HILBERT MIXTURE WASSERSTEIN DISTANCE FOR EFFICIENT GAUSSIAN MIXTURE COMPARISON

Anonymous authors
Paper under double-blind review

ABSTRACT

Mixture Wasserstein compares Gaussian mixtures by optimizing both Gaussian factor alignment and component matching for every pair, making repeated comparisons costly. We introduce Hilbert Mixture Wasserstein (HMW), a metric built from a reusable ordered representation of each mixture. We encode each Gaussian by its mean and Cholesky factor and sort these vectors in a deterministic Hilbert order. Component weights set the lengths of consecutive subintervals of a unit interval. Overlaps between the two mixtures' subintervals specify a sparse coupling, with costs evaluated from the full feature vectors. Comparison is linear in the prepared representations' combined size, and equally weighted mixtures with a common component count admit exact finite Euclidean embeddings. HMW upper-bounds mixture Wasserstein, and we decompose the squared-distance gap into the costs of fixing factor alignment and component matching. We give conditions for a small gap and show that changes in component order can keep HMW bounded away from zero as mixture Wasserstein tends to zero. Constructed examples distinguish covariance response from component matching, and timing experiments examine scaling with mixture size and dimension. We also examine representation reuse in image retrieval and ordered grouping in density reduction and 3D Gaussian Splatting scene compression.

1 INTRODUCTION

Gaussian mixtures represent heterogeneous distributions through local means, covariances, and weights. Their comparison is central to mixture interpolation, density approximation, and 3D Gaussian Splatting (3DGS) scene compression (Chen et al., 2019; Delon & Desolneux, 2020; Zhang et al., 2024; Wang et al., 2025). Repeated queries require repeatedly matching components across mixtures, which can dominate computation. We seek a distance that instead reuses a representation prepared independently for each mixture.

Optimal transport (OT) minimizes the cost of moving probability mass (Peyré & Cuturi, 2019), using a pairwise cost matrix for discrete measures (Ahuja et al., 1993). Entropic regularization enables matrix scaling through Sinkhorn iterations (Cuturi, 2013; Altschuler et al., 2017), while Sinkhorn divergences correct the resulting self-interaction bias (Feydy et al., 2019). Stochastic dual optimization (Genevay et al., 2016), low-rank couplings (Scetbon et al., 2021), and sparse kernels (Li et al., 2023) further reduce computation or storage.

A second route solves transport by sorting each one-dimensional projection, then averages over directions to obtain a sliced Wasserstein distance (Bonneel et al., 2015). Generalized projections (Kolouri et al., 2019), optimized directions, direction distributions, and subspaces adapt the comparison to informative variations (Deshpande et al., 2019; Nguyen et al., 2021; Paty & Cuturi, 2019). More recent work improves direction selection and Monte Carlo estimation (Nguyen et al., 2024; Nguyen & Ho, 2024). For Gaussian mixtures, projection affects the response to both location and covariance differences.

The Bures–Wasserstein (BW) formula gives the transport cost between individual Gaussians (Dowson & Landau, 1982; Gelbrich, 1990; Bhatia et al., 2019). Mixture Wasserstein (MW) optimizes a coupling between these components (Chen et al., 2019; Delon & Desolneux, 2020), with extensions to other mixture families (Dusson et al., 2026). Sliced transport has also been used to learn Gaussian mixtures (Kolouri et al., 2018). For comparing their component measures, mixture sliced Wasserstein

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.

(MSW) replaces Gaussian costs by projected costs, and double-sliced mixture Wasserstein (DSMW) further reduces matching to scalar transport (Piening & Beinert, 2025). Even after rescaling sliced distances to match BW for uniform covariance expansion, they respond less strongly to shape changes that expand some directions and contract others. Section 2 quantifies this difference with exact integration over directions.

Ordering simplifies correspondence while retaining costs in the original space. Projected Wasserstein obtains correspondences from projected orderings (Rowland et al., 2019), while Hilbert curve projection (HCP) orders multivariate supports and matches cumulative masses (Li et al., 2024). Linear transport representations make repeated comparisons inexpensive by embedding distributions relative to a common reference (Wang et al., 2013; Liu et al., 2025). For Gaussian mixtures, fixing a component order and factor representation removes the need to optimize alignment and matching for each pair.

