q)\|_2, \tag{66}$$

$$\mathrm{Exp}_C^0(U) = \phi^{-1}\left(\phi(C) + D\phi_C[U]\right), \tag{67}$$

$$\mathrm{Log}_C^0(Q) = (D\phi_C)^{-1}\left[\phi(Q) - \phi(C)\right], \tag{68}$$

$$\overline{C}_w = \phi^{-1}\left(\sum_i w_i \phi(C_i)\right), \quad w_i \ge 0,\ \sum_i w_i = 1. \tag{69}$$

These expressions follow directly from the global Euclidean coordinates and avoid both the base-point-dependent Jacobian of the exponential map and the base-point–mean redundancy of the AIRM-SPD wrapped Gaussian(De Surrel et al., 2025).

C.3 Closed-Form Operations under the Adaptive Geometry

The pullback-geometry result is stated in Section 3. Its explicit operators are

$$\mathrm{Exp}_C^\theta(U) = \Psi_\theta^{-1}\left(\Psi_\theta(C) + D\Psi_{\theta,C}[U]\right), \tag{70}$$

$$\mathrm{Log}_C^\theta(Q) = (D\Psi_{\theta,C})^{-1}\left[\Psi_\theta(Q) - \Psi_\theta(C)\right], \tag{71}$$

$$\Gamma_{C\to Q}^\theta(U) = (D\Psi_{\theta,Q})^{-1} D\Psi_{\theta,C}[U], \tag{72}$$

$$\overline{C}_w^\theta = \Psi_\theta^{-1}\left(\sum_i w_i \Psi_\theta(C_i)\right). \tag{73}$$

In exact arithmetic, all of these outputs belong to $\mathcal{C}_d$.

When $T_\theta = \mathrm{Id}$, g^θ reduces to the fixed off-log metric. A fixed OLM network is therefore one point in this geometric family. CorNet constructs multinomial logistic regression (MLR), fully connected, and convolutional layers under fixed correlation geometries(Chen et al., 2026). In contrast, CorTG uses an invertible transport to select a data-adaptive geometric representative and shares its parameters with a normalized probabilistic model. This representative has the gauge freedom described in the main text and cannot be interpreted as a unique geometry from likelihood alone.

C.4 Intrinsic Centers and Model Probability Regions

The intrinsic-center and probability-region result is summarized in Section 3. Its explicit center, learned-volume density, and probability regions are as follows. Define

$$M_k = \Psi_\theta^{-1}(\mu_k). \tag{74}$$

If $\Sigma_k = \sigma_k^2 I$, the density relative to the learned volume can be written as

$$q_k^{(\theta)}(C) = (2\pi\sigma_k^2)^{-m/2} \exp\left[-\frac{d_\theta(C, M_k)^2}{2\sigma_k^2}\right]. \tag{75}$$

For a general $\Sigma_k \succ 0$, if $C \sim Q_k$, then

$$R_k^2(C) = (\Psi_\theta(C) - \mu_k)^\top \Sigma_k^{-1} (\Psi_\theta(C) - \mu_k) \sim \chi_m^2. \tag{76}$$

Letting $\chi^2_{m,1-\alpha}$ denote the $1-\alpha$ quantile of χ_m^2, the set

$$\mathcal{E}_{k,\alpha} = \{C \in \mathcal{C}_d : R_k^2(C) \le \chi^2_{m,1-\alpha}\} \tag{77}$$

satisfies $Q_k(\mathcal{E}_{k,\alpha}) = 1-\alpha$. In the isotropic case, it reduces to the learned geodesic ball

$$d_\theta(C, M_k) \le \sigma_k \sqrt{\chi^2_{m,1-\alpha}}. \tag{78}$$

C.5 Reusable Coordinate-Based Network Layers

The following layers are defined in the chosen global isometric coordinates $z = \Psi_\theta(C)$ and mapped back to the correlation manifold through Ψ_θ^{-1}. They are therefore constraint-preserving, manifold-valued network layers. Distance, exponential and logarithmic maps, parallel transport, and the induced Fréchet mean are intrinsic to the learned metric. By contrast, linear maps, coordinate-wise activations, and scaled dot-product attention (Vaswani et al., 2017) depend on the selected coordinate frame.

Manifold fully connected layer and dimension change. A correlation fully connected layer from $\mathcal{C}_d$ to $\mathcal{C}_{d'}$ is

$$\mathrm{CorrFC}_{A,b}(C) = \Psi_{\theta,d'}^{-1}\left(A\Psi_{\theta,d}(C) + b\right), \quad A \in \mathbb{R}^{m' \times m}, \quad m' = \frac{d'(d'-1)}{2}. \tag{79}$$

The output remains a valid full-rank correlation matrix even when A is rectangular, such a layer is used only for discriminative representation. If $d' = d$ and A is invertible, the layer is itself a diffeomorphism. When it lies on the data-transformation path of the evaluated variable, $\log|\det A|$ must be included in the likelihood.

Activation and residual layers. The discriminative branch may use

$$\mathrm{CorrAct}(C) = \Psi_\theta^{-1}(\rho(\Psi_\theta(C))), \tag{80}$$

where ρ is a coordinate-wise Gaussian error linear unit (GELU) (Hendrycks & Gimpel, 2016). Its residual form

$$C^+ = \mathrm{Exp}_C^\theta\left((D\Psi_{\theta,C})^{-1} F(\Psi_\theta(C))\right) = \Psi_\theta^{-1}\left(z + F(z)\right) \tag{81}$$

remains on the manifold under the exact inverse map. At this point, two computational paths must be distinguished. The data-transformation path from the observation h_t to the evaluated variable z_t must be invertible, and every Jacobian on that path must be included. The conditional parameter network that generates $(\mu_{t,k}, \Sigma_{t,k})$ from the history need not be invertible. It may use GELU, pooling, or standard attention provided that its inputs are strictly historical and its output is deterministic given that history.

Manifold normalization. The exact Fréchet mean of a mini-batch is $\overline{C}_B^{\theta} = \Psi_\theta^{-1}(\overline{z})$, where $\overline{z} = B^{-1}\sum_i z_i$. Define

$$\widetilde{z}_i = G\,(V_B + \eta I)^{-1/2}(z_i - \overline{z}) + b, \qquad \widetilde{C}_i = \Psi_\theta^{-1}(\widetilde{z}_i), \tag{82}$$

to obtain full-covariance manifold batch normalization. When $B < m$, we use a diagonal or grouped V_B. If normalization lies on the data-transformation path, we use ActNorm (Kingma & Dhariwal, 2018) fixed after training and include its Jacobian, rather than statistics that depend on other samples in the same batch. The conditional parameter network is not subject to this invertibility restriction.

Convolution and weighted pooling. For a local receptive field $\mathcal{R}(t)$, the intrinsic convolution with a nonnegative normalized kernel w_{ts} is

$$P_t = \Psi_\theta^{-1}\left(\sum_{s\in\mathcal{R}(t)} w_{ts}\Psi_\theta(C_s)\right). \tag{83}$$

By Eq. (73), this layer is exactly the weighted Fréchet mean under the learned metric.

Coordinate-based geodesic attention and strictly historical conditioner. Let $z_t = \Psi_\theta(C_t)$. For a purely discriminative branch, we use the current point as the query and add a learned-geodesic-distance bias to the dot-product similarity:

$$e_{ts}^{\mathrm{disc}} = \frac{(W_Q z_t)^\top (W_K z_s)}{\sqrt{p}} - \tau\|z_t - z_s\|_2^2 + b_{t-s}, \tag{84}$$

$$\widetilde{C}_t = \Psi_\theta^{-1}\left(z_t + W_O\sum_s \mathrm{softmax}_s(e_{ts}^{\mathrm{disc}})W_V z_s\right). \tag{85}$$

Eq. (85) depends on the current z_t and therefore cannot be used directly to construct an autoregressive conditional density for z_t. The generative conditioner instead uses the right-shifted input $[\mathrm{BOS}, z_1, \ldots, z_{T-1}]$. Let $\bar{z}_t = g_\omega(\mathrm{BOS}, z_{<t})$ be the output of the causal network at position t. Then $\bar{z}_t$ is measurable with respect to the history-generated σ-algebra $\sigma(z_{<t})$. For $s < t$, define

$$e_{ts}^{\mathrm{gen}} = \frac{(W_Q \bar{z}_t)^\top (W_K z_s)}{\sqrt{p}} - \tau\|\bar{z}_t - z_s\|_2^2 + b_{t-s}, \qquad \tau = \mathrm{softplus}(\widetilde{\tau}) \geq 0, \tag{86}$$

$$\alpha_{ts}^{\mathrm{gen}} = \frac{\exp e_{ts}^{\mathrm{gen}}}{\sum_{j<t}\exp e_{tj}^{\mathrm{gen}}}, \qquad c_t = \bar{z}_t + W_O\sum_{s<t}\alpha_{ts}^{\mathrm{gen}} W_V z_s. \tag{87}$$

When $t = 1$, c_1 is determined solely by the BOS state. The query, keys, and values in Eq. (87) contain neither current nor future variables, so c_t depends only on $z_{<t}$. This conditioner only outputs context for the distribution parameters and does not modify the evaluated z_t. It therefore need not be invertible and introduces no additional change-of-variables Jacobian.

C.6 Distribution-Family Inclusion and Low-Rank Capacity

Proposition 3.1 and its population-risk consequence are stated in Section 3. That section also states the exact one- and two-class absorption result and the multiclass commuting criterion. Here we retain the constructive initialization and its implication for rank selection.

This structure suggests first using the shared transport for approximate joint diagonalization and then modeling the remaining dependence with class-specific low-rank terms. Given covariance estimates $\{\Sigma_k\}_{k=1}^K$, a shared linear initialization can solve

$$\min_{B\in\mathrm{GL}(m):\, B\Sigma_1 B^\top = I} \sum_{k=2}^{K} \left\|\mathrm{off}(B\Sigma_k B^\top)\right\|_F^2. \tag{88}$$

The Gaussian heads $\mathrm{Diag}(v_k) + U_k U_k^\top$ are then trained jointly. Eq. (88) provides a more direct basis for selecting r than the original covariance dimension. Here, r controls the residual complexity that still requires class-specific structure after the shared congruence transformation. Approximate joint diagonalization and diagonal-plus-low-rank decompositions of a single matrix have each been studied systematically(Afsari, 2008; Drton et al., 2007; Saunderson et al., 2012). The noncommutative rank results in the main text next derive a necessary cross-class bound that arises when these components are coupled through a shared congruence transformation.

D PROOFS OF FORMAL RESULTS

This appendix gives the formal proofs retained separately from their statements in Section 3 and Appendices B– C, following their order of presentation. It uses the notation introduced in Section 3 and developed in those appendices. The scale-invariance proposition is justified directly where it is stated.

