THE WRISTBAND GAUSSIAN LOSS: DETERMINISTIC, COMPOSABLE LATENTS VIA A SPHERE–INTERVAL DECOMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce the *Wristband Gaussian Loss*, a deterministic batch regularizer for learning standard-Gaussian representations without per-step Gaussian reference sampling, KL penalties or an inner transport solve. Gaussianity is equivalent to joint uniformity of direction and CDF-transformed radius on a sphere–interval product. The full reflected-Neumann population energy has a unique Gaussian target, and its excess equals a squared maximum mean discrepancy (MMD), with analytic proofs and an accompanying Lean 4 formalization. Independent Gaussian blocks compose, and joint Gaussian codes support block resampling. GPU-friendly three-image pairwise and truncated spectral repulsion cost $O(N^2d)$ and $O(NdK)$, respectively, for K radial modes and fixed angular truncation. Optional radial and moment penalties and Gaussian-null calibration balance the finite-batch training terms. At $N = 8192$, $d = 128$, $K = 16$, three B300 input seeds give a $1.86 \pm 0.06\times$ spectral-over-pairwise speedup for full-loss forward/backward evaluation. Five-seed direct tests compare Gaussianization with analytic MMD, exact empirical assignment, refined Sinkhorn, and SIGReg. On learning representations: CIFAR-100, Shapes3D, and conditional-resampling experiments evaluate the trade-off between Gaussian representation structure and task performance. In the primary CIFAR-100 cohort, Wristband has lower held-out projector MMD^2 than analytic MMD, with mean linear-probe accuracies of 29.56% and 30.88%, respectively.

1 INTRODUCTION

A deterministic encoder with a standard-Gaussian output offers a useful interface: independent Gaussian blocks concatenate to a larger Gaussian, and individual blocks can be replaced by fresh Gaussian draws. These operations motivate Gaussianization, but their usefulness also depends on preserving task information and learning an accurate decoder or predictor. We develop a Gaussian-target regularizer with GPU-friendly pairwise and spectral evaluations, and measure computation, distribution matching, and retained task information.

Wristband extends the hyperspherical uniformity objective of Wang and Isola (2020) to this Gaussian interface. The construction has three ingredients: uniform directions, the correct Gaussian radius law, and independence between direction and radius. Mapping squared radius through its Gaussian CDF turns the latter two requirements into joint uniformity on $S^{d-1} \times [0, 1]$. A reflected radial kernel handles the interval boundaries. Its full population energy gap is a squared MMD with a Gaussian-adapted kernel, making the statistical target explicit.

Computational efficiency is central. For N embeddings in d dimensions, our three-image pairwise repulsion costs $O(N^2d)$, while a spectral approximation costs $O(NdK)$ for angular degrees $\ell \in \{0, 1\}$ and K radial modes. Both use matrix products and reductions without an inner assignment solve. For comparison, constructing a dense squared-Euclidean cost matrix and solving its $N \times N$ assignment with a cubic Hungarian implementation costs $O(N^2d + N^3)$ (Mills-Tettey et al., 2007). Analytic Gaussian-target MMD and linear-cost MMD approximations also offer efficient alternatives (§2). The improvement we establish is a structured spectral reduction of Wristband's own pairwise cost, with measured speed/fidelity trade-offs. Radial sorting and full-covariance auxiliaries add $O(N \log N + Ndq + q^3)$ work, where $q = \min(N, d)$.

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.

Contributions.

1. **A Gaussian-target extension of uniformity.** We characterize Gaussianity by the joint direction–radius law and derive a full population objective uniquely minimized by the Gaussian law, with excess equal to squared MMD (§3–4).

2. **GPU-friendly pairwise and spectral computation.** We provide $O(N^2d)$ pairwise repulsion and an $O(NdK)$ Neumann–harmonic approximation, with measured large-batch acceleration and explicit timing scope. We give a radial truncation bound and an explicit higher-order angular blind spot (§5).

3. **Gaussianity–utility evaluation.** Direct X and copula tests isolate the loss; CIFAR-100 probes, Shapes3D edits, and a synthetic resampling task measure downstream trade-offs. Auxiliary-loss and coupling-removal ablations separate the loss from its example architecture (§8).

2 Related work

Uniformity and Gaussianization. Wang and Isola (2020) study a pairwise repulsive kernel on the unit sphere. Wristband retains this angular construction and adds radius and radius–direction dependence. Concatenating separately normalized spherical blocks fixes their relative energies, whereas a uniform direction in the larger space has a random Beta-distributed energy split. Independent standard-Gaussian blocks instead concatenate exactly. VICReg (Bardes et al., 2022) regularizes variances and cross-coordinate covariances. Radial Gaussianization exploits elliptical symmetry (Lyu and Simoncelli, 2009). Radial-VCReg (Kuang et al., 2025) adds a radial entropy term, with a Gaussian characterization under elliptical symmetry. Correct moments and radius alone do not characterize arbitrary laws. E2MC studies entropy-based refinement of pretrained embeddings (Chakraborty et al., 2025); our target is a specific joint Gaussian law. SIGReg, introduced with LeJEPA (Balestriero and LeCun, 2025), also targets an isotropic Gaussian through randomly projected characteristic functions. For fixed numbers of directions and quadrature nodes its computation is linear in batch size. Wristband instead uses the direction–radius decomposition and reflected kernel; its finite spectral features have a different approximation scope. Linear-cost Gaussian regularization itself is therefore not a novelty claim of this work. KerJEPA (Zimmermann et al., 2025) places such regularizers in a broader kernel-discrepancy framework and relates SIGReg to sliced MMD. VISReg (Wu et al., 2026) separates variance control from sliced-Wasserstein shape matching. Wristband's distinction is its reflected direction–radius kernel and structured Neumann–harmonic approximation.

Distribution-matching objectives. MMD (Gretton et al., 2012) is used in the WAE-MMD formulation of Wasserstein autoencoders (Tolstikhin et al., 2018). Analytic Gaussian-target MMD (Rustamov, 2021) and Cramér–Wold autoencoders (Knop et al., 2020) avoid target sampling, with $O(N^2d)$ pairwise distance arithmetic. Null-standardizing an MMD penalty also precedes this work (Rustamov, 2021). Linear-pair MMD costs $O(Nd)$ for an RBF kernel; FastMMD (Zhao and Meng, 2015) uses $O(NdR)$ work with R Fourier features. These have distinct variance and approximation trade-offs. Population sliced Wasserstein (Kolouri et al., 2019) identifies a joint law, while finite projection sets can miss structure. Entropic OT admits GPU-friendly matrix scaling (Cuturi, 2013); Sinkhorn divergences subtract entropic self-costs (Feydy et al., 2019). This debiasing is distinct from finite-reference sampling: expected mini-batch transport is its own objective (Fatras et al., 2020). Online Sinkhorn, Batch Greenkhorn, and mini-batch partial transport provide further scalable alternatives (Mensch and Peyré, 2020; Kostić et al., 2022; Nguyen et al., 2022). Our comparisons concern the stated implementations and tuning budgets, without establishing a ranking of these method families.

Latent-variable models. VAEs and β-VAEs (Kingma and Welling, 2014; Higgins et al., 2017) use stochastic latent training; diffusion and flow matching (Ho et al., 2020; Lipman et al., 2023) learn generative models. Wristband supplies a deterministic batch regularizer and can be combined with an encoder architecture of choice. We borrow RealNVP-style affine couplings (Dinh et al., 2017) as an optional invertible postprocessor, without a change-of-variables likelihood. The loss, rather than a new autoencoder architecture, is our contribution.

3 THE WRISTBAND MAP

3.1 DEFINITION AND EQUIVALENCE THEOREM

Let $d \geq 2$ and let Q be a Borel probability measure on $\mathbb{R}^d$ with $Q(\{0\}) = 0$. Define the *wristband map*

$$W = S^{d-1} \times [0,1] \qquad \Phi : \mathbb{R}^d \setminus \{0\} \to W, \qquad \Phi(x) \;=\; \left(\frac{x}{\|x\|},\; F_{\chi^2_d}\big(\|x\|^2\big) \right), \tag{1}$$

where $F_{\chi^2_d}(s) = \gamma(d/2, s/2)/\Gamma(d/2)$ is the CDF of the chi-squared distribution with d degrees of freedom. Write σ_{d-1} for the uniform probability measure on S^{d-1} and $\mathrm{Unif}[0,1]$ for Lebesgue measure on $[0,1]$. The pushforward $\Phi_\# Q$ denotes the distribution of $\Phi(X)$ when $X \sim Q$.