We introduce *Hilbert Mixture Wasserstein* (HMW), using a canonical recursive Hilbert rule to order Gaussian mean and Cholesky features. Each component occupies a subinterval of a unit interval whose length equals its weight, producing a feature-valued step function. HMW is the distance between the two step functions, evaluated by weighting feature differences by their interval overlaps. Only a linear number of component pairs overlap, making comparison linear in the prepared representation size. Equal weights and a common component count give exact finite Euclidean embeddings.

HMW upper-bounds MW, and the squared-distance gap separates into the added costs of fixed factor alignment and ordered coupling. We identify configurations where the two distances coincide and bound their difference under small feature perturbations that preserve weights and component orders. Small parameter changes can also swap positions in the order and keep HMW bounded away from zero as MW tends to zero, giving HMW a strictly finer topology. Experiments examine distance geometry, representation reuse, and ordered grouping for density reduction and 3DGS scene compression.

2 Problem and Geometric Motivation

2.1 Gaussian Mixture Transport

We regard each Gaussian component, including its mean and covariance, as one atom. Let $\mathcal{G}_d^+ = \{\mathcal{N}(\mu, \Sigma) : \mu \in \mathbb{R}^d, \Sigma \in \mathbb{R}^{d\times d}, \Sigma \succ 0\}$ and represent mixtures by

$$\mathcal{P} = \sum_{i=1}^{K} a_i \delta_{G_i}, \qquad \mathcal{Q} = \sum_{j=1}^{M} b_j \delta_{F_j},$$

where $G_i, F_j \in \mathcal{G}_d^+$, δ_G is unit mass at G, and $a \in \mathbb{R}^K$, $b \in \mathbb{R}^M$ are positive probability vectors after removing zero weights and merging identical atoms. Write $\Pi(a, b) = \{\pi \in \mathbb{R}^{K\times M} : \pi \geq 0, \pi\mathbf{1} = a, \pi^\mathsf{T}\mathbf{1} = b\}$ for their couplings, with $\mathbf{1}$ an all-ones vector. For $G = \mathcal{N}(\mu, \Sigma)$ and $F = \mathcal{N}(\nu, \Lambda)$, the Gaussian squared transport cost and mixture distance are (Dowson & Landau, 1982; Gelbrich, 1990; Delon & Desolneux, 2020)

$$c_{\mathrm{BW}}(G, F) = \|\mu - \nu\|^2 + \mathrm{tr}\Big(\Sigma + \Lambda - 2(\Sigma^{1/2}\Lambda\Sigma^{1/2})^{1/2}\Big),$$
$$D^2 := \mathrm{MW}_{\mathrm{BW}}^2(\mathcal{P}, \mathcal{Q}) = \min_{\pi\in\Pi(a,b)} \sum_{ij} \pi_{ij} c_{\mathrm{BW}}(G_i, F_j).$$

Here $d_{\mathrm{BW}} = \sqrt{c_{\mathrm{BW}}}$, $\|\cdot\|$ is the Euclidean norm, and $\|\cdot\|_F$ is the Frobenius norm. The metric D compares the discrete measures of Gaussian components, rather than transporting samples from the mixture densities directly. All covariances are symmetric positive definite (SPD).

2.2 Covariance Response under Projection

Projection reduces the cost of comparing Gaussians, but it can change the relative importance of covariance scale and shape. Population MSW averages the squared 2-Wasserstein cost between one-dimensional Gaussian projections under a common component coupling. DSMW further projects each Gaussian's scalar mean and standard deviation, then matches the resulting scalars separately for

each direction pair (Piening & Beinert, 2025). We use normalized uniform directions; Appendix A.1 gives the definitions. For $d \geq 2$, consider $\mathcal{N}(0, \sigma^2 I)$ and $\mathcal{N}(0, (\sigma I + \epsilon S)^2)$, where $\sigma > 0$, $S \in \mathbb{R}^{d\times d}$ is symmetric, and the perturbed root remains SPD. Write $S = sI/\sqrt{d} + S_0$, with $\operatorname{tr} S_0 = 0$. The first term changes scale equally in every direction, while the traceless term changes shape. To compare local sensitivity, divide each squared distance by ϵ^2 and let $\epsilon \to 0$. This removes the common perturbation magnitude, leaving a coefficient that measures the response to S. Denote these limits by q_{BW}, q_{MSW}, and q_{DSMW}, respectively. They satisfy

$$q_{\mathrm{BW}} = s^2 + \|S_0\|_F^2, \qquad d \cdot q_{\mathrm{MSW}} = 2d \cdot q_{\mathrm{DSMW}} = s^2 + \frac{2}{d+2}\|S_0\|_F^2. \tag{1}$$