D.1 GLOBAL OFF-LOG DIFFEOMORPHISM AND INVERSE CHARACTERIZATION

Proof. Fix h and let $M(a,h) = \mathrm{Diag}(a) + \mathrm{hvec}^{-1}(h)$. The gradient of the strictly convex objective stated in Section 2 is $\mathrm{diag}(\exp M) - \mathbf{1}$. The function $\mathrm{tr}\exp(\cdot)$ is strictly convex on the symmetric matrix space, and its Hessian remains positive definite when restricted to nonzero diagonal directions. Next, let $M_a = \max_i a_i$. If $M_a \to \infty$, the objective is bounded below by $e^{M_a} - dM_a \to \infty$. If M_a is bounded while $\|a\| \to \infty$, at least one component tends to $-\infty$, so $-\mathbf{1}^\top a \to \infty$. The objective is therefore coercive and has a unique minimizer. Its first-order condition gives $\mathrm{diag}(\exp M(a^\star, h)) = \mathbf{1}$, so the inverse characterized in the main text lies in $\mathcal{C}_d$. Conversely, for any $C \in \mathcal{C}_d$, the choices $h = \phi(C)$ and $a = \mathrm{diag}(\log C)$ satisfy the same first-order condition. Uniqueness gives bijectivity. The Hessian is invertible, and the implicit function theorem ensures that $a^\star(h)$ and the inverse map are smooth. Thus, ϕ is a diffeomorphism. □

D.2 SPECTRAL BOUNDS FOR THE CONSTRAINED OFF-LOG SCHUR COMPLEMENT

Proof. Let $M = Q\,\mathrm{Diag}(\mu_1, \ldots, \mu_d)Q^\top$, so $C = Q\,\mathrm{Diag}(e^{\mu_1}, \ldots, e^{\mu_d})Q^\top$. By Eq. (65), the eigenvalues of $\mathcal{L}_C$ in the rotated orthonormal basis of symmetric matrices are

$$\ell_{pp} = e^{\mu_p}, \qquad \ell_{pq} = \frac{e^{\mu_p} - e^{\mu_q}}{\mu_p - \mu_q} \quad (p < q), \tag{89}$$

For repeated eigenvalues, use the continuous extension $\ell_{pq} = e^{\mu_p}$. The mean value theorem for the scalar exponential gives $\alpha \le \ell_{pq} \le \beta$, hence $\alpha I \preceq \mathcal{L}_C \preceq \beta I$. Principal compression preserves the Loewner order, which gives the bound for A_C. We now detail the bound for S_C. Partition $\mathcal{L}_C$ according to $\mathcal{D}_d \oplus \mathbb{S}_d$ and apply the standard block-matrix inverse formula to Eq. (25) to obtain

$$\mathcal{L}_C^{-1} = \begin{bmatrix} A_C^{-1} + A_C^{-1}B_C S_C^{-1} B_C^\top A_C^{-1} & -A_C^{-1}B_C S_C^{-1} \\ -S_C^{-1}B_C^\top A_C^{-1} & S_C^{-1} \end{bmatrix}.$$

Because $A_C \succ 0$, the Schur complement S_C and the block inverse above are well defined. In particular, the lower-right block of $\mathcal{L}_C^{-1}$ is exactly S_C^{-1}. Equivalently, for any $H \in \mathbb{S}_d$, applying $\mathcal{L}_C^{-1}$ and then orthogonally projecting back to $\mathbb{S}_d$ yields the operator S_C^{-1}. Therefore,

$$S_C^{-1} = P_{\mathbb{S}_d}\mathcal{L}_C^{-1}\big|_{\mathbb{S}_d}, \tag{90}$$

where $P_{\mathbb{S}_d}$ is the orthogonal projection onto the hollow subspace. Inverting $\alpha I \preceq \mathcal{L}_C \preceq \beta I$ reverses the Loewner order and gives $\beta^{-1} I \preceq \mathcal{L}_C^{-1} \preceq \alpha^{-1} I$. Compressing this inequality to $\mathbb{S}_d$ and applying Eq. (90) yields $\beta^{-1} I_m \preceq S_C^{-1} \preceq \alpha^{-1} I_m$. Inverting once more gives $\alpha I_m \preceq S_C \preceq \beta I_m$, and Eq. (27) follows.

Finally, differentiating the constraint equation gives $A_C\,\mathrm{d}a + B_C\,\mathrm{d}h = 0$. Substituting it into $\mathrm{hvec}(\mathrm{off}(\mathrm{d}C)) = B_C^\top \mathrm{d}a + D_C \mathrm{d}h$ yields

$$\mathrm{hvec}(\mathrm{off}(\mathrm{d}C)) = S_C\,\mathrm{d}h. \tag{91}$$

Because $\mathrm{hvec}(\mathrm{off}(\mathrm{d}C)) = \sqrt{2}\,\mathrm{d}c$, Eq. (28) follows. □

D.3 OFF-LOG VOLUME AND SHRINKAGE STABILITY

Proof. Eq. (28) gives $|\det J_\phi| = 2^{m/2}/\det S_C$. Combining this identity with the block-determinant formula $\det \mathcal{L}_C = \det A_C \det S_C$ yields Eq. (19). In Eq. (89), the product of the d eigenvalues along diagonal directions is $\det C$. The remaining m hollow directions give Eq. (20). Finally, using $\varepsilon \le \lambda(C_\varepsilon) \le (1-\varepsilon)d + \varepsilon$ in Theorem B.2 completes the proof. □

D.4 Closed-form operations under the adaptive geometry

Proof. The pullback definition makes Ψ_θ a global Riemannian isometry from $(\mathcal{C}_d, g^\theta)$ to Euclidean space. Completeness, zero curvature, and simple connectedness are preserved by the isometry. The distance stated in Section 3 and Eqs. (70)–(73) are the pullbacks through Ψ_θ^{-1} of Euclidean straight lines, translations, and weighted means. Finally, $\Psi_\theta^{-1} = \phi^{-1} \circ T_\theta^{-1}$, and the global inverse characterization in Section 3 ensures that the image of ϕ^{-1} is exactly $\mathcal{C}_d$. Thus, in exact arithmetic, positive definiteness and the unit diagonal require neither projection nor clipping. □

D.5 Normalization, sampling, and conditional identifiability under fixed transport

Proof. Because ϕ gives isometric coordinates from $(\mathcal{C}_d, g^0)$ to $\mathbb{R}^m$, we have $\mathrm{dvol}_{g^0} = \mathrm{d}h$. The change of variables $z = T_\theta(h)$ reduces the integral of Eq. (10) to a Euclidean Gaussian integral and its inverse gives Eq. (11). If two distributions are equal, pushing them forward to $\mathbb{R}^m$ through the bijection $T_\theta \circ \phi$ gives two identical nondegenerate Gaussians. Identifiability of Euclidean Gaussians then proves the theorem. □

D.6 Duality between fixed reference and learned volumes

Proof. The pullback definition makes the learned metric matrix the Gram matrix of JT_θ in off-log coordinates. Taking its determinant gives the volume ratio stated in Section 3. Multiplying the fixed-volume density by the reciprocal ratio cancels the flow Jacobian and leaves the Gaussian density in Ψ_θ coordinates. □

D.7 Intrinsic center and model probability regions

Proof. By the pullback-geometry result in Section 3, Ψ_θ is a global isometry to Euclidean space. Thus, $d_\theta(C, M_k) = \|\Psi_\theta(C) - \mu_k\|_2$, which proves Eq. (75). Moreover, $\Psi_\theta(C) \sim \mathcal{N}(\mu_k, \Sigma_k)$, so the standardized squared Mahalanobis distance follows χ^2_m. This gives the exact probability mass and its isotropic special case. □

D.8 Fréchet center and strong geodesic convexity

Proof. In the global isometric coordinates $z = \Psi_\theta(C)$, Eq. (30) reduces to minimizing $\mathbb{E}\|z - u\|_2^2$. Its unique solution is $u = \mathbb{E}[z] = \mu_k$. The same isometry maps g^θ-geodesics to Euclidean lines, while the Euclidean Hessian of Eq. (31) is always Σ_k^{-1}. Pulling back this quadratic form and applying the Rayleigh quotient bounds gives Eqs. (32)–(33). □

D.9 Affine absorption identity

Proof. Gaussian change of variables gives

$$\mathcal{N}(Bz + b; B\mu_k + b, B\Sigma_k B^\top) = |\det B|^{-1} \mathcal{N}(z; \mu_k, \Sigma_k), \tag{92}$$

Moreover, $|\det JT'_\theta| = |\det B|\,|\det JT_\theta|$, so the two factors cancel and yield the same density. The corresponding metric is $g'_C(U, V) = \langle BD\Psi_{\theta,C}[U], BD\Psi_{\theta,C}[V]\rangle$, which is necessarily equal to g^θ only when $B^\top B = I$. □

D.10 Full-covariance absorption for one or two classes

Proof. The single-class result follows directly from Eq. (92). The two-class result uses the orthogonal eigendecomposition of the symmetric positive-definite matrix A. Means only need to be transformed simultaneously according to Eq. (35) and impose no additional constraint on simultaneous diagonalization. □

D.11 Necessary and sufficient condition for shared multiclass diagonalization

Proof. If Eq. (38) holds, the family of real symmetric matrices $\{A_k\}_{k=2}^K$ can be simultaneously diagonalized by one orthogonal matrix U. Choosing $B = U^\top \Sigma_1^{-1/2}$ suffices. Conversely, suppose that every $D_k = B\Sigma_k B^\top$ is diagonal and define $Q = D_1^{-1/2} B \Sigma_1^{1/2}$. Since $QQ^\top = I$, the matrix Q is orthogonal, and $QA_kQ^\top = D_1^{-1/2} D_k D_1^{-1/2}$ is diagonal. All A_k can therefore be simultaneously diagonalized by the same orthogonal matrix and consequently commute pairwise. □

D.12 Fixed-head likelihood stabilizer and metric equivalence class

Proof. For any $z = T_1(h)$, the chain rule gives

$$q_{2,k}(h) = p_k(Gz)|\det JG(z)|\,|\det JT_1(h)|. \tag{93}$$

Because T_1 is surjective, class-wise equality in Eq. (41) is exactly equivalent to Eq. (40). On the other hand,

$$JT_2(h)^\top JT_2(h) = JT_1(h)^\top JG(z)^\top JG(z) JT_1(h), \tag{94}$$

Thus, the two pullback metrics are equal if and only if $JG(z)^\top JG(z) = I$ for all z. In this case, G preserves the length of every piecewise smooth curve. Applying G and its inverse to the line segment between two points gives $\|Gx - Gy\|_2 = \|x - y\|_2$. Every global Euclidean isometry has the form $G(z) = Oz + b$. The converse is immediate. □

D.13 Nonlinear radial twist preserving a standard Gaussian

Proof. In polar coordinates (r, ϑ) on the x plane, Eq. (46) is $(r, \vartheta, w) \mapsto (r, \vartheta + a(r^2), w)$. Replacing a by $-a$ gives its inverse. Since the rotation angle is a smooth function of $z_1^2 + z_2^2$, the map is smooth at the origin as well. It preserves radius and has unit Jacobian with respect to the volume element $r\,\mathrm{d}r\,\mathrm{d}\vartheta\,\mathrm{d}w$, so it preserves the standard Gaussian measure. On the other hand, on this two-dimensional plane,

$$G_a^*(\mathrm{d}r^2 + r^2\mathrm{d}\vartheta^2) = \big(1 + 4r^4 a'(r^2)^2\big)\mathrm{d}r^2 + 4r^3 a'(r^2)\mathrm{d}r\,\mathrm{d}\vartheta + r^2\mathrm{d}\vartheta^2. \tag{95}$$

Whenever $a' \not\equiv 0$, this metric differs from the Euclidean metric. □

D.14 Multiclass identifiability phase transition for shared-covariance Gaussians

Proof. Let $v_k = \Sigma^{-1/2}(\mu_k - \mu_1)$, where $v_1 = 0$. After whitening, the kth density is $\gamma_m(x - v_k)$. If H preserves all these densities, divide the preservation equation for class k by that for class 1. The common Jacobian cancels, giving

$$\exp\left(v_k^\top H(x) - \tfrac{1}{2}\|v_k\|_2^2\right) = \exp\left(v_k^\top x - \tfrac{1}{2}\|v_k\|_2^2\right). \tag{96}$$

Thus, $v_k^\top H(x) = v_k^\top x$ for every k, or equivalently $P_W H(x) = P_W x$. It follows that $H(w, y) = (w, F_w(y))$. Its Jacobian is block lower triangular with determinant $\det D_y F_w(y)$. Substituting this form into the preservation equation for class 1 and canceling the standard Gaussian factor on W gives Eq. (50). Conversely, that equation preserves the W component and the standard Gaussian measure on each fiber, so it preserves every $\gamma_m(x - v_k)$.