Theorem 1 (Wristband equivalence)**.** *For $d \geq 2$ and Q as above,*

$$\Phi_\# Q \;=\; \sigma_{d-1} \otimes \mathrm{Unif}[0,1] \quad \Longleftrightarrow \quad Q \;=\; \mathcal{N}(0, I_d). \tag{2}$$

The proof follows Gaussian polar decomposition and the probability integral transform (App. A). The $d = 1$ equivalence also holds with a Rademacher sign on S^0; continuous spherical harmonics motivate our restriction. The result is machine-verified in Lean 4 §6.

3.2 COMPOSABILITY AND RESAMPLING

If the joint code $(Z_a, Z_b) \sim \mathcal{N}(0, I_{d_a+d_b})$, its blocks are independent standard Gaussians. Replacing Z_b by independent $\varepsilon \sim \mathcal{N}(0, I_{d_b})$ preserves the decoder's input law; marginal Gaussianity alone is insufficient. Recovering a data conditional also requires an accurate predictive model and sufficient context. Dependent observations generally cannot form independent, lossless deterministic blocks. Our MNIST example therefore uses a larger bottom-half context code and a smaller sampled top-half residual (§8.3), evaluated as conditional completion rather than exact posterior recovery or causal identification.

4 THE WRISTBAND GAUSSIAN LOSS

The full reflected population energy targets $\mathcal{N}(0, I_d)$. We distinguish this ideal objective from the finite kernel approximations and auxiliary penalties used for training. Let $\{x_i\}_{i=1}^N \subset \mathbb{R}^d \setminus \{0\}$, $N \geq 2$, be a batch of nonzero embeddings, and write $u_i = x_i/\|x_i\|$, $s_i = \|x_i\|^2$, $t_i = F_{\chi^2_d}(s_i)$.

4.1 THE WRISTBAND REPULSION KERNEL

The main term is a soft pairwise repulsion in (u, t)-space.

$$K_W(u, t;\, u', t') \;=\; k_{\mathrm{ang}}(u, u') \cdot k_{\mathrm{rad}}(t, t') \tag{3}$$

The angular component of the kernel is the chordal Gaussian:

$$k_{\mathrm{ang}}(u, u') = \exp\!\big\{-\beta\alpha^2 \|u - u'\|^2\big\} = e^{-2\beta\alpha^2} \exp\!\big\{(2\beta\alpha^2) u^\top u'\big\}, \tag{4}$$

Here $\alpha > 0$ balances angular and radial scales and $\beta > 0$ sets locality. The chordal default $\alpha^2 = 1/12$ matches $\mathbb{E}(T - T')^2 = 1/6$ to $\alpha^2 \mathbb{E}\|U - U'\|^2 = 2\alpha^2$ under the product-uniform law. Historical direct point-cloud experiments instead use $\alpha = 0.8$, $\beta = 64$.

Let $t, t' \in [0,1]$. The infinite-image Neumann reflection of the radial Gaussian is

$$k_{\mathrm{rad}}(t, t') = \sum_{m \in \mathbb{Z}} e^{-\beta(t - t' - 2m)^2} + \sum_{m \in \mathbb{Z}} e^{-\beta(t + t' - 2m)^2}. \tag{5}$$

Define the three-image reflected kernel, truncated from the expression above:

$$k_{\mathrm{img}}(t, t') \;=\; e^{-\beta(t - t')^2} + e^{-\beta(t + t')^2} + e^{-\beta(t + t' - 2)^2}. \tag{6}$$

The three radial terms correspond to the real point t', its reflection $-t'$ across the boundary 0, and its reflection $2 - t'$ across the boundary 1.

For global reduction, the pairwise repulsion uses $K_W^{\mathrm{img}}(u,t;u',t') = k_{\mathrm{ang}}(u,u')k_{\mathrm{img}}(t,t')$ and a numerical floor $\epsilon > 0$:

$$\mathcal{L}_{\mathrm{rep}} = \frac{1}{\beta}\log\left(\frac{\sum_{i,j=1}^{N} K_W^{\mathrm{img}}(u_i,t_i;\,u_j,t_j) - N}{3N^2 - N} + \epsilon\right). \tag{7}$$

The subtraction removes real self-interactions; reflected self-images are retained and affect finite-batch gradients. Some neural examples use a per-point variant that averages rowwise logarithms instead of taking a logarithm after the global reduction.

Why reflection. The full reflection sum restores the missing kernel mass near the boundaries of $[0,1]$ and has zero normal derivative there. The three-image form in (6) approximates this Neumann kernel; it does not satisfy the boundary condition exactly. The omitted radial kernel tail is uniformly at most $2e^{-\beta}/(1-e^{-3\beta})$, or 6.71×10^{-4} at $\beta = 8$, by bounding the omitted images by $2\sum_{j\geq 1} e^{-\beta j^2}$.

For a Borel probability measure P on W, define $\mathcal{E}(P) = \iint_{W\times W} K_W(w,w')\,dP(w)\,dP(w')$.

Theorem 2 (Energy minimization)**.** *For the kernel* (3) *with* k_{rad} *as in* (5)*,* $\beta, \alpha > 0$*, and* $d \geq 2$*, every Borel probability measure* P *on* W *satisfies*

$$\mathcal{E}(P) \;\geq\; \mathcal{E}(\sigma_{d-1} \otimes \mathrm{Unif}[0,1]), \tag{8}$$

with equality iff $P = \sigma_{d-1} \otimes \mathrm{Unif}[0,1]$.

What the population loss measures. Let $\mu_0 = \sigma_{d-1} \otimes \mathrm{Unif}[0,1]$. The full kernel has constant μ_0-potential $C = \mathcal{E}(\mu_0)$. Expanding the MMD gives

$$\mathcal{E}(P) - C = \mathrm{MMD}^2_{K_W}(P,\mu_0). \tag{9}$$

The harmonic–cosine modes have positive eigenvalues and their linear span is dense in continuous functions on W; hence this MMD vanishes only at μ_0. Thus the unlogged population energy controls a kernel integral probability metric. Its spectral weighting can assign different importance to different defects; it supplies no universal numerical TV, KL, or Euclidean-W_2 bound. Vanishing full-kernel excess does imply weak convergence to the Gaussian; for laws with finite second moments, a vanishing population moment penalty additionally yields W_2 convergence (Simon-Gabriel et al., 2023; Villani, 2009). The identity does not turn the calibrated composite into an absolute distance, and uniqueness of the optimum does not guarantee optimizer convergence.

4.2 Optional accelerators: radial Wasserstein and moment terms

The full infinite-image reflected repulsion is the only term required for the population characterization (Theorems 1–2; cf. §6). We optionally add radial and moment terms intended to stabilize finite-sample optimization:

$$\mathcal{L}_{\mathrm{rad}} = \frac{1}{N}\sum_{i=1}^{N}\left(t_{(i)} - \tfrac{i-1/2}{N}\right)^2, \tag{10}$$

$$\mathcal{L}_{\mathrm{mom}} = W_2^2\Big(\mathcal{N}(\hat{\mu},\hat{\Sigma}),\,\mathcal{N}(0,I_d)\Big) \;=\; \|\hat{\mu}\|^2 + \sum_{i=1}^{d}\left(\sqrt{\hat{\lambda}_i} - 1\right)^2, \tag{11}$$

Here $t_{(i)}$ are radial order statistics, $\hat{\mu}, \hat{\Sigma}$ are batch moments, and $\hat{\lambda}_i$ are covariance eigenvalues. The exact W_2^2 from the empirical radial law to $\mathrm{Unif}[0,1]$ is $\mathcal{L}_{\mathrm{rad}} + 1/(12N^2)$. The population counterparts vanish at the target; finite Gaussian batches generally have nonzero values. Auxiliary weights are optimization choices, not a requirement of Theorem 2. The implementation normalizes the moment term by d and uses a blended radial variant averaging $12\mathcal{L}_{\mathrm{rad}}$ with the mean squared difference of standard-normal quantiles at $t_{(i)}$ and $(i-1/2)/N$, clipping only the former probabilities.