Multiplying squared MSW by d and squared DSMW by $2d$ matches their response to BW for uniform expansion ($S_0 = 0$). After this calibration, a traceless shape perturbation still receives only $2/(d+2)$ of its BW weight. Varying both the direction and magnitude of S can therefore make a perturbation farther from the reference under MW but closer under the sliced distances (Figure 2). Appendix A.1 gives the expansion and remainder. For mixtures, the component cost is one part of the comparison; the other is how mass is matched between components.

3 Hilbert Mixture Wasserstein

HMW represents each mixture by an ordered sequence of Gaussian features, with each feature occupying an interval proportional to its mass. Comparing two representations then requires only their interval overlaps.

3.1 Gaussian factor representation

For $G = \mathcal{N}(\mu, \Sigma)$, let $\Sigma = L_\Sigma L_\Sigma^{\mathsf{T}}$ be its Cholesky factorization with positive diagonal, and define

$$\Phi(G) = \big(\mu, \operatorname{vecl}(L_\Sigma)\big) \in \mathbb{R}^p, \qquad p = d + d(d+1)/2. \tag{2}$$

The operator vecl lists lower-triangular entries row by row, including the diagonal and without rescaling. The mean and Cholesky factor uniquely determine the Gaussian, giving the squared feature cost

$$c_\Phi(G, F) = \|\Phi(G) - \Phi(F)\|^2 = \|\mu - \nu\|^2 + \|L_\Sigma - L_\Lambda\|_F^2.$$

For $Z \sim \mathcal{N}(0, I_d)$, the shared-noise pair $X = \mu + L_\Sigma Z$, $Y = \nu + L_\Lambda Z$ has marginals G, F and satisfies $\mathbb{E}\|X - Y\|^2 = c_\Phi(G, F)$. Gaussian transport instead optimizes the relative factor alignment (Bhatia et al., 2019),

$$c_{\mathrm{BW}}(G, F) = \|\mu - \nu\|^2 + \min_{O^{\mathsf{T}}O = I} \|L_\Sigma - L_\Lambda O\|_F^2.$$

Fixing $O = I$ replaces pairwise alignment by Euclidean distance between factors computed once per Gaussian. The resulting cost depends on the coordinate axes because Cholesky factors do not generally rotate with the Gaussian as a whole.

3.2 Canonical order and mixture representation

We remove zero weights, merge identical atoms, normalize mass, and enumerate the features lexicographically. A fixed deterministic median-recursive Hilbert implementation then orders each feature set separately (Li et al., 2024). It partitions along successive coordinates and traverses the subsets with Hilbert orientations, using unscaled features and unweighted median splits. Here canonical means that, with this implementation fixed, the order is independent of input labels and the comparison partner. The recursion may reach singleton subsets before visiting every coordinate, whereas the cost c_Φ uses the entire feature vector. Appendix B.1 details tie handling and coordinate access.

Write $x_i = \Phi(G_i)$ and $y_j = \Phi(F_j)$. Let $\mathcal{S}$ denote this sorting procedure, which takes a list of feature vectors and returns their index permutation. Thus $\alpha = \mathcal{S}(x_1, \ldots, x_K)$ means that x_{α_r} occupies position r in the source order; define $\beta = \mathcal{S}(y_1, \ldots, y_M)$ similarly. Assign each ordered component a consecutive subinterval of $(0, 1]$ whose length equals its weight. Specifically, set $A_0 = B_0 = 0$, $A_r = \sum_{u \leq r} a_{\alpha_u}$, $B_s = \sum_{v \leq s} b_{\beta_v}$, $I_r = (A_{r-1}, A_r]$, and $J_s = (B_{s-1}, B_s]$. The ordered representation is the step function

$$Z_{\mathcal{P}}(u) = x_{\alpha_r} \quad \text{for } u \in I_r, \qquad U(\mathcal{P}, \mathcal{Q}) := \mathrm{HMW}(\mathcal{P}, \mathcal{Q}) = \|Z_{\mathcal{P}} - Z_{\mathcal{Q}}\|_{L^2(0,1)}. \tag{3}$$

Define $Z_{\mathcal{Q}}$ analogously on the intervals J_s. The overlap length assigns transport mass to each component pair, giving the weighted northwest-corner (NW) plan

$$\pi^{\mathrm{H}}_{\alpha_r \beta_s} = |I_r \cap J_s| = \max\{0, \min(A_r, B_s) - \max(A_{r-1}, B_{s-1})\}. \tag{4}$$

The plan preserves both marginals and has at most $K + M - 1$ positive entries (Lemma 4). The squared norm is the overlap-weighted sum $U^2 = \sum_{ij} \pi^{\mathrm{H}}_{ij} c_\Phi(G_i, F_j)$. Figure 1 illustrates the ordering and overlap construction.