When $q = 0$, no residual fiber remains, so $H = \mathrm{Id}$. When $q = 1$, every one-dimensional diffeomorphism F_w must be strictly monotone. Strict monotonicity of the standard Gaussian cumulative distribution function (CDF) implies that the increasing case is only $F_w(y) = y$, while the decreasing case is only $F_w(y) = -y$. Smoothness and connectedness of the parameter domain W exclude sign changes with w. Thus, only the identity and a global reflection remain in whitened coordinates. Conjugating the latter back to the original latent coordinates using Eq. (49) gives linear part A. Direct calculation yields $A^\top A = \Sigma^{-1/2} S_W \Sigma S_W \Sigma^{-1/2}$, which is Eq. (51).

When $q \geq 2$, fix a two-dimensional subspace of $W^\perp$ and define

$$F_w(y) = \left(R_{a(w,\rho^2)} y_{1:2},\, y_{3:q}\right), \qquad \rho^2 = y_1^2 + y_2^2, \tag{97}$$

where a is any smooth function. The same polar-coordinate calculation as in the proposition shows that this map preserves fiber radius and volume. Negating the rotation angle gives its inverse, so it satisfies Eq. (50). A nonconstant a generally yields a non-isometric map. The infinite-dimensional freedom in a gives the stated nonlinear subfamily. Applying Theorem B.10 then yields the metric-identifiability result. □

D.15 Normalization conditions for the temporal joint density

Proof. Conditioned on the history $z_{<t}$, the quantities c_t, $\mu_{t,k}$, and $\Sigma_{t,k}$ are constants. Therefore, $\mathcal{N}(z_t; \mu_{t,k}, \Sigma_{t,k})$ integrates to one with respect to z_t. The chain factorization $q(z_{1:T} \mid k) = \prod_t q(z_t \mid z_{<t}, k)$, followed by successive integration from $t = T$ down to 1, shows that the joint density integrates to one. Applying the complete change-of-variables formula at every time point then yields a normalized density on the correlation manifold. □

D.16 Distribution-family inclusion and optimal risk

Proof. Set $s_k = t_k = 0$ and $L_k = I$ in every coupling layer, and either disable Eq. (39) or fix it to the identity gauge. Then $T_\theta = \mathrm{Id}$, and Eq. (10) reduces exactly to a single Gaussian in the base coordinates with the same $\Sigma \in \mathfrak{S}_r$. The set inclusion immediately yields the relation between the KL infima. □

D.17 Noncommutative rank obstruction under a shared affine transport

Proof. The Woodbury identity gives

$$\widetilde{\Sigma}_a^{-1} = D_a^{-1} - V_a V_a^\top, \quad V_a = D_a^{-1} U_a \big(I + U_a^\top D_a^{-1} U_a\big)^{-1/2}. \tag{98}$$

where $\mathrm{rank}(V_a) \le r_a$. Therefore,

$$\widetilde{\Sigma}_k \widetilde{\Sigma}_a^{-1} = \Lambda_k + E_k, \quad \Lambda_k = D_k D_a^{-1}, \quad E_k = U_k U_k^\top \widetilde{\Sigma}_a^{-1} - D_k V_a V_a^\top. \tag{99}$$

where Λ_k is diagonal and $\mathrm{rank}(E_k) \le r_k + r_a$. To obtain the rank bound for the commutator, first expand it term by term:

$$\begin{aligned}[\Lambda_i + E_i, \Lambda_j + E_j] &= [\Lambda_i, \Lambda_j] + [\Lambda_i, E_j] + [E_i, \Lambda_j] + [E_i, E_j] \\ &= [\Lambda_i, E_j] + [E_i, \Lambda_j] + [E_i, E_j],\end{aligned}.$$

The second equality uses the fact that the diagonal matrices Λ_i and Λ_j commute, so $[\Lambda_i, \Lambda_j] = 0$. The column spaces of the individual terms satisfy

$$\begin{aligned}\mathrm{col}([\Lambda_i, E_j]) &= \mathrm{col}(\Lambda_i E_j - E_j \Lambda_i) \subseteq \Lambda_i \,\mathrm{col}(E_j) + \mathrm{col}(E_j), \\ \mathrm{col}([E_i, \Lambda_j]) &= \mathrm{col}(E_i \Lambda_j - \Lambda_j E_i) \subseteq \mathrm{col}(E_i) + \Lambda_j \,\mathrm{col}(E_i), \\ \mathrm{col}([E_i, E_j]) &= \mathrm{col}(E_i E_j - E_j E_i) \subseteq \mathrm{col}(E_i) + \mathrm{col}(E_j).\end{aligned}$$

The last inclusion follows from $\mathrm{col}(E_i E_j) \subseteq \mathrm{col}(E_i)$ and $\mathrm{col}(E_j E_i) \subseteq \mathrm{col}(E_j)$. Thus, the column space of the quadratic term $[E_i, E_j]$ is already covered by the subspaces appearing in the first two terms. Using $\mathrm{col}(X + Y) \subseteq \mathrm{col}(X) + \mathrm{col}(Y)$ then gives

$$\begin{aligned}\mathrm{col}\big([\Lambda_i + E_i, \Lambda_j + E_j]\big) \subseteq\ & \Lambda_i \,\mathrm{col}(E_j) + \mathrm{col}(E_j) \\ & + \mathrm{col}(E_i) + \Lambda_j \,\mathrm{col}(E_i),\end{aligned} \tag{100}$$

For any linear map L, $\dim(L \,\mathrm{col}(E_k)) \le \mathrm{rank}(E_k)$. Moreover, the dimension of a sum of subspaces does not exceed the sum of their dimensions. Eq. (100) therefore gives

$$\begin{aligned}\mathrm{rank}([\Lambda_i + E_i, \Lambda_j + E_j]) &\le 2\,\mathrm{rank}(E_i) + 2\,\mathrm{rank}(E_j) \\ &\le 2(r_i + r_j + 2r_a).\end{aligned}$$

Finally,

$$\widetilde{\Sigma}_k \widetilde{\Sigma}_a^{-1} = B R_{k|a} B^{-1}. \tag{101}$$

Thus, the two commutators are similar and have the same rank, which proves Eq. (55). □

D.18 Testable necessary lower bound for a common low-rank order

Proof. When $r_i, r_j, r_a \le r$, Eq. (55) bounds the rank of every commutator by $8r$. Taking the maximum over all anchor classes and class pairs and using the integrality of r proves the result. □

D.19 Robust version of the noncommutative rank obstruction

Proof. Let $\widehat{K}_{ij} = [\widehat{R}_{i|a}, \widehat{R}_{j|a}]$. By Theorem B.14, $\operatorname{rank}(\widehat{K}_{ij}) \le 8r$. The Eckart–Young theorem therefore gives $\tau_{8r}(K_{ij}) \le \|K_{ij} - \widehat{K}_{ij}\|_F$. Define $\Delta_k = R_{k|a} - \widehat{R}_{k|a}$. Then

$$K_{ij} - \widehat{K}_{ij} = [\Delta_i, R_{j|a}] + [\widehat{R}_{i|a}, \Delta_j]. \tag{102}$$

This decomposition follows directly from $R_{i|a} = \widehat{R}_{i|a} + \Delta_i$. For the first term, the triangle inequality and $\|AX\|_F \le \|A\|_2 \|X\|_F$ and $\|XA\|_F \le \|X\|_F \|A\|_2$ yield

$$\begin{aligned} \|[\Delta_i, R_{j|a}]\|_F &\le \|\Delta_i R_{j|a}\|_F + \|R_{j|a}\Delta_i\|_F \\ &\le 2\|\Delta_i\|_F \|R_{j|a}\|_2 \le 2\delta \|R_{j|a}\|_2. \end{aligned}$$

Likewise, $\|[\widehat{R}_{i|a}, \Delta_j]\|_F \le 2\delta \|\widehat{R}_{i|a}\|_2$. Applying the triangle inequality once more and noting that

$$\|\widehat{R}_{i|a}\|_2 \le \|R_{i|a}\|_2 + \|\Delta_i\|_2 \le \|R_{i|a}\|_2 + \delta,$$

gives

$$\begin{aligned} \|K_{ij} - \widehat{K}_{ij}\|_F &\le 2\delta \|R_{j|a}\|_2 + 2\delta \|\widehat{R}_{i|a}\|_2 \\ &\le 2\delta \big(\|R_{i|a}\|_2 + \|R_{j|a}\|_2 + \delta \big). \end{aligned}$$

Combining this result with the Eckart–Young bound proves Eq. (58). Rewriting it as $2\delta^2 + 2S_{ij}\delta - \tau_{8r}(K_{ij}) \ge 0$ and taking the nonnegative root of the quadratic inequality yields Eq. (59). □

D.20 Global scale–correlation decomposition of SPD matrices and its Jacobian

Proof. A positive-definite matrix has a strictly positive diagonal, so the forward map in Section 3 is well defined. Eq. (61) is plainly its inverse, and positive diagonal congruence preserves positive definiteness. Hence the map is a smooth global diffeomorphism. Ordering the diagonal elements before the off-diagonal elements makes $\partial(v, c)/\partial \operatorname{vech}(S)$ block lower triangular. Its two diagonal blocks satisfy

$$\frac{\partial v_i}{\partial S_{ii}} = S_{ii}^{-1} = e^{-v_i}, \qquad \frac{\partial C_{ij}}{\partial S_{ij}} = (S_{ii}S_{jj})^{-1/2} = e^{-(v_i+v_j)/2}. \tag{103}$$

Therefore,

$$\log \left| \det \frac{\partial(v, c)}{\partial \operatorname{vech}(S)} \right| = -\sum_i v_i - \frac{1}{2} \sum_{i<j} (v_i + v_j) = -\frac{d+1}{2} \sum_i v_i. \tag{104}$$

The inverse Jacobian is its reciprocal. □

E Supplementary Experiments

This appendix includes the complete data/protocol summary and fixed-geometry CorNet results supporting the cross-domain benchmark in the main paper. The remaining studies examine scale-aware modeling and temporal generation, extend the main strict-nesting result to sampling and coverage, and examine ablation, structured capacity, computational complexity, and interpretability under controlled protocols. Unless stated otherwise, confidence intervals use paired bootstrap resampling, with random seeds or datasets as the sampling units (Efron & Tibshirani, 1993). We apply Holm correction within each preregistered comparison family (Demšar, 2006; Holm, 1979), and probability quality is measured by posterior NLL, the Brier score (Brier, 1950), and 15-bin ECE (Guo et al., 2017).