4.3 SELF-CALIBRATION VIA z-SCORING

Here each $\mathcal{L}_*$ denotes its configured implemented component, including the radial variant and stabilized moment normalization by d. Its scale depends on N, d, and β. We draw M Gaussian batches at initialization to estimate each component mean $\bar{\mu}_*$ and variance $\hat{v}_*$. The implementation uses effective scales $\bar{s}_* = \sqrt{\max(\hat{v}_*, \delta)}$, with numerical variance floor $\delta > 0$. For $* \in \{\mathrm{rep}, \mathrm{rad}, \mathrm{mom}\}$, define

$$\widetilde{\mathcal{L}}_* = \frac{\mathcal{L}_*(\{x_i\}) - \bar{\mu}_*}{\bar{s}_*}.$$

The training-time numerator is

$$S = w_{\mathrm{rep}}\widetilde{\mathcal{L}_{\mathrm{rep}}} + w_{\mathrm{rad}}\widetilde{\mathcal{L}_{\mathrm{rad}}} + w_{\mathrm{mom}}\widetilde{\mathcal{L}_{\mathrm{mom}}}.$$

The weights $w_{\mathrm{rad}}, w_{\mathrm{mom}}$ correspond to code options `lambda_rad`, `lambda_mom`; outer task weights use λ. We divide by the similarly floored null scale $\bar{s}_S = \sqrt{\max(\widehat{\mathrm{Var}}(S), \delta)}$:

$$\mathcal{L}_{\mathrm{wb}} \;=\; \frac{S}{\bar{s}_S}. \tag{12}$$

This accounts for component correlations. The null mean is approximately zero; variance is approximately one only when the final floor is inactive. A score near zero is not a Gaussianity certificate or a calibrated p-value. Scores depend on batch size, kernel, reduction, and enabled components (App. C).

Mini-batches. Uniform random mini-batches give unbiased estimates of the *unlogged off-diagonal* kernel energy, so cross-mode pairs are sampled rather than structurally excluded. Logs, self-images, auxiliaries and batch-specific calibration prevent equating average training scores with a full-dataset score. Rare modes can yield high sampling variance; a longitudinal comparison remains open.

5 SPECTRAL APPROXIMATION AND ITS SCOPE

The full Neumann kernel has a cosine radial expansion and a spherical-harmonic angular expansion. Retaining radial indices $k = 0, \ldots, K-1$ and angular degrees $\ell \in \{0, 1\}$ gives

$$c_{0,k} = N^{-1} \sum_i \cos(k\pi t_i), \qquad c_{1,k} = \sqrt{d}\, N^{-1} \sum_i u_i \cos(k\pi t_i), \tag{13}$$

$$\mathcal{E}_{\mathrm{sp}} = \lambda_0 \sum_{k=0}^{K-1} a_k c_{0,k}^2 + \lambda_1 \sum_{k=0}^{K-1} a_k \|c_{1,k}\|^2, \tag{14}$$

$$\mathcal{L}_{\mathrm{rep}}^{\mathrm{sp}} = \beta^{-1} \log\!\left(\frac{\mathcal{E}_{\mathrm{sp}}}{\lambda_0 a_0}\right). \tag{15}$$

The $\sqrt{d}$ factor accounts for the degree-one orthonormal basis. We evaluate normalized coefficient ratios directly, avoiding the historical implementation's absolute floors on tiny kernel constants. The spectral objective retains diagonal terms and uses constant-mode normalization, so its finite-batch reduction differs from (7). Computing the $N \times K$ cosine matrix and its product with the $N \times d$ direction matrix costs $O(NdK)$ work and $O(Nd + NK + dK)$ tensor storage. Only the summary state is $O(dK)$; the implemented training path is not streaming. Pseudocode is in App. B.

The radial coefficients obey $a_k/a_0 = 2e^{-\pi^2 k^2/(4\beta)}$ for $k \geq 1$. Choose K from a bound on the *sum* of omitted coefficients (App. B); increasing β requires more radial modes at fixed tolerance. This does not control angular error. A finite angular expansion is noncharacteristic: if directions are uniform on the vertices $\{\pm 1/\sqrt{d}\}^d$ and independent of a correct Gaussian radius, every retained nonconstant coefficient vanishes, as do population radial and moment penalties, yet the law is not Gaussian. Increasing K cannot repair this blind spot. Sign-symmetric radius–direction dependence with a uniform radial coordinate and zero conditional mean direction is also invisible to these features. The pairwise path has no angular truncation; its three-image radial approximation remains distinct from the full-kernel theorem. Spectral evaluation requires $d \geq 3$, chordal angles, and global reduction.

6 LEAN 4 FORMALIZATION

Theorems 1 and 2 are machine-checked in Lean 4 (`mathlib` v4.28.0). The wristband equivalence (`wristbandEquivalence`) is `sorry`-free as an iff; energy minimization and uniqueness are established for the infinite-image Neumann kernel modulo a small inventory of imported textbook axioms and routine analytic support lemmas.

7 USING WRISTBAND IN A DETERMINISTIC AUTOENCODER

A deterministic Gaussian autoencoder (DGAE) uses an encoder E to map data to a code; an optional invertible affine-coupling postprocessor f shapes that code before regularization. The decoder receives $f^{-1}(z)$, where $z = f(E(x))$:

$$\mathcal{L} = \lambda_{\text{rec}} \,\text{MSE}(D(f^{-1}(z)), x) + \lambda_{\text{wb}} \mathcal{L}_{\text{wb}}(\{z_i\}), \tag{16}$$

where MSE averages over samples and input coordinates. Since $f^{-1}(f(E(x))) = E(x)$, the reconstruction path cancels the postprocessor algebraically. Wristband supplies its distributional training signal; no base-density likelihood or log-Jacobian is used. Setting f to the identity gives a direct encoder regularizer. The synthetic example uses learnable-key Euclidean attention (App. E); the architecture is not required by the loss.

8 EXPERIMENTS

We separate direct Gaussianization, computational fidelity, and downstream utility. The X and RAC tests optimize points without an encoder, decoder, or coupling network. Learned-representation tests measure task quality alongside Gaussianity rather than assuming that one predicts the other. Here "full Wristband" means implemented repulsion plus radial and moment auxiliaries, not exact evaluation of the infinite-image population kernel.

8.1 DIRECT GAUSSIANIZATION WITH NEW-SEED CONFIRMATION

Table 1 reports 135 confirmations on five new paired seeds after selection on seeds $0, 1$. The original 115 cases and 20-case SIGReg extension use matched clouds, budgets and evaluation. X places points on coordinate axes; RAC preserves sampled Gaussian radial and angular marginals while coupling radius ranks to direction entropy. Analytic MMD has the lowest mean sliced error in all four problems. Other diagnostics reveal different trade-offs: full Wristband has lower radial Kolmogorov–Smirnov (KS) discrepancy than MMD in both 10D problems, but substantial radius–angle coupling remains in RAC. At $d = 128$, the absolute Spearman correlation between radius and angular entropy is 0.544 ± 0.037 for full Wristband, 0.690 ± 0.054 for spectral, and 0.170 ± 0.016 for MMD; iid Gaussian controls give 0.023 ± 0.023. This coordinate-specific diagnostic complements sliced error rather than certifying independence. The supplementary code and records specify the X and radius–angle copula (RAC) constructions, metrics, solver caps and selection settings. Assignment uses SciPy's modified Jonker–Volgenant solver (Crouse, 2016). A separate 40-case reference-policy check reduces assignment's high-dimensional sliced error when Gaussian references are fixed, while other metrics disagree; its paired records are supplied in the supplement.

8.2 MEASURED COMPUTATION AND APPROXIMATION SCOPE

Table 2 and Fig. 1 compare isolated objectives over three input seeds. At $N = 8192$, stable spectral full Wristband is $1.86 \pm 0.06\times$ faster than full pairwise Wristband (mean and sample SD of paired time ratios). At $N = 1024$, their times are similar, while analytic MMD and SIGReg are faster. SIGReg also has comparable large-batch time to full spectral Wristband. The radial and moment auxiliaries reduce the spectral speedup relative to repulsion alone. These measurements compare the stated objectives and implementations; they exclude network computation and do not match optimization quality. The supplementary records also include a separate cuGraph assignment pilot with qualified CPU/GPU comparisons under different timing boundaries.