When both mixtures have n distinct atoms with weights $1/n$, concatenate the ordered features as

$$\boldsymbol{z}_{\mathcal{P}} = n^{-1/2}(x_{\alpha_1}^{\mathsf{T}}, \ldots, x_{\alpha_n}^{\mathsf{T}})^{\mathsf{T}} \in \mathbb{R}^{np}. \tag{5}$$

The two interval partitions now coincide, so only components at the same rank are matched and $U(\mathcal{P}, \mathcal{Q}) = \|\boldsymbol{z}_{\mathcal{P}} - \boldsymbol{z}_{\mathcal{Q}}\|$.

Computational cost. For $N = K + M$ atoms, factorization takes $O(Nd^3)$ time and balanced Hilbert recursion with linear-time selection uses $O(N \log N)$ scalar comparisons. Prepared comparison takes $O(Np)$ time, with $O(Nd^2)$ total storage (Appendix B.1).

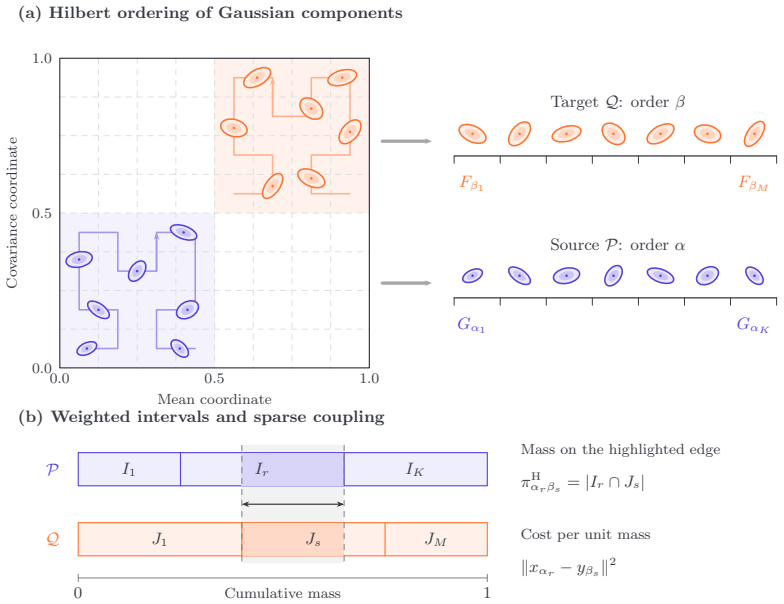


Figure 1: Gaussian ordering and sparse transport. (a) Ellipses illustrate components in a normalized mean–covariance plane. Normalization is for display only; the algorithm orders the unscaled features in (2). (b) Weighted interval overlaps define masses, with edges priced in the complete feature space.

4 Metric Geometry and Comparison with MW

We assume SPD covariances and exact arithmetic throughout, with proofs in Appendix A.

4.1 Metric Representation and Topology

The shared-noise coupling gives $c_{\mathrm{BW}} \le c_\Phi$ for Gaussian atoms. Averaging over the ordered plan yields $D \le U$ (Theorem 5), so a small HMW distance ensures a small MW distance.

Theorem 1 (Metric under canonical ordering)**.** *Fix the coordinates and a deterministic ordering that assigns each finite Gaussian-atom measure a unique list after merging duplicates and removing zero weights, independently of input labels and the comparison partner. Then U in* (3) *is a metric on finite probability measures on $\mathcal{G}_d^+$.*

Injectivity of the ordered representation identifies equal measures, and its L^2 norm gives the triangle inequality. The argument also applies to other canonical orderings. The inequality $D \le U$ implies

that convergence in HMW entails convergence in MW. The converse can fail when a small change in Gaussian parameters swaps components' positions in the canonical order. The step functions then compare different components over intervals of positive mass, so HMW can stay bounded away from zero even as MW tends to zero. Appendix A.7 constructs such two-atom sequences for every $d \geq 1$ when means and covariances may both vary. Thus the HMW topology is strictly finer than the MW topology.