E.1 Scale-Quotient Robustness to Positive Diagonal Gains

Channel gains can vary across subjects, devices, and acquisition conditions while preserving correlation structure. We test whether CorTG's positive-diagonal invariance is retained numerically and compare its robustness with SPD baselines under increasing gain shift.

Data are generated from three class-specific correlation centers. Each sample contains $d = 6$ variables and 32 observations, with $100/33/100$ train/validation/test windows per class. Training and validation gains are log-normal with standard deviation 0.1. At test time, the same held-out windows receive sample- and channel-specific gains with $\sigma_g \in \{0, 0.5, 1, 1.5, 2\}$. Within each seed, all strengths scale a common standard-normal log-gain direction, yielding paired perturbation trajectories. CorTG is trained for at most 120 epochs, selected by validation posterior NLL, with patience 20.

We apply the congruence-equivariant shrinkage $R_\lambda(S) = S + \lambda D_S$ from Eq. (64) before projection, since the usual λI ridge would break the invariance. The primary metric is the normalized area under the accuracy curve (robustness AUC),

$$\mathrm{AUC}_{\mathrm{rob}} = \frac{1}{\sigma_{\max} - \sigma_{\min}} \int_{\sigma_{\min}}^{\sigma_{\max}} \mathrm{Acc}(\sigma)\, \mathrm{d}\sigma, \tag{105}$$

and endpoint drop $\mathrm{Acc}(0) - \mathrm{Acc}(2)$. Both are computed within seed and aggregated over ten seeds. Baselines are log-Euclidean SPD classification (Arsigny et al., 2007), affine-invariant Riemannian metric minimum distance to mean (AIRM-MDM), and AIRM tangent-space logistic regression (AIRM-TS-LR) (Pennec et al., 2006; Barachant et al., 2010; 2012).

Table 4: Results over ten seeds for the channel-gain shift experiment (%, mean ± standard deviation). $\mathrm{AUC}_{\mathrm{rob}}$ is the normalized area under the accuracy curve in Eq. (105). A lower endpoint drop is better. Bold indicates the best value in each column.

Method	$\mathrm{Acc}(0)$	$\mathrm{AUC}_{\mathrm{rob}}$	Endpoint drop
CorTG	87.33 ± 2.29	$\mathbf{87.33 \pm 2.29}$	$\mathbf{0.00 \pm 0.00}$
LogEuclidean-SPD	89.40 ± 2.07	52.01 ± 2.00	53.73 ± 2.93
AIRM-MDM	$\mathbf{93.73 \pm 1.25}$	73.48 ± 2.34	42.40 ± 3.18
AIRM-TS-LR	93.30 ± 1.29	73.31 ± 3.59	42.63 ± 4.43

Table 4 shows that CorTG preserves accuracy and macro area under the receiver operating characteristic curve (ROC-AUC) at every gain level and seed, giving zero endpoint drop. At $\sigma_g = 2$, LogEuclidean-SPD, AIRM-MDM, and AIRM-TS-LR fall to 35.67%, 51.33%, and 50.67% accuracy, whereas CorTG's paired $\mathrm{AUC}_{\mathrm{rob}}$ advantages are 35.32, 13.86, and 14.02 points. Their 20,000-sample paired-bootstrap 95% confidence intervals (CIs) are $[33.81, 37.05]$, $[12.33, 15.64]$, and $[12.80, 15.35]$, with positive differences in all seeds. Macro ROC-AUC similarly remains 95.95% for CorTG but ends at 52.23%, 69.24%, and 68.84% for the baselines.

A coefficient-level audit confirms the mechanism. Across 300 test windows per seed, five gain levels, and ten seeds, the maximum absolute correlation error is 6.66×10^{-16}. At $\sigma_g = 2$, the mean, 95th-, and 99th-percentile errors are 5.50×10^{-17}, 1.67×10^{-16}, and 2.62×10^{-16}, well below the 10^{-9} tolerance. The flat curves therefore reflect removal of the positive-diagonal action, not weak perturbations or regenerated samples.

E.2 Task-Level Benefits of Scale-Quotient Modeling

We next compare correlation-only and joint scale–correlation modeling on two three-class problems: correlation signal with nuisance scale, and scale signal with identical class correlations.

Each covariance uses $d = 5$ variables and 96 observations, and each class has $120/40/120$ train/validation/test samples over ten paired seeds. The correlation-signal test set receives independent channel gains. In the scale-signal task, classes share identity correlation and differ only in marginal scale, so any $\mathrm{Corr}(S)$-only method should operate at three-class chance. We compare CorTG with Joint-CorTG, which retains scale, and with AIRM-MDM, tangent-space linear and quadratic discriminant analysis (TS-LDA and TS-QDA) (Pennec et al., 2006; Barachant et al., 2010; 2012), and the homoscedastic and heteroscedastic wrapped discriminant analysis variants (Ho-WDA and He-WDA) (De Surrel et al., 2025). Both proposed models use at most 400 epochs, validation posterior-NLL selection, and patience 20.

All 40 learned models pass the prespecified convergence and probability audits. Ho-/He-WDA use default SteepestDescent with limits of 10,000 iterations and 1200 s, and every run reaches the iteration limit. Their 20 correlation-signal probability outputs underflow natively and are recomputed with an equivalent log-domain expression, while the 20 scale-signal outputs pass directly.

Table 5: Balanced accuracy over ten seeds (%, mean ± standard deviation). Bold indicates the best value in each setting.

Method	Correlation-signal	Scale-signal
CorTG	$\mathbf{99.72 \pm 0.29}$	33.25 ± 2.45
Joint-CorTG	98.97 ± 0.68	92.39 ± 1.78
AIRM-MDM	82.33 ± 2.47	$\mathbf{100.00 \pm 0.00}$
TS-LDA	86.75 ± 2.20	$\mathbf{100.00 \pm 0.00}$
TS-QDA	66.53 ± 2.98	$\mathbf{100.00 \pm 0.00}$
Official Ho-WDA	83.42 ± 2.08	$\mathbf{100.00 \pm 0.00}$
Official He-WDA	74.17 ± 5.28	$\mathbf{100.00 \pm 0.00}$

On correlation signal, CorTG reaches $99.72 \pm 0.29\%$ balanced accuracy, versus $83.42 \pm 2.08\%$ and $74.17 \pm 5.28\%$ for Ho-WDA and He-WDA. The paired gains are 16.31 and 25.56 points (bootstrap 95% CIs $[15.17, 17.67]$ and $[22.11, 28.33]$), positive in all ten seeds, with $p_{\mathrm{Holm}} = 0.0059$ for both two-sided Wilcoxon tests. Joint-CorTG reaches $98.97 \pm 0.68\%$. Thus, the invariance in Section E.1 translates to task-level robustness when scale is nuisance.

On scale signal, correlation-only CorTG obtains 33.25%, -0.08 points from chance (bootstrap 95% CI $[-1.56, 1.31]$). A two one-sided tests (TOST) procedure (Schuirmann, 1987) with the prespecified ± 5-point margin gives $p = 6.67 \times 10^{-5}$, as expected when class information lies in marginal scale.

Restoring scale raises Joint-CorTG to 92.39%, a 59.14-point gain (CI $[57.47, 60.83]$), positive in all seeds with $p_{\mathrm{Holm}} = 0.0059$. Scale-aware baselines reach 100%. These results support the quotient model for nuisance-scale tasks and the joint extension when marginal variance is predictive.

E.3 Cross-Domain Benchmark Details

Table 6 summarizes the five datasets, matrix construction, and outer protocols used in Section 4.1, while Table 7 reports all five fixed CorNet geometries. The main comparison uses the best geometry for each dataset and metric.

Hyperspectral protocol. For Indian Pines, Pavia University, and Salinas, we follow the matrix construction of the source benchmark (De Surrel et al., 2025). PCA with five components is estimated once from the complete image without using class labels. A 5×5 SCM is then computed for every valid 25×25 window with unit stride. Each window receives the majority label of its source pixels, with background label 0 retained in the label set. This construction yields 14,641, 185,176, and 94,184 matrices and 12, 6, and 17 effective classes for Indian Pines, Pavia University, and Salinas, respectively. The resulting matrix counts correspond to

$$(145-24)^2, \qquad (610-24)(340-24), \qquad (512-24)(217-24).$$

Following the source protocol, random five-fold partitions are formed over the derived matrices with the prescribed random seed. Because unit-stride windows can overlap across folds, we keep the matrix construction, fold assignments, and evaluation procedure identical for every method, thereby providing a controlled protocol-matched comparison (Liang et al., 2017).

Table 6: Data and protocol overview for the cross-domain matrix benchmarks. N is the number of matrices entering outer evaluation after invalid samples are removed, d is the matrix order, and $m = d(d-1)/2$ is the dimension of the correlation manifold. N/d indicates the sample-size range relative to matrix order. For the hyperspectral datasets, PCA-5 and stride-one 25×25 local SCM construction follow the source benchmark. PCA and SCM denote principal component analysis and sample covariance matrix, respectively.

Dataset	Domain	Protocol	Dim. $d \times d\,(m)$	# matrices N	# classes	N/d	Matrix construction
Indian Pines	Hyperspectral	random5	5×5 (10)	14,641	12	2,928	PCA-5, 25×25 SCM
Pavia University	Hyperspectral	random5	5×5 (10)	185,176	6	37,035	PCA-5, 25×25 SCM
Salinas	Hyperspectral	random5	5×5 (10)	94,184	17	18,837	PCA-5, 25×25 SCM
Textile	Texture	random5	10×10 (45)	16,000	2	1,600	10-filter feature SCM
BreizhCrops	Multispectral	random5	13×13 (78)	122,048	6	9,388	temporal feature covariance + OAS

Table 7: Complete results for the single-matrix CorNet (Chen et al., 2026) under five fixed correlation geometries. The aggregation procedure matches Table 1. Bold marks the best result for each dataset and metric, determined from the unrounded values.