A separate synthetic ResNet-18/projector panel includes network and loss forward/backward computation at $d = 256$. All 60 cases completed. At $N = 8192$ per view, full pairwise and spectral

Table 1: New-seed direct Gaussianization: sliced $W_2^2/2$ ($\times 10^3$), mean $\pm$ sample SD over paired cloud seeds 100–104. Hyperparameters were frozen using seeds $0, 1$; all 135 confirmations completed 1000 Adam steps with $N = 1024$. OT training references are resampled each step. Spectral is unavailable at $d = 2$.

Method	X-2D	X-10D	RAC-10D	RAC-128D
Wristband full	1.84 ± 0.3	2.13 ± 0.12	2.03 ± 0.098	2.37 ± 0.035
Wristband repulsion	2.04 ± 0.36	4.14 ± 1	2.59 ± 0.038	2.36 ± 0.045
Wristband spectral	—	7.32 ± 0.27	2.36 ± 0.093	2.37 ± 0.059
Analytic MMD	1.66 ± 0.3	1.93 ± 0.087	1.92 ± 0.094	2.26 ± 0.053
Assignment OT	1.68 ± 0.36	15.7 ± 0.76	15.7 ± 0.78	185 ± 0.75
Refined Sinkhorn	1.7 ± 0.34	3.87 ± 0.3	2.23 ± 0.093	3.02 ± 0.058
SIGReg	3.05 ± 0.15	1.94 ± 0.087	1.96 ± 0.094	2.33 ± 0.05

Table 2: Loss forward/backward time (ms), $d = 128$, NVIDIA B300. Mean $\pm$ sample SD over three input-seed medians, each from five repetitions after two warmups. Full Wristband includes auxiliaries; CPU JV includes cost construction and transfers. SIGReg times fresh directions; OT uses fixed Gaussian references. Setup, optimizer and network computation are excluded. The completed JV $N = 8192$ extension uses larger runtime caps.

Method	$N = 1024$	$N = 8192$
Wristband, full	5.124 ± 0.165	8.773 ± 0.071
Stable spectral, full	4.852 ± 0.234	4.717 ± 0.201
Analytic Gaussian MMD	2.417 ± 0.099	7.179 ± 0.095
SIGReg (fresh directions)	1.411 ± 0.044	4.703 ± 0.046
Sinkhorn, refined	23.862 ± 3.899	128.336 ± 0.716
SciPy CPU JV assignment	123.981 ± 5.411	23065 ± 678

Wristband take 747.83 ± 0.27 and 739.05 ± 0.31 ms, respectively: a paired ratio of 1.0119 ± 0.0003, with network work dominating total time. This diagnostic excludes optimization, data loading, augmentation and learning; its protocol and per-method records are supplied in the supplement.

Auxiliary and angular-stress ablations. Under the historical direct protocols, repulsion-only runs had equal or lower barycentric-reference scores than the full loss in all four settings, with X-10D within seed variation. A separate 64D test used antipodal pairs on 32 orthogonal two-dimensional planes, midpoint-quantile radii, and a random global rotation to null the retained degree-one statistics without a trivial low-rank covariance defect. At $N = 4096$, $K = 16$, full pairwise and spectral losses reached barycentric W_2 z-scores of 2.67 ± 0.25 and 14.8 ± 0.30, respectively. These historical scores standardize distance to a recursively averaged Gaussian-sample reference against fresh Gaussian batches; they are not population Gaussian distances. The supplement records their provenance. The analytic counterexample in §5 establishes the angular limitation independently of these optimization runs.

8.3 Autoencoding and conditional completion

On a standardized 15D five-component mixture, an 8D deterministic autoencoder reaches reconstruction MSE 0.241 ± 0.006; its calibrated Wristband score falls from 3405.1 ± 88.2 to 6.28 ± 1.79. The latter is a reduction in the configured statistic, not a distance-to-normal certificate. In a 2×2 ablation (coupling/no coupling, full/repulsion-only loss), removing the postprocessor improved both reported metrics within each objective. Totals for different objectives were calibrated separately and must not be compared across them. Architecture and protocol details are in App. D.

For MNIST, separate encoders map bottom and top 14×28 halves to 18D context and 3D residual blocks. Wristband acts on their joint 21D code. Training combines full reconstruction with a weak expected pixelwise-L_1 loss over 32 residual draws. The five-seed full-reconstruction and conditional losses are 0.0369 ± 0.0007 and 0.0533 ± 0.0008; the calibrated score is 1.90 ± 0.33. At inference the residual is replaced by fresh Gaussian draws. Figures 2–3 show completions with the observed

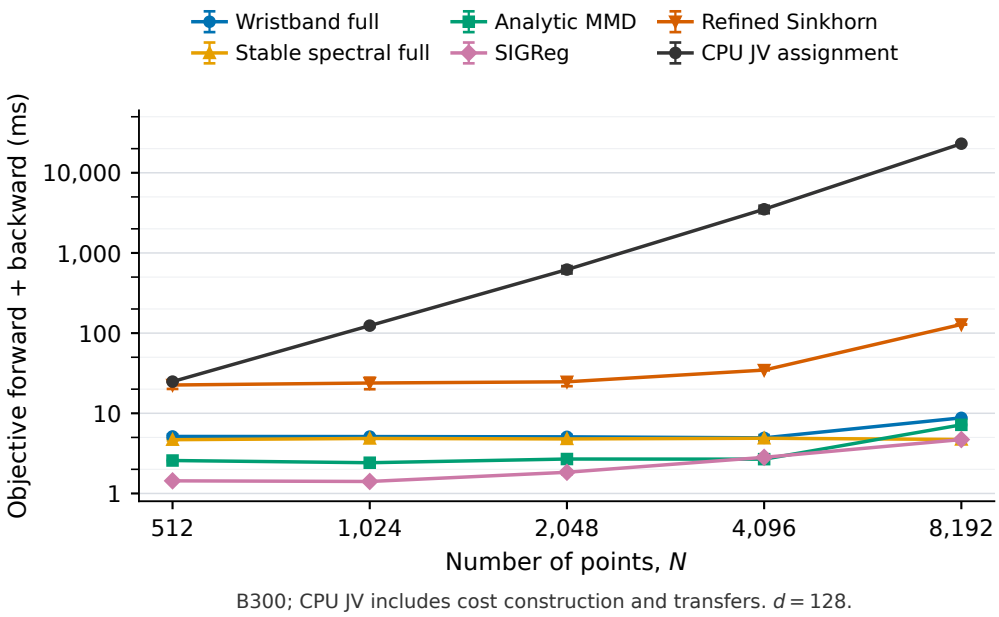

Figure 1: Measured loss-computation scaling at $d = 128$: mean and sample SD over three input-seed medians, with the measurement boundaries of Table 2. All points, including CPU assignment at $N = 8192$, are measured; lines connect points without fitting or extrapolation. CPU assignment is shown alongside GPU objectives.

half clamped for display. This illustrates conditional use of the interface, without measuring exact posterior calibration.

8.4 Representation utility and semantic editing

We compare seven regularization choices under a common VICReg-style self-supervised learning (SSL) recipe with ResNet-18 on CIFAR-100 (He et al., 2016; Krizhevsky, 2009). The objective is $25\,\mathrm{MSE}(Z_1, Z_2) + \lambda[R(Z_1) + R(Z_2)]/2$ for two augmented views. Screens select λ by validation probe accuracy after 50 SSL/20 probe epochs, before 200/100-epoch confirmation on paired seeds 100–102. An expanded coefficient screen yields a separate four-method confirmation on seeds 103–105, with stronger Sinkhorn regularization. Table 3 reports official-test backbone-probe accuracy and held-out projector Gaussianity, which concern different representations. MMD^2 U omits empirical self-pairs and uses an analytic Gaussian target with five RBF bandwidths. Wristband has the lowest measured projector MMD^2; assignment has the highest mean probe accuracy. Wristband's pairwise implementation takes less SSL-plus-probe time than assignment or refined Sinkhorn, and more than MMD, SIGReg or VICReg. These selected operating points expose a Gaussianity–utility trade-off, not an accuracy-matched speed comparison. The cohorts use the same official test set and are not pooled; three seeds give descriptive uncertainty. Selection configurations and individual results are supplied in the supplement.