4.2 Factor and coupling excess over MW

To isolate the effect of component matching, keep the Hilbert coupling and replace its feature costs by exact BW costs,

$$V^2 = \sum_{ij} \pi_{ij}^{\mathrm{H}} c_{\mathrm{BW}}(G_i, F_j), \qquad D^2 \leq V^2 \leq U^2.$$

The intermediate distance V is also a metric (Appendix A.4) and requires factor alignment only on matched pairs. With $\Delta_{ij} = c_{\Phi}(G_i, F_j) - c_{\mathrm{BW}}(G_i, F_j)$,

$$U^2 - D^2 = \underbrace{\sum_{ij} \pi_{ij}^{\mathrm{H}} \Delta_{ij}}_{A = U^2 - V^2} + \underbrace{V^2 - D^2}_{B}. \tag{6}$$

Here A compares Cholesky and BW costs on the same matched pairs, while B compares the Hilbert coupling with an optimal coupling under the same BW costs. Both are nonnegative. Thus A measures the effect of the Gaussian factor representation, and B measures the effect of component matching. For diagonal covariances, the two component costs agree, so $A = 0$ and the entire gap comes from matching. Appendix A.5 examines how these component costs differ under small covariance perturbations.

The distances are exactly equal in a simple geometric setting. Suppose all covariances are diagonal and all feature vectors lie on one straight line. If both Hilbert orders follow the same direction along that line, matching cumulative masses in order is optimal, so $A = B = 0$ and HMW equals MW. For small perturbations of these features that preserve the weights and component orders, Corollary 10 bounds the distance gap by the size of the perturbations. Appendix A.6 also gives a more general bound under conditions on the BW component costs that make ordered matching optimal or nearly optimal (Theorem 8).

5 Distance geometry and computational cost

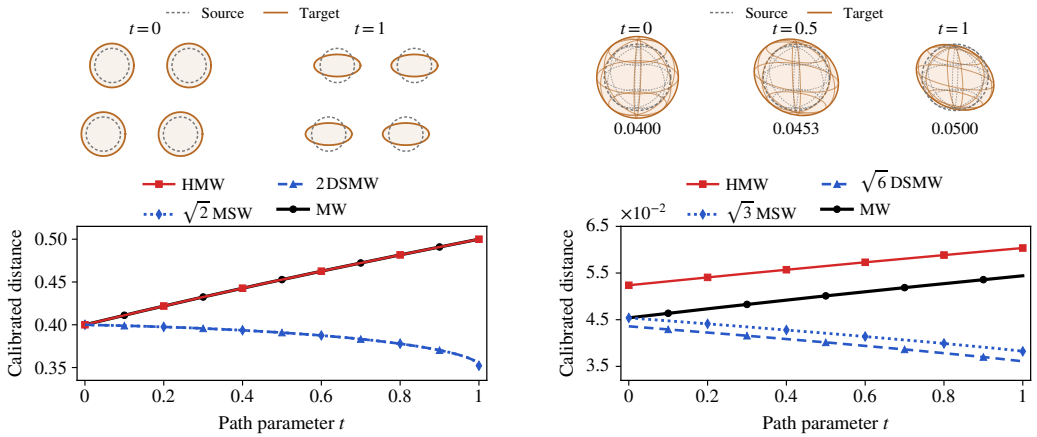
5.1 Covariance response and component pairing

Figure 2 examines how the distances respond when Gaussian components change shape. In each panel, the source and target have four components with weights $1/4$. The source is fixed with identity component covariances. The target means and weights remain fixed while its common covariance varies over 161 equally spaced values of $t \in [0, 1]$. At $t = 0$, every target component expands equally in all coordinate directions relative to the source. At $t = 1$, one direction expands and another contracts. The Euclidean norm of the standard-deviation change is then 1.25 times its value at $t = 0$. All covariances are diagonal, so Cholesky and BW component costs agree and $A = 0$. For both panels, we plot $\sqrt{d}\,\mathrm{MSW}$ and $\sqrt{2d}\,\mathrm{DSMW}$, using the fixed calibration for uniform covariance expansion in Section 2.