Dataset	CorNet-ECM	CorNet-LECM	CorNet-OLM	CorNet-LSM	CorNet-PHCM
			(a) Accuracy (%) ↑		
Indian Pines	77.32 ± 0.53	77.62 ± 0.48	$\mathbf{80.50 \pm 0.37}$	80.42 ± 0.44	78.09 ± 0.34
Pavia University	91.12 ± 0.09	91.99 ± 0.11	91.79 ± 0.11	91.79 ± 0.11	$\mathbf{92.03 \pm 0.11}$
Salinas	75.39 ± 0.22	76.33 ± 0.17	77.78 ± 0.27	78.78 ± 0.19	$\mathbf{79.30 \pm 0.05}$
Textile	$\mathbf{89.53 \pm 0.26}$	85.72 ± 0.83	89.43 ± 0.36	88.56 ± 0.29	89.24 ± 0.31
BreizhCrops	66.39 ± 0.33	$\mathbf{67.17 \pm 0.31}$	65.26 ± 0.24	62.02 ± 0.30	62.92 ± 0.37
			(b) Posterior NLL ↓		
Indian Pines	0.682 ± 0.011	0.682 ± 0.013	0.618 ± 0.011	$\mathbf{0.604 \pm 0.012}$	0.682 ± 0.013
Pavia University	0.243 ± 0.002	0.232 ± 0.003	0.224 ± 0.002	0.225 ± 0.002	$\mathbf{0.213 \pm 0.003}$
Salinas	0.696 ± 0.004	0.661 ± 0.003	0.575 ± 0.004	0.561 ± 0.004	$\mathbf{0.547 \pm 0.004}$
Textile	$\mathbf{0.269 \pm 0.007}$	0.342 ± 0.008	0.286 ± 0.003	0.305 ± 0.004	0.281 ± 0.006
BreizhCrops	0.784 ± 0.004	$\mathbf{0.766 \pm 0.005}$	0.823 ± 0.003	0.911 ± 0.002	0.904 ± 0.003
			(c) ECE ↓		
Indian Pines	0.061 ± 0.006	0.062 ± 0.006	0.074 ± 0.002	$\mathbf{0.058 \pm 0.003}$	0.073 ± 0.006
Pavia University	0.021 ± 0.000	0.015 ± 0.001	0.014 ± 0.001	$\mathbf{0.014 \pm 0.001}$	0.015 ± 0.001
Salinas	0.020 ± 0.001	0.018 ± 0.002	0.019 ± 0.003	$\mathbf{0.011 \pm 0.002}$	0.020 ± 0.001
Textile	$\mathbf{0.053 \pm 0.003}$	0.067 ± 0.007	0.074 ± 0.009	0.078 ± 0.003	0.063 ± 0.004
BreizhCrops	0.045 ± 0.003	$\mathbf{0.045 \pm 0.003}$	0.058 ± 0.002	0.074 ± 0.004	0.074 ± 0.005

E.4 Strictly Causal Temporal Likelihood and Free Rollout

We test strict causality, one-step likelihood, and free rollout on a two-class nonlinear autoregressive process in off-log coordinates. Each class contains 240 length-10 sequences of 4×4 correlation matrices, split $60\%/20\%/20\%$ over ten seeds. Temporal-CorTG and a right-shifted causal gated recurrent unit Gaussian (Causal-GRU-Gaussian) (Cho et al., 2014) use rank-2 conditional DPLR heads, whereas Independent-Gaussian models time steps independently. Training and selection use class-conditional NLL. For a common measure, NLL is converted through the off-log Jacobian to Lebesgue measure over independent correlation coefficients and normalized by sequence length. Free-rollout quality is measured by unbiased radial-basis-function maximum mean discrepancy (RBF-MMD)^2 on full sequence vectors (Gretton et al., 2012), with class-specific bandwidths estimated from training data and fixed across models.

Table 8: Results over ten seeds for the strictly causal temporal experiment (mean ± standard deviation). NLL is measured with respect to the independent-correlation-coefficient measure and normalized by time step. Rollout MMD^2 is first averaged over the two classes and then aggregated across seeds. Arrows indicate that lower is better.

Method	NLL/T ↓	rollout MMD^2 ↓	Accuracy (%)
Temporal-CorTG	-6.170 ± 0.065	$\mathbf{0.0013 \pm 0.0027}$	100.0 ± 0.0
Causal-GRU-Gaussian	$\mathbf{-6.214 \pm 0.065}$	0.0065 ± 0.0043	100.0 ± 0.0
Independent-Gaussian	-5.815 ± 0.088	–	100.0 ± 0.0

Perturbing current and future coordinates at every test position changes c_t by exactly 0 across all seeds, confirming that the conditioner accesses neither z_t nor $z_{>t}$.

Table 8 shows that Temporal-CorTG lowers per-step NLL by 0.355 ± 0.042 relative to Independent-Gaussian, while Causal-GRU-Gaussian is 0.043 ± 0.036 lower than Temporal-CorTG. Both paired Wilcoxon comparisons give $p = 0.0020$ and $p_{\mathrm{Holm}} = 0.0039$, establishing the value of historical conditioning while showing a small one-step advantage for the GRU parameterization.

In free rollout, Temporal-CorTG lowers MMD^2 by 0.0052 ± 0.0051 relative to Causal-GRU-Gaussian ($p = 0.0098$). We retain raw unbiased MMD^2 estimates, which may be slightly negative, and pair comparisons across classes and seeds. Together with the one-step NLL results, this shows complementary strengths in immediate density prediction and long-horizon distribution matching. All methods reach 100% accuracy, so the comparison isolates dynamic probability structure.

E.5 Joint Evaluation of Density, Sampling, and Probability Regions

We extend the strictly nested study in Section 4.2 from test NLL to free samples and nominal probability regions.

We use the primary condition $(r, n, \gamma) = (2, 2000, 2)$: three rank-2 DPLR Gaussians passed through a shared four-layer invertible transport, with 2000 samples per class, $60\%/20\%/20\%$ splits, and ten paired seeds. A common identity initialization branches into frozen `continued_identity` and trainable `learned_flow`, with validation class-conditional NLL selecting between them. All seeds pass branching, convergence, and reconstruction checks and select learned flow. Besides test NLL under $\mathrm{dvol}_{g^0} = \mathrm{d}h$, we report class-averaged unbiased RBF-MMD^2 (Gretton et al., 2012), using training-estimated bandwidths shared across methods, and mean absolute coverage error over 50%, 90%, and 95% latent chi-squared regions.

Table 9: Recovery of density, samples, and probability regions under a strongly nonlinear true distribution (mean $\pm$ standard deviation over ten seeds). Lower NLL, unbiased MMD^2, and mean absolute coverage error are better. Finite-sample estimates of unbiased MMD^2 may be slightly negative. Bold indicates the best non-Oracle model.

Method	Test NLL ↓	MMD^2 ↓	Coverage error ↓
Oracle	11.443 ± 1.972	-0.000072 ± 0.000610	0.00825 ± 0.00399
CorTG	$\mathbf{11.592 \pm 2.008}$	$\mathbf{0.000094 \pm 0.000747}$	0.02192 ± 0.00581
Identity-DPLR	13.076 ± 1.966	0.001192 ± 0.001183	0.03728 ± 0.00685
CorTG-fixed	11.777 ± 2.014	0.000845 ± 0.001055	$\mathbf{0.01750 \pm 0.00619}$

Relative to the strictly nested Identity-DPLR, CorTG reduces NLL, MMD^2, and coverage error by 1.483 (bootstrap 95% CI $[1.299, 1.668]$), 1.10×10^{-3} ($[0.71, 1.44] \times 10^{-3}$), and 0.0154 ($[0.0112, 0.0195]$). All three one-sided Wilcoxon tests give $p_{\mathrm{Holm}} = 0.0088$ over nine comparisons. With head capacity fixed, transport therefore improves density fit, sample matching, and probability regions.

Against CorTG-fixed, CorTG reduces NLL by 0.185 (CI $[0.156, 0.215]$, $p_{\mathrm{Holm}} = 0.0088$) and MMD^2 by 7.51×10^{-4} (CI $[3.14 \times 10^{-4}, 1.20 \times 10^{-3}]$, $p_{\mathrm{Holm}} = 0.0137$). A full-covariance Gaussian therefore does not replace nonlinear reorganization of higher-order shape.

Coverage provides a complementary calibration check. CorTG-fixed has error 0.01750 ± 0.00619, versus 0.02192 ± 0.00581 for CorTG, with a CI of $[-0.00695, -0.00208]$ for "CorTG-fixed minus CorTG." CorTG's empirical coverages are 47.91%, 87.68%, and 93.25% at nominal 50%, 90%, and 95%. Together, the results show consistent gains over the matched low-rank head, improved NLL and MMD^2 over the full-covariance baseline, and strong probability-region coverage from both models.

E.6 Single-Factor Ablation

We ablate one factor at a time on a known nonlinear distribution: three rank-2 DPLR Gaussians mapped through a shared four-layer transport to $d = 5$ correlations. Each class has 350 samples with $60\%/20\%/20\%$ splits over ten seeds. Learnable variants modify a depth-6, rank-2 LU-CorTG reference while holding data, class-conditional NLL, optimization, and validation-NLL selection fixed. Component tests are paired by seed, with Holm correction over six two-sided Wilcoxon tests.

Relative to deep LU, ZCA normalization, Jacobian metric regularization, and their combination reduce coordinate NLL by 0.263 ± 0.251, 0.175 ± 0.133, and 0.382 ± 0.233, respectively, and all give $p_{\mathrm{Holm}} = 0.0117$. Their combination yields the best NLL and accuracy among learnable transports.

Changing the DPLR head from $r = 2$ to $r = 0$ changes NLL by only 0.002 ± 0.062 ($p_{\mathrm{Holm}} = 0.625$). By contrast, shallow LU is 0.284 ± 0.173 lower than deep LU, whereas Householder mixing is 0.366 ± 0.272 higher, and both give $p_{\mathrm{Holm}} = 0.0117$. No accuracy difference among learnable variants is significant at uncorrected 0.05, indicating that these choices primarily affect probability shape.

Table 10: Single-factor ablation on a nonlinear correlation distribution (mean ± standard deviation over ten seeds). NLL is measured with respect to the global off-log base-coordinate measure. Parameter counts include the Gaussian head for each class. Bold indicates the overall best result, and underlining indicates the best learnable transport.

Configuration	Parameters	Coordinate NLL ↓	Accuracy (%) ↑
CorTG-fixed	197	$\mathbf{8.561 \pm 0.966}$	$\mathbf{99.05 \pm 1.08}$
Deep LU, $r = 2$	112.9k	10.264 ± 1.020	95.29 ± 1.95
Shallow LU, $r = 2$	18.9k	9.979 ± 0.935	95.33 ± 2.02
Deep LU, $r = 0$	112.9k	10.266 ± 0.993	95.71 ± 2.06
Deep LU, $r = 2$, ZCA	112.9k	10.001 ± 0.950	95.90 ± 2.21
Deep LU, $r = 2$, metric reg.	112.9k	10.089 ± 1.031	95.86 ± 1.81
Deep Householder, $r = 2$	111.9k	10.630 ± 1.029	93.81 ± 2.89
Deep LU, $r = 2$, ZCA + metric reg.	112.9k	$\underline{9.882 \pm 0.950}$	$\underline{96.10 \pm 2.05}$

CorTG-fixed, the 197-parameter full-covariance model, attains 8.561 ± 0.966 NLL and $99.05 \pm 1.08\%$ accuracy, with NLL 1.321 below the best learned transport. Against deep LU, its paired advantages are 1.703 ± 0.578 NLL and 3.76 ± 1.04 accuracy points ($p = 0.0020$ for both). At this sample size, the result highlights the strength of a compact fixed-coordinate reference and supports ZCA and metric regularization as effective choices for the learnable transport.

Numerically, unit-diagonal residuals stay below 9.9×10^{-8} and off-log round-trip error below 4.5×10^{-15}. On an RTX 5090 with batch size 32, median batch latencies at $d = 3, 5, 8, 12$ are 0.592, 1.219, 2.928, and 6.502 s. Forward coordinates, transport, and DPLR density operate at millisecond scale, whereas Newton-based inverse off-log reconstruction dominates runtime.

E.7 Noncommutative Rank Obstruction and DPLR Capacity Selection

We test whether the noncommutative lower bound in Eq. (23) predicts the minimum class-specific DPLR rank remaining after a shared affine transport.

For each of ten seeds, four rank-2 DPLR Gaussians in $m = 15$ off-log coordinates are mapped through a common unknown invertible linear transform. Each class contains 500 samples, divided into $60\%/20\%/20\%$ training, validation, and test splits. The candidate models share the same matrix-exponential affine transport and differ only in the head rank $r \in \{0, 1, 2, 4\}$. Training maximizes the class-conditional likelihood, and checkpoints are selected by validation NLL. Convergence is declared when the improvement remains below 10^{-3} for 50 epochs within a budget of 3000 epochs. Coordinate NLL is measured with respect to $\mathrm{dvol}_{g^0} = \mathrm{d}h$.