On Shapes3D (Burgess and Kim, 2018), deterministic autoencoders use ground-truth factor labels to edit codes by centroid displacement. The five-seed summary includes the seed-0 selection pilot; selected weights are at the lower grid boundaries. Wristband and sliced Wasserstein give similar edit/preservation rates (Table 4); Wristband has lower mean paired-image MSE, while sliced Wasserstein has lower MMD^2. Unregularized codes give higher edit and factor-probe accuracy. These controlled edits do not establish unsupervised disentanglement or causal identification. A synthetic two-factor conditional-resampling study further tests replacing one latent by Gaussian noise. Its retained author-reported aggregates show a task/error trade-off, without statistically established gains over tuned MMD variants. The supplementary historical application code specifies this protocol.

Table 3: CIFAR-100 confirmations: mean ± sample SD of three paired seeds per separate cohort, 200 SSL/100 probe epochs. Probe uses the raw frozen backbone; MMD^2 U uses raw projector outputs on 1,024 test images. Times exclude setup/calibration and feature metrics. Expanded-range Sinkhorn uses $\varepsilon = 0.02$; Wristband remains fixed.

Method	λ	Test probe (%)	MMD^2 U (10^{-3})	SSL+probe (min)
Original validation range; confirmation seeds 100–102				
No distribution penalty	0	1.99 ± 0.57	382.78 ± 0.00	15.60 ± 0.04
VICReg	30	29.93 ± 0.17	30.37 ± 2.29	16.04 ± 0.01
Analytic MMD	300	30.88 ± 1.52	0.56 ± 0.03	16.19 ± 0.42
Exact assignment	30	34.26 ± 0.25	288.13 ± 0.28	54.61 ± 2.94
Refined Sinkhorn	30	15.48 ± 0.76	380.69 ± 0.80	30.16 ± 4.31
SIGReg	3	31.01 ± 1.08	0.37 ± 0.02	16.35 ± 0.55
Wristband	0.1	29.56 ± 0.72	0.09 ± 0.03	18.59 ± 0.68
Expanded validation range; confirmation seeds 103–105				
VICReg	30	30.80 ± 0.46	31.71 ± 4.27	15.97 ± 0.39
Exact assignment	30	33.67 ± 0.67	287.58 ± 0.12	54.49 ± 0.44
Refined Sinkhorn	1000	31.52 ± 0.34	167.71 ± 0.18	30.10 ± 1.78
Wristband	0.1	30.35 ± 0.48	0.08 ± 0.01	18.36 ± 0.20

Table 4: Shapes3D controlled edits: mean ± sample s.d. over five seeds. Edits use ground-truth labels. A separately trained factor oracle measures target-factor accuracy (Edit) and mean accuracy on the five unchanged factors (Preserve). Paired MSE compares to the image with the intended factor changed.

Regularizer	Edit (%, ↑)	Preserve (%, ↑)	Paired MSE (↓)
None	93.37 ± 5.09	98.90 ± 0.81	0.00337 ± 0.00046
MMD	61.18 ± 6.72	89.33 ± 1.15	0.01014 ± 0.00171
Sliced Wasserstein	91.70 ± 3.06	97.93 ± 1.05	0.00266 ± 0.00071
Wristband	91.48 ± 2.74	97.97 ± 0.68	0.00202 ± 0.00070

9 Discussion and limitations

The full-kernel theorem characterizes a population target; finite kernel approximations and calibrated training scores require external Gaussianity diagnostics. Numerical floors also require care near zero embeddings. The experiments separate distribution matching, task utility and computation: expanded-range CIFAR confirmations improve Sinkhorn's probe performance, while network computation reduces the benefit of faster loss evaluation. They do not establish ImageNet-scale performance, exhaustive tuning, accuracy-matched speedups or superiority over baseline families. A limited $\beta \in [4, 16]$ sweep changes synthetic DGAE reconstruction MSE by at most 8.9%, without establishing general sensitivity. Attention-module ablation and longitudinal mini-batch/full-dataset validation remain open. Conditional resampling requires an accurate decoder or predictor and is not an identified individual causal counterfactual.

10 Conclusion

Wristband turns Gaussian representation learning into joint sphere–interval uniformity, with a unique full-kernel population target, analytic proofs and an accompanying Lean 4 formalization. Reflected boundary handling and Gaussian-null calibration give a configurable deterministic regularizer. Its spectral approximation reduces repulsion cost from $O(N^2 d)$ to $O(NdK)$, trading angular identifiability for faster evaluation. Joint Gaussian codes provide a composable interface for block resampling; downstream usefulness also depends on retained information and an accurate predictive model. The experiments quantify computation and Gaussianity–utility trade-offs, supporting Wristband as a practical option for learning Gaussian representations.

AI USE STATEMENT

Generative AI tools, including Claude and Codex, were used for mathematical derivation and proof assistance, understanding technical concepts, implementing standard and original methods, facilitating or running experiments, and prose editing. Derivation assistance included exploring alternative kernel forms and the spectral approximation. AI tools assisted with reviewing the writing, literature and source checks, manuscript revision, experiment design, generation of synthetic point clouds, experiment code and result analysis.

ETHICS STATEMENT

The experiments use public benchmark data and synthetic distributions. Gaussian latent structure does not establish fairness, privacy, semantic independence, or valid causal interventions. The evaluations here do not assess those properties or deployment in consequential applications.

REPRODUCIBILITY STATEMENT

The map proof and spectral coefficients, radial truncation bound and pseudocode are in Apps. A and B. Calibration and autoencoder architectures are in Apps. C–E. The anonymous supplement supplies experiment code and frozen configurations, tuning and confirmation records, saved direct-study arrays, the Lean development, and CPU-only table replay tools. Its guides distinguish replaying recorded summaries from regenerating inputs or retraining models. `PROTOCOL_GUIDE.md` maps the results to their configurations and records. It preserves failed and capped runs; datasets and model checkpoints are omitted, with public sources and regeneration procedures documented.

REFERENCES

R. Balestriero and Y. LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

A. Bardes, J. Ponce, and Y. LeCun. VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning. In *ICLR*, 2022.

C. Burgess and H. Kim. 3D Shapes Dataset. `https://github.com/google-deepmind/3d-shapes`, 2018.

D. Chakraborty, Y. LeCun, T. G. J. Rudner, and E. Learned-Miller. Improving Pre-trained Self-Supervised Embeddings Through Effective Entropy Maximization. In *AISTATS*, PMLR 258:433–441, 2025.

D. F. Crouse. On implementing 2D rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4):1679–1696, 2016. doi:10.1109/TAES.2016.140952.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 26, 2013.

F. Dai and Y. Xu. *Approximation Theory and Harmonic Analysis on Spheres and Balls*. Springer Monographs in Mathematics. Springer, New York, 2013. doi:10.1007/978-1-4614-6660-4.

E. De Micheli. Integral representation for Bessel's functions of the first kind and Neumann series. *Results in Mathematics*, 73:61, 2018. doi:10.1007/s00025-018-0826-5.

L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using Real NVP. In *ICLR*, 2017.

E. Dorovatas, G. Paraskevopoulos, and A. Potamianos. Auto-compressing networks. In *NeurIPS*, 38, 2025.

V. Dutordoir, N. Durrande, and J. Hensman. Sparse Gaussian Processes with Spherical Harmonic Features. In *ICML*, PMLR 119:2793–2802, 2020.

K. Fatras, Y. Zine, R. Flamary, R. Gribonval, and N. Courty. Learning with minibatch Wasserstein: Asymptotic and gradient properties. In *AISTATS*, PMLR 108:2131–2141, 2020.

J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouvé, and G. Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *AISTATS*, PMLR 89:2681–2690, 2019.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *J. Machine Learning Research*, 13:723–773, 2012.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *ICLR*, 2014.

S. Knop, P. Spurek, J. Tabor, I. Podolak, M. Mazur, and S. Jastrzębski. Cramer-Wold Auto-Encoder. *J. Machine Learning Research*, 21(164):1–28, 2020.