Panel (a) uses identical mean locations in dimension two. Each source component has unit standard deviations, while the target standard deviations change from approximately $(1.283, 1.283)$ to $(1.354, 0.646)$. Matching components at the same location is optimal and agrees with the Hilbert matching, so $B = 0$. MW and HMW therefore coincide and increase from 0.400 to 0.500. The calibrated sliced distances instead decrease from 0.400 to 0.352. Expansion and contraction can partly offset each other after projection, so the sliced distances assign less cost to the final shape change even though its full Gaussian transport cost is larger.

Panel (b) uses distinct fixed mean sets in dimension three. Hilbert pairs source 2 with target 4 and source 4 with target 2, instead of matching equal indices, adding a positive mean-transport cost

(Appendix B.4). As the covariance changes, MW and HMW both increase, with HMW remaining larger. Their squared-distance difference $B = U^2 - D^2$ is constant because the matching stays fixed and every component pair receives the same covariance contribution at each t. Both calibrated sliced distances decrease along this path as well. Appendices B.4–B.5 give the mixture parameters, matching checks, and numerical integration rules.



(a) Shared mean locations ($d = 2$).

(b) Distinct mean locations ($d = 3$).

Figure 2: Distances from uniform expansion to directional expansion and contraction. Sliced distances use the stated dimension-based calibration. Upper shapes compare source (gray) and target (orange). Panel (a) shows all four components at actual scale. Panel (b) shows their shared covariance as centered 3D ellipsoids, with a fixed view and scale and standard-deviation changes magnified tenfold. Numbers beneath them give the actual covariance BW distances, excluding mean costs. The curves use the original covariances and mean sets.

5.2 Computation across mixture sizes and dimensions

Figure 3 compares online evaluation times for HMW, exact MW, MSW, and DSMW as mixture size and dimension increase. We generate paired mixtures with equal component counts, random full covariances, and nonuniform weights. Targets perturb source means, covariances, and weights. Mean shifts are scaled so their average squared length equals the average squared BW covariance distance. All methods receive the same raw inputs and run on one CPU core. Timings include method-specific preprocessing and transport computation, excluding data generation and warm-up. MSW uses 16 directions and DSMW 16 direction pairs, sampled by normalizing vectors with independent Uniform$(0, 1)$ entries as in the released implementation.

HMW has the lowest displayed times across both panels. The largest HMW and DSMW instances have 10^5 components.

5.3 Reusing ordered representations

Repeated comparison allows each stored mixture to be prepared once. We examine this setting using handwritten digit images (Alpaydin & Kaynak, 1998), each encoded by 16 equally weighted local Gaussians for joint pixel position and intensity in $\mathbb{R}^3$. Each component contributes three mean coordinates and six Cholesky coordinates, so (5) gives a 144-dimensional HMW vector per image. The dataset contains 300 queries and nested galleries of 100 and 500 images. Timings use 30 queries balanced across digits and cohorts.

Starting from the Gaussian encodings, we prepare the gallery representations before processing queries. Table 1 separates preparation of one image from comparison of one prepared query against the stored gallery. HMW uses Euclidean distances between its vectors. For DSMW, each direction pair produces 16 sorted scalar values; concatenating the lists for L direction pairs gives $16L$ coordinates. MW and MSW optimize a coupling for every query–gallery pair. MW forms the full BW component-