Table 11: DPLR rank sweep under a shared affine transport (mean ± standard deviation over ten seeds). The theoretical necessary lower bound is $r_{\min} = 2$. Parameter counts include the shared affine transport and four class-conditional Gaussian heads. Epoch denotes the mean training epoch at which the common convergence criterion is satisfied.

Rank r	Parameters	Coordinate NLL ↓	Accuracy (%) ↑	Epoch
0	360	13.769 ± 0.465	79.53 ± 4.93	1225.0
1	420	13.589 ± 0.489	81.88 ± 4.48	1192.3
2	480	$\mathbf{13.517 \pm 0.458}$	82.93 ± 3.97	834.2
4	600	13.543 ± 0.467	82.90 ± 4.19	707.7

After symmetric anchor whitening, the population covariances give $\kappa_{\mathrm{nc}} = 12$, while the training-fold empirical covariances give $\kappa_{\mathrm{nc}} = 14$, for every seed. Both diagnostics yield the same necessary rank because

$$\left\lceil \frac{12}{8} \right\rceil = \left\lceil \frac{14}{8} \right\rceil = 2.$$

The resulting even ranks are also consistent with the skew-symmetry of the whitened commutators. Using only training-fold empirical covariances, the maximum τ_{8r} is 25.618 ± 8.165 at $r = 0$ and 0.859 ± 0.157 at $r = 1$. These values imply relative-operator Frobenius-error lower bounds of

0.727 ± 0.125 and 0.0334 ± 0.0074, respectively. For $r \geq 2$, $8r \geq m$ and the spectral tail is zero. Together with the explicit rank-2 data-generating model, the corrected diagnostic identifies $r = 2$ as the smallest realizable rank.

Taking $r = 2$ as the reference, the coordinate-NLL gaps are 0.252 at $r = 0$ (bootstrap 95% CI $[0.216, 0.286]$) and 0.072 at $r = 1$ ($[0.040, 0.105]$). Both one-sided Wilcoxon tests remain significant after Holm correction, with $p_{\mathrm{Holm}} = 0.0020$. At $r = 4$, the NLL is only 0.026 higher than at $r = 2$ (CI $[0.011, 0.042]$, $p = 0.0195$), which lies entirely within the prespecified ± 0.05 equivalence range. Accuracy follows the same transition, increasing from 79.53% at $r = 0$ to 82.93% at $r = 2$ and remaining at 82.90% for $r = 4$.

All 40 runs satisfy the common convergence criterion between epochs 475 and 2291, ruling out optimization-budget failure as the source of the low-rank gap. The mean convergence epoch decreases from 1225.0 and 1192.3 for $r = 0$ and $r = 1$ to 834.2 and 707.7 for $r = 2$ and $r = 4$, respectively. The empirical transition therefore occurs at the rank predicted by the corrected commutator diagnostic, supporting its use for selecting DPLR capacity.

E.8 Mechanisms of Invertible Transport and Visual Interpretation of the Learned Geometry

We isolate how transport affects density, posteriors, and local geometry on a controlled task with known generative coordinates. Three $m = 10$ rank-2 DPLR Gaussians are mapped to $d = 5$ off-log coordinates by a shared four-layer transport at $\gamma = 2$. Each class has 1000 samples with $60\%/20\%/20\%$ splits over ten seeds. Frozen `equal-budget identity` and trainable `learned transport` branches share the moment initialization, class-conditional NLL, mini-batches, optimizer, budget, and patience 30. All runs pass checks for branch identity, invertible cycles, autodifferentiated log determinants, density decomposition, and posterior Jacobian cancellation.

Diagnostics cover four levels: radial Kolmogorov–Smirnov (KS) statistics and 95% coverage for Gaussianity, Fisher trace ratio, Silhouette (Rousseeuw, 1987), and true-class margin for class organization, exact Jacobian spectra, $\log \kappa(J_T)$, $\log |\det J_T|$, and density decomposition on 128 fixed probes per seed, and test R^2 plus distance Spearman after a training-fold affine alignment to the oracle representative. The last comparison removes the global affine gauge while retaining nonlinear structure.

Table 12: Paired results over ten seeds in the controlled interpretability experiment (mean $\pm$ standard deviation). "Favorable change" is defined so that positive values always indicate improvement. Changes in accuracy and coverage error are measured in percentage points. p_{Holm} is obtained from two-sided Wilcoxon tests paired by seed, with Holm correction over eight preregistered global metrics.

Metric	Equal-budget identity	Learned transport	Favorable change	p_{Holm}
Test class-conditional NLL ↓	13.036 ± 1.999	11.628 ± 2.025	1.408 ± 0.288	0.0156
Accuracy (%) ↑	78.32 ± 6.21	82.98 ± 5.41	4.67 ± 1.87	0.0156
Mean radial KS ↓	0.0992 ± 0.0135	0.0655 ± 0.0146	0.0337 ± 0.0172	0.0156
95% absolute coverage error (pp) ↓	3.77 ± 1.71	2.37 ± 1.13	1.40 ± 1.42	0.0430
Fisher trace ratio ↑	0.3185 ± 0.3366	0.3243 ± 0.3367	0.0058 ± 0.0098	0.1309
Silhouette ↑	0.0562 ± 0.0522	0.0601 ± 0.0524	0.00384 ± 0.00367	0.0410
Affine-gauge-aligned oracle R^2 ↑	0.9553 ± 0.0170	0.9748 ± 0.0128	0.0195 ± 0.0087	0.0156
Aligned oracle distance Spearman ↑	0.9530 ± 0.0149	0.9775 ± 0.0093	0.0245 ± 0.0090	0.0156

Table 12 and Figure 3 show all-seed improvements of 1.408 nats in NLL and 4.67 accuracy points, with absolute paired rank-biserial correlation 1 for both. Radial KS falls from 0.0992 to 0.0655, and 95% coverage error from 3.77 to 2.37 points. Silhouette also improves significantly after correction, while Fisher trace ratio remains similar. The combined results show that transport improves within-class shape, relative evidence, and local class organization.

Across the 128 probes per seed, latent true-class Gaussian log-density changes by -0.083 ± 0.142 nats, while the shared log-Jacobian contributes 1.507 ± 0.336, yielding a fixed-volume density gain of 1.424 ± 0.271. The density gain therefore comes mainly from volume reallocation. Because the

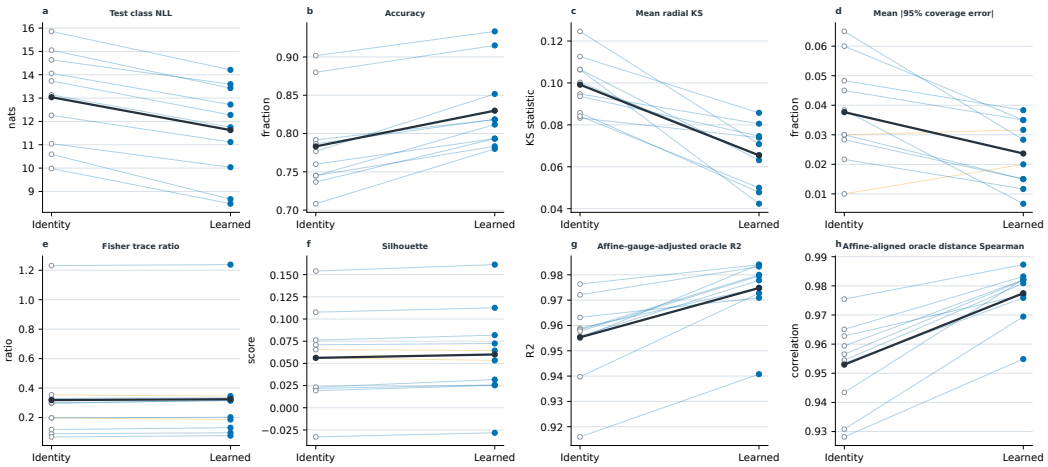


Figure 3: Paired changes over ten seeds. Light blue and gold denote positive and negative signed changes, and black connects means. Transport improves NLL, accuracy, radial KS, and both gauge-aligned recovery metrics for every seed. Low-dimensional views are descriptive, while inference uses original-dimensional metrics.

shared Jacobian cancels in posterior ratios, classification changes must arise from relative Gaussian evidence: true-class margin increases by 0.280 ± 0.428 nats, posterior NLL falls from 0.552 ± 0.157 to 0.406 ± 0.122, and ECE from 0.0642 ± 0.0257 to 0.0348 ± 0.0125.

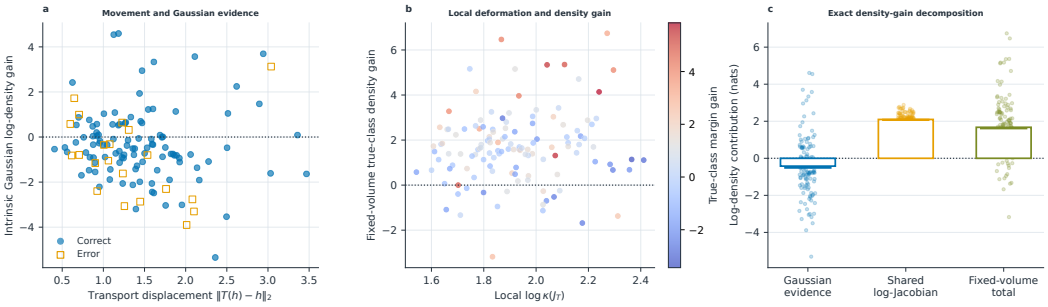


Figure 4: Local mechanisms for seed 41, selected because its NLL gain (1.455) is closest to the median (1.422). Left: displacement and latent-evidence change. Middle: $\log \kappa(J_T)$, fixed-volume density gain, and margin change. Right: density-gain decomposition into Gaussian evidence and shared log-Jacobian.

Figure 4 shows heterogeneous local behavior. Across seeds, margin gain correlates 0.226 ± 0.112 with fixed-volume density gain, correlations of condition number and log-Jacobian with that gain are 0.081 ± 0.096 and 0.080 ± 0.117, and displacement correlates -0.012 ± 0.111 with latent-evidence gain. No single local distortion statistic explains the effect. Globally, gauge-aligned oracle R^2 and distance Spearman improve for every seed, indicating closer agreement with the chosen generative representative without implying geometric uniqueness under the remaining nonlinear gauge freedom.

E.9 Neurophysiological Evidence in Cross-Session Brain–Computer Interfaces

We analyze cross-session motor-imagery evidence and feature dependence by coordinate replacement. Subject 1 of BCI Competition IV 2a (`BNCI2014_001`) is processed with the Mother of All

BCI Benchmarks (MOABB) framework (Tangermann et al., 2012; Jayaram & Barachant, 2018): left/right-hand trials, 22 channels, 8–32 Hz filtering, and 128 Hz resampling. Session `0train` supplies training and validation, and all 144 `1test` trials are held out. Each trial yields 13 OAS-covariance windows (Chen et al., 2010) of 1.0 s at 0.25 s stride. A rank-2, depth-4 CorTG is evaluated over ten seeds.