S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. Rohde. Generalized sliced Wasserstein distances. In *NeurIPS*, 2019.

V. R. Kostić, S. Salzo, and M. Pontil. Batch Greenkhorn Algorithm for Entropic-Regularized Multimarginal Optimal Transport: Linear Rate of Convergence and Iteration Complexity. In *ICML*, PMLR 162:11529–11558, 2022.

A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Y. Kuang, Y. Dagade, D. Chakraborty, E. Learned-Miller, R. Balestriero, T. G. J. Rudner, and Y. LeCun. Radial-VCReg: More informative representation learning through radial Gaussianization. In *NeurIPS Workshop on Symmetry and Geometry in Neural Representations*, 2025.

Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling. In *ICLR*, 2023.

S. Lyu and E. P. Simoncelli. Nonlinear extraction of independent components of natural images using radial Gaussianization. *Neural Computation*, 21(6):1485–1519, 2009.

A. Mensch and G. Peyré. Online Sinkhorn: Optimal Transport distances from sample streams. In *NeurIPS*, 33:1657–1667, 2020.

G. Ayorkor Mills-Tettey, A. Stentz, and M. B. Dias. The Dynamic Hungarian Algorithm for the Assignment Problem with Changing Costs. Technical Report CMU-RI-TR-07-27, Carnegie Mellon University, July 2007.

K. Nguyen, D. Nguyen, T.-A. Vu-Le, T. Pham, and N. Ho. Improving mini-batch optimal transport via partial transportation. In *ICML*, PMLR 162:16656–16690, 2022.

R. M. Rustamov. Closed-form expressions for maximum mean discrepancy with applications to Wasserstein auto-encoders. *Stat*, 10(1):e329, 2021. doi:10.1002/sta4.329.

C.-J. Simon-Gabriel, A. Barp, B. Schölkopf, and L. Mackey. Metrizing weak convergence with maximum mean discrepancies. *Journal of Machine Learning Research*, 24(184):1–20, 2023.

I. Tolstikhin, O. Bousquet, S. Gelly, and B. Schölkopf. Wasserstein auto-encoders. In *ICLR*, 2018.

C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften, volume 338. Springer, 2009. doi:10.1007/978-3-540-71050-9.

T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *ICML*, PMLR 119:9929–9939, 2020.

H. Wu, R. Balestriero, and M. Levine. VISReg: Variance-Invariance-Sketching Regularization for JEPA training. *arXiv preprint arXiv:2606.02572*, 2026.

J. Zhao and D. Meng. FastMMD: Ensemble of circular discrepancy for efficient two-sample test. *Neural Computation*, 27(6):1345–1372, 2015. doi:10.1162/NECO_a_00732.

E. Zimmermann, H. Wiltzer, J. Szeto, D. Alvarez-Melis, and L. Mackey. KerJEPA: Kernel Discrepancies for Euclidean Self-Supervised Learning. *arXiv preprint arXiv:2512.19605*, 2025.

A Proof of Theorem 1

We provide the standard measure-theoretic proof; the corresponding statement is formalized in Lean 4 in the accompanying development.

($\Leftarrow$) $Q = \mathcal{N}(0, I_d) \implies \Phi_{\#}Q = \sigma_{d-1} \otimes \mathrm{Unif}[0,1]$. Let $X \sim \mathcal{N}(0, I_d)$. Write $R = \|X\|$ and $U = X/\|X\|$. The Gaussian polar decomposition gives $U \sim \sigma_{d-1}$, $R^2 \sim \chi^2_d$, and $U \perp\!\!\!\perp R$. By the probability integral transform, $T = F_{\chi^2_d}(R^2) \sim \mathrm{Unif}[0,1]$, and since T is a measurable function of R^2, we also have $U \perp\!\!\!\perp T$. Therefore $\Phi(X) = (U, T) \sim \sigma_{d-1} \otimes \mathrm{Unif}[0,1]$.

($\Rightarrow$) $\Phi_{\#}Q = \sigma_{d-1} \otimes \mathrm{Unif}[0,1] \implies Q = \mathcal{N}(0, I_d)$. Let $X \sim Q$ and $(U, T) = \Phi(X)$. By assumption, $U \sim \sigma_{d-1}$, $T \sim \mathrm{Unif}[0,1]$, and $U \perp\!\!\!\perp T$. Since $T \in (0,1)$ almost surely, define $S = F^{-1}_{\chi^2_d}(T)$ and $R = \sqrt{S}$, with arbitrary inverse-CDF values at the null-probability endpoints. Strict monotonicity gives $S = \|X\|^2$ almost surely. By the reverse probability integral transform, $S \sim \chi^2_d$, so R has the $\mathrm{Chi}(d)$ density

$$f_R(r) = \frac{1}{2^{d/2-1}\Gamma(d/2)} r^{d-1} e^{-r^2/2}, \qquad r > 0.$$

Moreover, R is a measurable function of T, hence $R \perp\!\!\!\perp U$. Since $X = RU$ almost surely by construction of Φ, (R, U) are the polar coordinates of X. Let $\omega_{d-1} = 2\pi^{d/2}/\Gamma(d/2)$ be the surface area of S^{d-1}. The joint density of (R, U) is $f_R(r)/\omega_{d-1}$, and the change of variables $x = ru$ has Jacobian r^{d-1}. Thus the induced density of X is

$$f_X(x) = \frac{f_R(\|x\|)}{\omega_{d-1}\|x\|^{d-1}} = \frac{1}{(2\pi)^{d/2}} e^{-\|x\|^2/2},$$

which is exactly the $\mathcal{N}(0, I_d)$ density.

B Spectral coefficients and Neumann pseudocode

For the chordal angular kernel, the spherical-harmonic eigenvalues under uniform probability measure are $\lambda_\ell = \Gamma(\nu+1)(2/c)^\nu e^{-c} I_{\nu+\ell}(c)$, where $\nu = (d-2)/2$, $c = 2\beta\alpha^2$, and I_ρ is the modified Bessel function (Dutordoir et al., 2020; Dai and Xu, 2013; De Micheli, 2018). The radial expansion is $k_{\mathrm{rad}}(t, t') = \sum_{k \geq 0} a_k \cos(k\pi t)\cos(k\pi t')$, with $a_0 = \sqrt{\pi/\beta}$ and $a_k/a_0 = 2e^{-\pi^2 k^2/(4\beta)}$ for $k \geq 1$. Thus the algorithm uses $r_0 = 1$ and $\rho = \lambda_1/\lambda_0 = I_{\nu+1}(c)/I_\nu(c)$.

Proof of Theorem 2. Let $Y_{\ell m}$ be spherical harmonics orthonormal under σ_{d-1}, with $Y_{00} = 1$, and set $p_{\ell m k} = \int Y_{\ell m}(u)\cos(k\pi t)\, dP(u,t)$. The uniformly convergent kernel expansion gives $\mathcal{E}(P) = \sum_{\ell,m,k} \lambda_\ell a_k p^2_{\ell m k} \geq \lambda_0 a_0 = \mathcal{E}(\mu_0)$. All coefficients are positive and $p_{000} = 1$. Equality forces every nonconstant harmonic–cosine moment to vanish. Their span is dense in $C(W)$, so equality implies $P = \mu_0$, proving uniqueness.

For $a = \pi^2/(4\beta)$ and integer $K \geq 1$, the omitted radial kernel obeys

$$\frac{|k_{\mathrm{rad}} - k_{\mathrm{rad},K}|}{a_0} \leq 2\sum_{k=K}^{\infty} e^{-ak^2} \leq \frac{2e^{-aK^2}}{1 - e^{-a(2K+1)}}.$$

The final inequality uses $(K+j)^2 \geq K^2 + j(2K+1)$ for integer $j \geq 0$. Increasing K until this bound meets a chosen tolerance controls radial kernel truncation, not angular error or downstream performance.

Algorithm 1 Spectral Neumann wristband repulsion (single forward pass; global reduction)

Require: batch $\{x_i\}_{i=1}^N \subset \mathbb{R}^d \setminus \{0\}$, $N \geq 2$, $d \geq 3$; precomputed $r_k = a_k/a_0$, $\rho = \lambda_1/\lambda_0$.