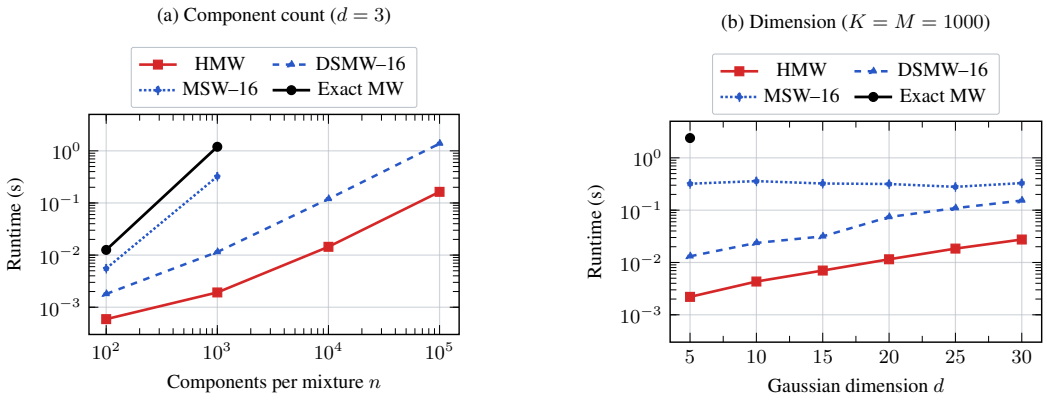


Figure 3: Recorded online runtimes with 16-direction sliced baselines. Panel (a) varies component count at $d = 3$; panel (b) varies dimension with $K = M = 1000$. Points are medians across three or five cases, except $n = 10^4, 10^5$, which use one case each. Dense methods stop at preset memory/dimension limits. MW uses one call per case in (b); other points use case-level timing medians. Appendix B.6 gives the protocol.

cost matrix, while MSW caches projected means and standard deviations for 16 directions and averages their Gaussian costs.

Table 1: Prepared comparisons on a common, HMW-selected encoding of 16 joint Gaussians with intensity scale 2. Times are median milliseconds. Preparation starts from Gaussian parameters; scans compare one prepared query with 100 or 500 stored mixtures and exclude ranking. Vector dimension applies to exact Euclidean representations at equal weights.

Method	Dimension	Preparation	Scan 100	Scan 500
HMW	144	0.441	0.011	0.036
MW	–	0.390	112.528	562.600
MSW, $L = 16$	–	0.482	31.811	159.119
DSMW, $L = 4$	64	0.483	0.010	0.018
DSMW, $L = 16$	256	0.603	0.014	0.068
DSMW, $L = 64$	1024	0.715	0.040	0.246
DSMW, $L = 256$	4096	0.949	0.185	0.799

The cached HMW scan takes 0.036 ms for 500 mixtures. MSW retains cost construction and coupling optimization for each pair and takes 159.119 ms. DSMW's scan time increases with its number of cached coordinates. With independently selected encodings, HMW and DSMW give 94.0% and 93.7%–95.3% top-one retrieval accuracy, respectively, on the larger gallery (Appendix B.3.1).

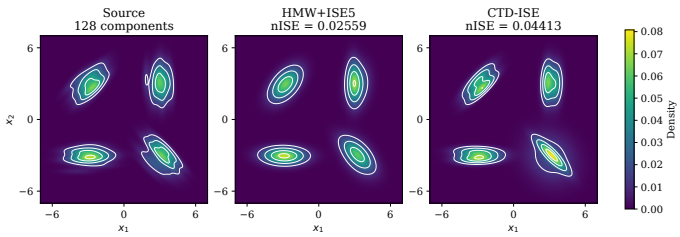
6 Applications

We use ordered grouping to replace Gaussian components by fewer representatives and assess how well they preserve mixture densities and rendered 3DGS scenes.

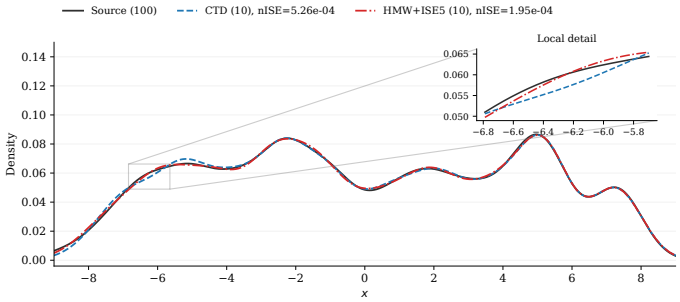
6.1 Gaussian mixture reduction

Gaussian mixture reduction approximates a source density f by a mixture density g with fewer components. Integrated squared error (ISE), $\int [f(x) - g(x)]^2 \, dx$, measures their total squared density discrepancy. We report normalized ISE (nISE), dividing this error by $\int f(x)^2 \, dx$ so that smaller values indicate a closer density approximation.

We divide the Hilbert-ordered components into consecutive groups whose component counts differ by at most one (Figure 4, schematic). Each representative takes its group's total weight and matches



(a) Two-dimensional density reduction ($128 \to 8$).



(b) One-dimensional scalar-order example, $100 \to 10$.

Figure 4: Selected density reductions favoring HMW+ISE5. The schematic combines consecutive pairs, preserving group mass. Both methods share moment-matched initialization. Panel (a) uses common color and contour levels; (b) abbreviates CTD-ISE as CTD. Appendices C.1 and C.2 give protocols and selection criteria.