We apply 32-step integrated gradients (IG) (Sundararajan et al., 2017) to true-versus-alternative posterior log-odds, using the training mean in off-log coordinates as baseline. Edge attribution is split equally between endpoints, and each class contributes 24 correct trials. Replacements remain valid correlation matrices through the analytic inverse map. We replace the top 5%, 10%, 20%, and 40% of edges and compare with 20 size-matched random replacements. Performance is $90.21 \pm 2.28\%$ accuracy, $97.54 \pm 0.79\%$ ROC-AUC, and 0.334 ± 0.044 posterior NLL. Maximum IG completeness residual is 2.81×10^{-4}, the 95th-percentile relative residual is below 2.38×10^{-4}, and edge-to-node allocation error is below 7.11×10^{-15}.

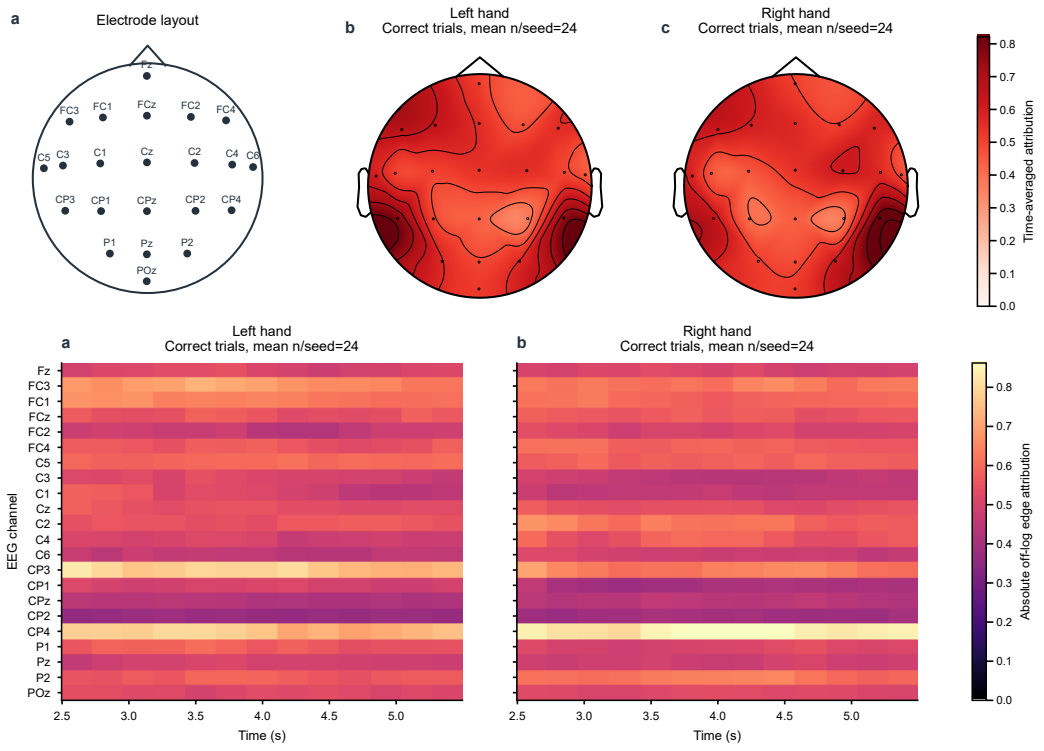


Figure 5: Class-level CorTG attribution. Top: electrode layout and time-averaged scalp maps for correct trials. Bottom: channel–time absolute off-log IG. Attributions are aggregated within seed and then across ten seeds, using a shared color scale and 24 correct trials per class and seed.

Figure 5 emphasizes central–parietal structure with class-dependent weighting. Left-hand imagery peaks at CP3 (0.766) and CP4 (0.751), whereas right-hand imagery peaks at CP4 (0.837) and CP3 (0.632). Evidence persists across adjacent windows. Across-seed Spearman stability is 0.615 ± 0.092 and 0.652 ± 0.105 for left and right imagery, while within-seed cross-class correlation is 0.796 ± 0.058. The classes therefore share a stable sensorimotor scaffold and differ primarily in its weighting.

Correct and incorrect channel–time maps correlate at 0.867 ± 0.043 (Figure 6), suggesting reweighting of a common scaffold rather than a separate error mechanism. The more diffuse error maps also have fewer trials (mean 14.1, range 9–20, versus 48 correct). Dominant positive edges are CP1–CP3 (0.169), P2–Pz (0.127), and CP3–FC3 (0.122) for left imagery, and P1–CP3 (0.157), P2–CP4 (0.107), and CP3–FC3 (0.105) for right imagery. The shared CP3–FC3 edge and class-specific rankings are consistent with a common sensorimotor scaffold under task-dependent reweighting.

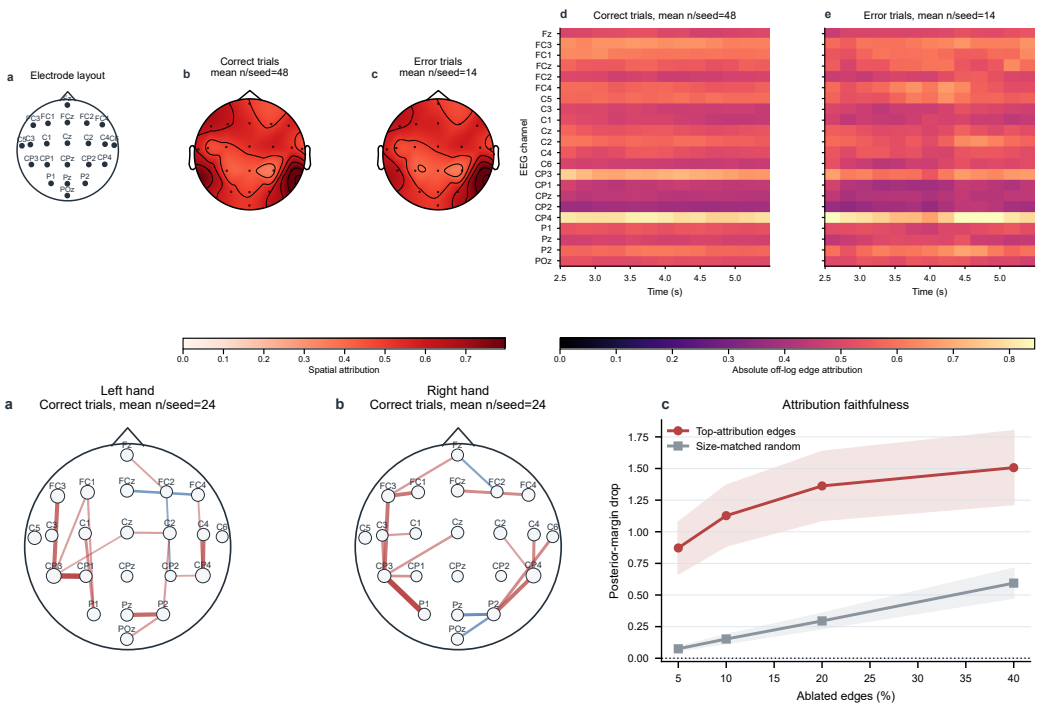


Figure 6: Correctness stratification and faithfulness. Top: scalp maps and channel–time IG for correct and incorrect trials (mean 48 and 14.1 per seed, with an incorrect range of 9–20). Bottom: top-16 signed edges for each class and posterior-margin loss after attribution-ranked versus size-matched random replacement. Red/blue indicates positive/negative margin contribution, and node area denotes absolute attribution. Curves show ten-seed means with Student-t 95% CIs.

Replacing the top 5%, 10%, 20%, and 40% of edges lowers the true-class margin by an additional 0.797 (CI $[0.606, 0.988]$), 0.975 ($[0.769, 1.181]$), 1.067 ($[0.856, 1.279]$), and 0.913 ($[0.738, 1.088]$) relative to matched random replacement. The effect is larger in every seed at every proportion, and one-sided exact Wilcoxon tests give $p_{\mathrm{Holm}} = 0.0039$ over four comparisons. The rankings therefore capture model-relevant dependencies rather than visualization artifacts.

The attribution maps consistently emphasize central–parietal correlation structure, and replacing the highest-ranked edges perturbs the model more than matched random replacements. These interventions establish model-level feature dependence, not biological causality. Larger multisubject studies are needed to determine whether the same patterns generalize beyond this subject and session pair.

E.10 Metric-Dependent Geodesic Projection on the Correlation Manifold

To show how coordinate choice changes geometric operations on $\mathcal{C}_3$, we extend the CorNet comparison with an CorTG panel (Chen et al., 2026). Every panel uses the same nine full-rank correlation matrices and reference curve. We compare ambient off-diagonal chords with the flat geodesics induced by OLM, LSM, and CorTG (Archakov & Hansen, 2021; Thanwerdas, 2024). In each global chart, a sample is matched to the nearest of 120 uniformly sampled curve points, joined to it by a straight segment, and mapped back to $\mathcal{C}_3$. Only the chart, induced metric, and nearest-point assignment differ across panels.

The CorTG panel composes the off-log chart with a fixed nonidentity transport: invertible LU affine mixing followed by smooth additive coupling with strength 0.85. Keeping this map fixed isolates its geometric effect from optimization. At zero strength, CorTG reduces exactly to OLM, with maximum coordinate discrepancy 0 on the audited samples. Every visualized matrix has unit diagonal and remains positive definite, with minimum eigenvalues of 0.506 on the 28×28 surface section and 0.687 across the nine samples.

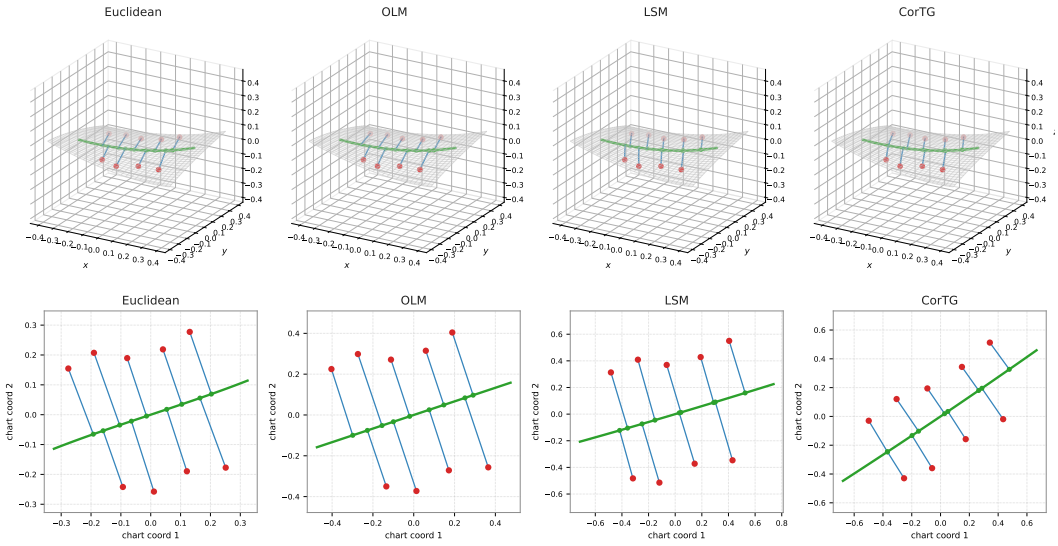


Figure 7: Metric-dependent projection of common matrices onto a fixed reference curve in $\mathcal{C}_3$. Top: a two-dimensional OLM-parameterized surface section (grey), the reference curve (green), samples (red), and projection paths (blue) under each geometry. Bottom: the same objects in the first two coordinates of each chart, with chart-specific axes.