$u_i \leftarrow x_i/\|x_i\|,\ t_i \leftarrow F_{\chi^2_d}(\|x_i\|^2)$ ▷ $O(Nd)$

$C_{i,k} \leftarrow \cos(k\pi t_i)$ for $k = 0, \ldots, K-1$ ▷ $O(NK)$

$c_{0,k} \leftarrow N^{-1} \sum_i C_{i,k}$ ▷ $O(NK)$

$c_{1,k} \leftarrow (\sqrt{d}/N) \sum_i u_i C_{i,k}$ ▷ $O(NdK)$

$R_K \leftarrow \sum_k r_k c_{0,k}^2 + \rho \sum_k r_k \|c_{1,k}\|^2$ ▷ $R_K \geq 1$

return $\mathcal{L}_{\mathrm{rep}}^{\mathrm{spec}} \leftarrow \beta^{-1} \log R_K$

C CALIBRATION CONSTANTS

Table 5 reports the null means and effective normalization denominators $(\bar{\mu}_*, \bar{s}_*)$, with 4096 Gaussian calibration batches. The implementation floors variances at $\delta = \sqrt{\epsilon_{\mathrm{machine}}}$ before taking their square roots. In float32 the resulting minimum denominator is $\epsilon_{\mathrm{machine}}^{1/4} = 0.01858136$, explaining the repeated 0.0186 entries. They are not unfloored empirical standard deviations. The same floor is applied to the variance of the weighted numerator S. On the calibration sample, $\widehat{\mathrm{Var}}(\mathcal{L}_{\mathrm{wb}}) = \hat{v}_S / \max(\hat{v}_S, \delta) \leq 1$; unit variance holds only if the floor is inactive. Thus a reported score of 6.28, for example, need not represent 6.28 actual Gaussian-null standard deviations. The recorded calibration-only times on an NVIDIA RTX PRO 6000 Blackwell Workstation Edition are 4.93s, 8.76s and 27.63s for $d = 8, 64, 256$, respectively. These initialization costs are separate from training-step timings.

Table 5: Null means and floored normalization denominators for $\mathcal{L}_{\mathrm{wb}}$ at $N = 1024$, $\beta = 8$, using 4096 calibration batches.

	$d = 8$		$d = 64$		$d = 256$	
Component	$\bar{\mu}$	$\bar{s}$	$\bar{\mu}$	$\bar{s}$	$\bar{\mu}$	$\bar{s}$
$\mathcal{L}_{\mathrm{rep}}$	−0.3486	0.0186	−0.3606	0.0186	−0.3619	0.0186
$\mathcal{L}_{\mathrm{rad}}$	0.0025	0.0186	0.0025	0.0186	0.0025	0.0186
$\mathcal{L}_{\mathrm{mom}}$	0.0032	0.0186	0.0170	0.0186	0.0659	0.0186

D HYPERPARAMETERS AND EXPERIMENTAL DETAILS

Historical DGAE and MNIST experiments use Adam or AdamW with cosine annealing and batch size 1024, with means and standard deviations over five seeds. Neural-model runs on Windows compile the model or train-step wrapper with `torch.compile(..., mode="default")`. The new CIFAR-100 cohorts instead use LARS, batch size 512, and three paired seeds per cohort. Direct point-cloud experiments optimize free point variables; frozen configurations and evaluation code for the new studies are supplied in the supplement.

Direct confirmation metrics. X samples a uniformly chosen axis with coordinate $\mathrm{Unif}[-\sqrt{3d}, \sqrt{3d}]$; RAC rank-couples sampled χ_d radii to uniform spherical directions by entropy $-\sum_j u_j^2 \log(u_j^2)$. Sliced error averages $W_2^2/2$ over 256 random projections and four fresh 1024-point Gaussian references. Radial KS compares CDF-transformed radii with $\mathrm{Unif}[0, 1]$. Selection minimizes final sliced error averaged over tuning seeds $0, 1$. Analytic MMD equally averages RBF bandwidths $\sqrt{d}\{0.25, 0.5, 1, 2, 4\}$.

CIFAR-100 metrics and selection. The shared split contains 45,000 training and 5,000 validation images. The raw frozen backbone has 512 dimensions and the projector 256. The held-out analytic-target MMD uses the same five RBF bandwidth multipliers as the direct study. Its U estimator excludes self-pairs and can be negative. Coefficient grids and tie-breaking rules are recorded in the supplement; selection uses validation probe accuracy, not minimum MMD.

15D mixture DGAE. The dataset has $100{,}000$ samples in $\mathbb{R}^{15}$ from a five-component mixture: Gaussian samples receive a sign-preserving $3/2$-power transform on one third of their coordinates, followed by component-specific linear transforms and shifts. The mixture is standardized coordinate-wise. The encoder uses a learnable-key Euclidean attention module with internal dimension 128, 64 heads, 256 basis points, a learned query transform, and an Auto-Compressing Network (ACN) head combiner (Dorovatas et al., 2025). The decoder uses the same module with input/output dimensions reversed and no query transform. The flow has 4 affine coupling layers, hidden width 32, 2 conditioner blocks, $s_{\max} = 2$ (the coupling log-scale bound), and deterministic pairwise permutations. Training uses 40 epochs, learning rate 3×10^{-4}, $\lambda_{\mathrm{rec}} = 1$, and $\lambda_{\mathrm{wb}} = 0.1$.

MNIST conditional sampling. The reported MNIST protocol uses a 90/10 train/test split from the standard training set with seed 42. Each image is split into top and bottom 14×28 halves. The bottom and top encoders are separate ACN networks mapping 392 pixels to $d_b = 18$ and $d_t = 3$, respectively, with hidden dimension 128 and 4 blocks. Each branch has its own 4-layer affine coupling flow with hidden width 64. The decoder applies the inverse flows, projects each branch to 128 features, concatenates them, maps to a $128 \times 7 \times 7$ seed tensor, and upsamples by transposed convolutions to 28×28. The conditional reconstruction loss matches inference-time usage by encoding only the bottom half and sampling the missing residual,

$$\mathcal{L}_{\mathrm{cf}}(x) = \frac{1}{J} \sum_{j=1}^{J} \ell(D(z_{\mathrm{bot}}, \varepsilon_j), x), \qquad \varepsilon_j \sim \mathcal{N}(0, I_3), \tag{17}$$

where ℓ is pixelwise L_1. We average the *loss*, not the output; expected reconstruction losses favor variance reduction when overweighted, so $\mathcal{L}_{\mathrm{cf}}$ receives a small weight. Reported settings are AdamW, learning rate 2×10^{-3}, $J = 32$, weights $(\lambda_{\mathrm{rec}}, \lambda_{\mathrm{cf}}, \lambda_{\mathrm{wb}}) = (1, 0.1, 0.1)$, and 20 warm-up epochs. The supplement distinguishes this reported protocol from defaults in the surviving example code.

Wristband defaults. The historical direct point-cloud defaults are $\beta = 64$, $\alpha = 0.8$, global reduction, $w_{\mathrm{rep}} = 1$, $w_{\mathrm{rad}} = 0.1$, and $w_{\mathrm{mom}} = 1$; new confirmations use the selected configurations recorded in the supplement. Historical DGAE/MNIST examples use $\beta = 8$, $\alpha = \sqrt{1/12}$ and pairwise `per_point` reduction. CIFAR-100 confirmations use pairwise `global` reduction, $\beta = 8$, $\alpha = \sqrt{1/12}$, radial/moment weights $0.1/1$, and 256 calibration batches of shape $(512, 256)$. Calibration uses Gaussian null batches of the training shape; the constants in App. C use 4096 batches. Spectral timing uses global reduction and $K = 16$.

E EXTENDED DGAE CONSTRUCTION

This appendix expands the architectural choices in §7; neither is required by the Wristband objective.

Learnable-key Euclidean attention. We use a multi-head softmax attention layer in which keys $\{k_j\}$ and values $\{v_j\}$ are free parameters rather than projections of input tokens. Given an input representation h, each head forms a query $q = W_q h + b$ (or $q = h$ when the dimensions already match). The Euclidean logits with inverse-temperature τ are

$$\ell_j(q) = \tau \left(q^\top k_j - \frac{1}{2} \|k_j\|^2 \right).$$

This equals $-\frac{\tau}{2}\|q - k_j\|^2$ up to the query-dependent constant $\frac{\tau}{2}\|q\|^2$, which cancels inside the softmax. Each basis point therefore acts as a learned RBF prototype, and attention becomes soft nearest-prototype retrieval.