Figure 7 shows that the metric changes both the selected projection points and the manifold paths, even though the inputs are identical. Ambient, OLM, and LSM coordinates produce different scales and orientations, while the nonidentity CorTG map further reshapes the coordinate distribution and changes the nearest-point assignments. This provides a direct geometric view of the learned pull-back metric defined in Section 3, and Appendix C.3 gives its distance, geodesic, and Fréchet-mean operators.

E.11 Computational Complexity Details

Table 13 complements the controlled CPU benchmark in Section 4.3 with the leading asymptotic training and inference costs.

Table 13: Leading asymptotic costs. Here, m and q are the numbers of free coordinates in a correlation matrix and a symmetric matrix, respectively, while n_{ep}, n_{geo}, and n_{it} denote epochs, candidate geometries, and optimization iterations.

Method	Training $\mathcal{O}(\cdot)$	Inference $\mathcal{O}(\cdot)$
CorTG	$Nd^3 + n_{\mathrm{ep}}N\left[L(mh + h^2 + m^2) + C(mr^2 + r^3)\right]$	$d^3 + L(mh + h^2 + m^2) + C(mr^2 + r^3)$
CorNet	$n_{\mathrm{geo}}Nd^3 + n_{\mathrm{geo}}n_{\mathrm{ep}}NmC$	$d^3 + mC$
Ho-WDA	$n_{\mathrm{it}}(Nd^3 + Nq^2 + q^3)$	$C(d^3 + q^2)$
He-WDA	$n_{\mathrm{it}}(Nd^3 + Nq^2 + Cq^3)$	$C(d^3 + q^2)$
Exact GP	$N^3 + mN^2$	Nm mean, N^2 variance

Derivation. We count dense arithmetic operations and omit constants and lower-order terms. Let N be the number of training matrices, d the matrix order, $m = d(d-1)/2$ the off-log coordinate dimension, $q = d(d+1)/2$ the SPD tangent dimension, and C the number of classes. For CorTG, constructing and caching the off-log coordinates requires one eigendecomposition per matrix, giving $O(Nd^3)$. In one transport block, the conditioning network and dense mixing cost $O(mh + h^2 + m^2)$ for width h. Evaluating one rank-r DPLR Gaussian uses $O(mr^2 + r^3)$ through the Woodbury identity and an $r \times r$ factorization. Thus, L blocks and C class heads cost $O\left(L(mh + h^2 + m^2) + C(mr^2 + r^3)\right)$ per sample. Multiplication by N samples and n_{ep} epochs

gives the training entry in Table 13; backpropagation has the same leading order. At inference, the off-log map contributes $O(d^3)$ and the transport and class heads contribute the same per-sample term.

For CorNet, each of the n_{geo} candidate geometries constructs coordinates in $O(Nd^3)$ and trains a linear C-class model in $O(n_{\mathrm{ep}}NmC)$. Validation selects one geometry, so prediction costs $O(d^3 + mC)$ per sample. Ho-WDA uses one q-dimensional covariance factorization per optimization step, whereas He-WDA uses one per class; this gives the respective q^3 and Cq^3 terms. Their shared samplewise operations cost $O(Nd^3 + Nq^2)$ per iteration, and evaluation against C class models costs $O(C(d^3 + q^2))$. Finally, an exact GP factors an $N \times N$ kernel matrix in $O(N^3)$ and solves for m outputs in $O(mN^2)$. For one test point, its predictive mean costs $O(Nm)$ and its variance costs $O(N^2)$.

F Algorithm Pseudocode

This section gives implementation-level pseudocode for the complete CorTG pipeline. A static observation is treated as a sequence of length one. The transport routine below includes every invertible operation on the evaluated data path (including a frozen train-split ZCA gauge layer when it is used) and returns their accumulated log-determinant. In contrast, the temporal conditioner only parameterizes a conditional density and therefore contributes no change-of-variables term.

Algorithm 2 Exact CorTG likelihood and posterior evaluation

Input: Covariance sequence $S_{1:T}$ with positive diagonal; mode in $\{\textsc{static}, \textsc{temporal}\}$; class priors $\{\pi_k\}_{k=1}^{K}$; shrinkage ε; invertible transport T_θ; DPLR heads; optional strictly causal conditioner $\mathcal{A}_\omega$
Output: Class log-likelihoods $g_{1:K}$ and posterior probabilities $p_{1:K}$

for $t = 1, \dots, T$ **do**
  $D_t \leftarrow \mathrm{Diag}(\mathrm{diag}\, S_t)$
  $C_t \leftarrow (1-\varepsilon) D_t^{-1/2} S_t D_t^{-1/2} + \varepsilon I_d$
  $h_t \leftarrow \mathrm{hvec}(\mathrm{off}(\log C_t))$
  $(z_t, \ell_t) \leftarrow \textsc{ForwardAndLogDet}(T_\theta, h_t)$ ▷ $\ell_t = \log|\det JT_\theta(h_t)|$
end for
if mode is TEMPORAL **then**
  $c_{1:T} \leftarrow \mathcal{A}_\omega([\mathrm{BOS}, z_1, \dots, z_{T-1}])$
  assert c_t depends only on $z_{<t}$; disable dropout in the generative conditioner
end if
for $k = 1, \dots, K$ **do**
  $g_k \leftarrow 0$
  for $t = 1, \dots, T$ **do**
    if mode is STATIC **then**
      $(\mu_{t,k}, b_{t,k}, U_{t,k}) \leftarrow (\mu_k, b_k, U_k)$
    else
      $(\mu_{t,k}, b_{t,k}, U_{t,k}) \leftarrow (f_{\mu,k}(c_t), f_{b,k}(c_t), f_{U,k}(c_t))$
    end if
    $d_{t,k} \leftarrow \mathrm{softplus}(b_{t,k}) + \delta$
    $\Sigma_{t,k} \leftarrow \mathrm{Diag}(d_{t,k}) + U_{t,k} U_{t,k}^\top$
    $g_k \leftarrow g_k + \log \mathcal{N}(z_t; \mu_{t,k}, \Sigma_{t,k}) + \ell_t$ ▷ evaluate the DPLR term with Woodbury
  end for
end for
$\log p_k \leftarrow \log \pi_k + g_k - \mathrm{LSE}_j(\log \pi_j + g_j)$ for every k ▷ LSE denotes log-sum-exp
return $g_{1:K}, \exp(\log p_{1:K})$

Algorithm 3 Joint generative–discriminative training of CorTG

Input: Training set $\mathcal{D}_{\mathrm{tr}} = \{(S^{(n)}_{1:T_n}, y_n)\}_{n=1}^{N}$; validation set; weights λ, β, γ; optimizer and stopping rule
Output: Best validated parameters $\Theta^{\star}$

Initialize the smooth invertible transport, DPLR heads, and (if needed) $\mathcal{A}_{\omega}$
Select ε and all preprocessing hyperparameters using training data only
for each training epoch **do**
  for each mini-batch $\mathcal{B} \subset \mathcal{D}_{\mathrm{tr}}$ **do**
    for all $(S^{(n)}_{1:T_n}, y_n) \in \mathcal{B}$ **do**
      $(g^{(n)}_{1:K}, p^{(n)}_{1:K}) \leftarrow \textsc{CorTGForward}(S^{(n)}_{1:T_n})$
    end for
    $\mathcal{L}_{\mathrm{disc}} \leftarrow -|\mathcal{B}|^{-1} \sum_{n \in \mathcal{B}} \log p^{(n)}_{y_n}$
    $\mathcal{L}_{\mathrm{gen}} \leftarrow -|\mathcal{B}|^{-1} \sum_{n \in \mathcal{B}} g^{(n)}_{y_n}$
    $\mathcal{R}_{\mathrm{scale}} \leftarrow \sum_{q,\ell} \|s_{q,\ell}\|_2^2$
    Estimate $\mathcal{R}_{\mathrm{metric}}$ from Eq. (15)
    $\mathcal{L} \leftarrow \mathcal{L}_{\mathrm{disc}} + \lambda \mathcal{L}_{\mathrm{gen}} + \beta \mathcal{R}_{\mathrm{scale}} + \gamma \mathcal{R}_{\mathrm{metric}}$
    Update all trainable parameters by back-propagating $\mathcal{L}$
  end for
  Update train-split gauge statistics, if used, and include their affine log-determinant
  Evaluate the pre-specified validation criterion and retain the best checkpoint
  if the stopping rule is met **then**
    break
  end if
end for
Freeze the selected checkpoint and all train-split preprocessing/gauge statistics
return $\Theta^{\star}$

Algorithm 4 Exact correlation-manifold sampling and optional covariance reconstruction

Input: Trained model $\Theta^\star$; mode in $\{\text{STATIC}, \text{TEMPORAL}\}$; class k; requested length T; Newton tolerance η; optional scale model p_v
Output: Correlation sequence $C_{1:T}$ and, when requested, covariance sequence $S_{1:T}$

$z_{<1} \leftarrow \varnothing$
for $t = 1, \dots, T$ **do**
  if mode is STATIC **then**
    $(\mu_{t,k}, d_{t,k}, U_{t,k}) \leftarrow (\mu_k, d_k, U_k)$
  else
    $c_t \leftarrow \mathcal{A}_\omega(\text{BOS}, z_{<t})$
    Obtain $(\mu_{t,k}, d_{t,k}, U_{t,k})$ from the class-k conditional head at c_t
  end if
  Draw $\xi_t \sim \mathcal{N}(0, I_m)$ and $\zeta_t \sim \mathcal{N}(0, I_r)$
  $z_t \leftarrow \mu_{t,k} + \text{Diag}(d_{t,k})^{1/2}\xi_t + U_{t,k}\zeta_t$
  $h_t \leftarrow T_\theta^{-1}(z_t)$; $H_t \leftarrow \text{hvec}^{-1}(h_t)$; initialize $\alpha \in \mathbb{R}^d$
  $M \leftarrow \text{Diag}(\alpha) + H_t$; $r \leftarrow \text{diag}(\exp M) - \mathbf{1}$
  while $\|r\|_\infty > \eta$ **do**
    for $j = 1, \dots, d$ **do**
      $A_{:j} \leftarrow \text{diag}(D \exp_M[E_{jj}])$
    end for
    Solve $A\Delta = -r$ and choose a damping factor $\rho \in (0, 1]$ by backtracking
    $\alpha \leftarrow \alpha + \rho\Delta$
    $M \leftarrow \text{Diag}(\alpha) + H_t$; $r \leftarrow \text{diag}(\exp M) - \mathbf{1}$
  end while
  $C_t \leftarrow \exp M$
  Append z_t to the causal history
  if the scale–correlation extension is requested **then**
    Draw $v_t \sim p_v(\cdot \mid C_t, k)$ and set $S_t \leftarrow \text{Diag}(e^{v_t/2})C_t\,\text{Diag}(e^{v_t/2})$
  end if
end for
return $C_{1:T}$ and, if constructed, $S_{1:T}$

Algorithms 2–3 evaluate likelihoods relative to the fixed reference volume $\text{dvol}_{g^0} = \text{d}h$. When a density with respect to the Lebesgue measure of the independent correlation entries is required, one additionally adds $\log|\det J_\phi(C_t)|$ at every time step. The inverse off-log solve in Algorithm 4 is required only for explicit manifold outputs, augmentation, or sampling, it is not part of the forward likelihood path.