Invertible affine-coupling postprocessor. A stack of RealNVP-style coupling layers with deterministic permutations and identity initialization supplies the optional map f. For full reconstruction its forward and inverse cancel algebraically. The regularizer acts on $f(E(x))$ and can still change the encoder; invertibility alone guarantees neither Gaussianity nor preservation of task utility. Section 8.3 summarizes an ablation removing this map.

F MNIST CONDITIONAL SAMPLING GRIDS



Figure 2: MNIST conditional sampling grid. The bottom half is observed and clamped for display; the top half is decoded from fresh residual samples. Similar observed bottoms can admit different digit completions. Figure 3 shows a larger grid.

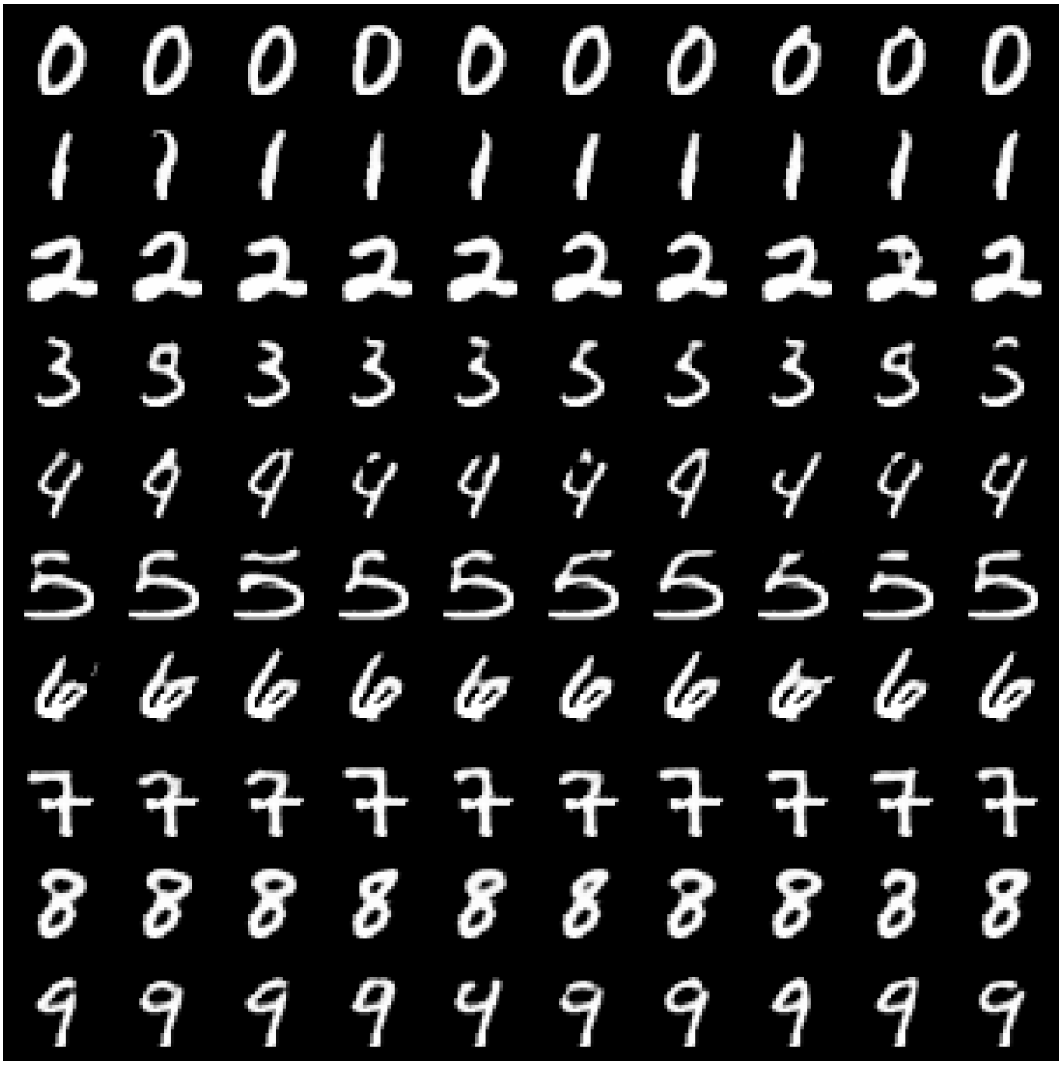

Figure 3: Larger MNIST conditional sampling grid using the setup of Figure 2. The bottom half is observed and clamped for display; the top half is decoded from fresh residual samples drawn from $\mathcal{N}(0, I_3)$.

G SUNSHINE DISTRIBUTION ADVERSARIAL EXAMPLE

The sunshine construction divides the angle of a two-dimensional Gaussian into 16 sectors and rotates alternate sectors onto their neighbors. It preserves radii while leaving angular gaps. These are historical illustrative runs.

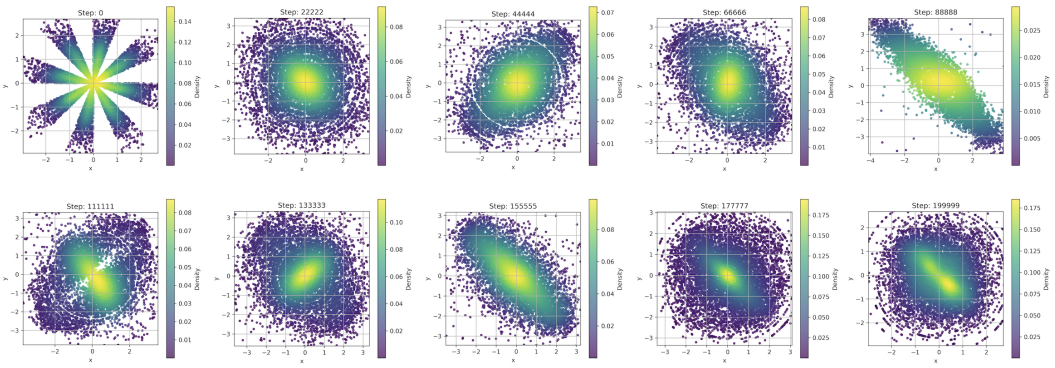

Figure 4: Radial-VCReg on the sunshine point cloud, with snapshots through 200k optimization steps. The displayed angular structure remains visible despite radial regularization. This is an empirical failure example, not a claim that every elliptically symmetric non-Gaussian law is outside the method's scope.

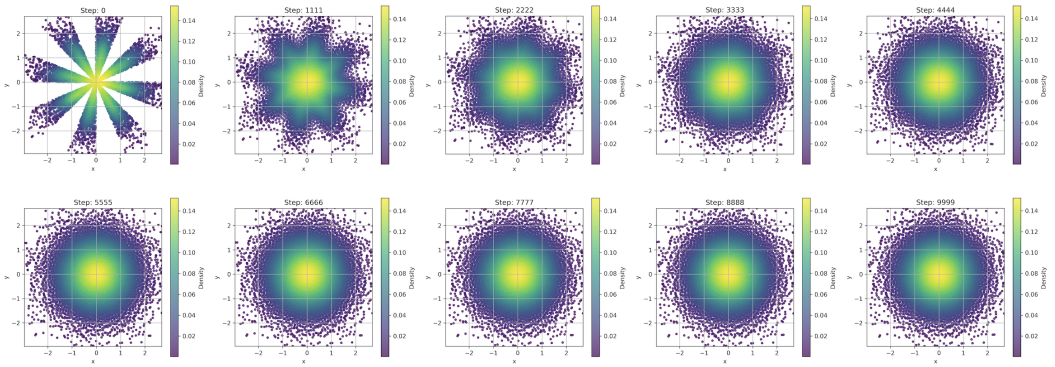

Figure 5: Wristband on the same sunshine construction. The displayed structure dissipates before 10k steps in this run. Visual agreement does not certify exact Gaussianity.