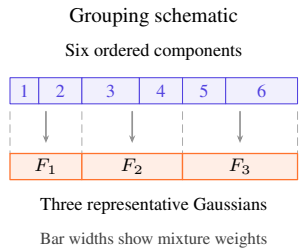


Figure 4(a) reduces 128 equally weighted, elongated Gaussian components around four centers to eight representatives. Panel (b) reduces 100 unequally weighted univariate components to ten. Their common source variance permits ordering by mean. Both methods retain the four modes in (a), while HMW+ISE5 better preserves the lower-right peak height. In (b), the curves nearly coincide, and the inset shows where HMW+ISE5 follows the source more closely. HMW+ISE5 has lower nISE in both examples.

the group's mean and covariance. This moment matching includes both component covariances and the spread of their means. HMW+ISE5 then holds groups and weights fixed and jointly refines representative means and covariances for at most five iterations. Its objective sums source-weighted ISE costs between each source Gaussian and its representative, an upper bound on the full mixture ISE. Appendix C.1 gives the optimization details and the error bound.

Our baseline minimizes composite transportation divergence (CTD) (Zhang et al., 2024), which sums costs between Gaussians weighted by transported mass. We use ISE for each Gaussian pair, giving CTD-ISE. Its hard-assignment implementation alternates component reassignment with updates of representative weights, means, and covariances. Both methods start from the same moment-matched mixture. CTD uses at most three reassignment/update cycles in panel (a) and 1000 in panel (b). Continuing (a) to convergence leaves its result unchanged.

6.2 3D Gaussian Splatting scene compression

A 3D Gaussian Splatting (3DGS) scene (Kerbl et al., 2023) renders images from Gaussian primitives carrying position, shape, opacity, and appearance. GHAP (Gaussian Herding across Pens) (Wang et al., 2025) compresses 3DGS scenes by minimizing CTD within KD-tree partitions of Gaussian centers. Our spatial extension, labeled Hilbert, orders centers in $\mathbb{R}^3$ and forms blocks from consecutive groups. Each group initializes one representative. Refinement within blocks uses GHAP's geometric cost, the sum of squared differences in means and covariances.

We target 10%, 1%, and 0.1% retention of the original primitives in the trained Train and Truck scenes from Tanks and Temples (Knapitsch et al., 2017). Both methods' outputs undergo the same 5000-iteration fine-tuning on training images, updating geometry and appearance without changing primitive counts. Table 2 reports quality over all 38 Train and 32 Truck held-out views before and after fine-tuning, separating compression loss from subsequent recovery. PSNR (dB) measures pixel fidelity, SSIM structural similarity, and LPIPS perceptual discrepancy. Better reconstructions have higher PSNR and SSIM and lower LPIPS. Appendix C.3 gives the protocol and timings.

Table 2: **Rendering quality before and after fine-tuning.** Metrics cover all test views. Bold marks the better displayed value within each method pair and stage, including ties.

		Before fine-tuning			After fine-tuning		
Retained	Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Train: 38 test views							
10%	GHAP	14.676	0.4638	0.4962	21.193	0.7494	0.3138
	Hilbert	**15.297**	**0.4856**	**0.4843**	**21.413**	**0.7525**	**0.3106**
1%	GHAP	8.800	0.2314	0.6717	18.582	0.6034	0.4700
	Hilbert	**9.336**	**0.2649**	**0.6669**	**18.745**	**0.6059**	**0.4675**
0.1%	GHAP	5.636	0.0269	**0.7314**	15.213	0.4794	0.6067
	Hilbert	**6.102**	**0.0695**	0.7342	**15.690**	**0.4862**	**0.6026**
Truck: 32 test views							
10%	GHAP	14.474	0.5238	0.4611	24.527	0.8468	0.2273
	Hilbert	**15.531**	**0.5534**	**0.4465**	**24.556**	**0.8495**	**0.2215**
1%	GHAP	7.219	0.2204	0.6789	**21.513**	0.7106	0.3919
	Hilbert	**7.714**	**0.2546**	**0.6708**	**21.513**	**0.7128**	**0.3894**
0.1%	GHAP	4.873	0.0182	0.7642	17.639	0.5453	**0.5853**
	Hilbert	**4.904**	**0.0202**	**0.7635**	**17.764**	**0.5468**	0.5857

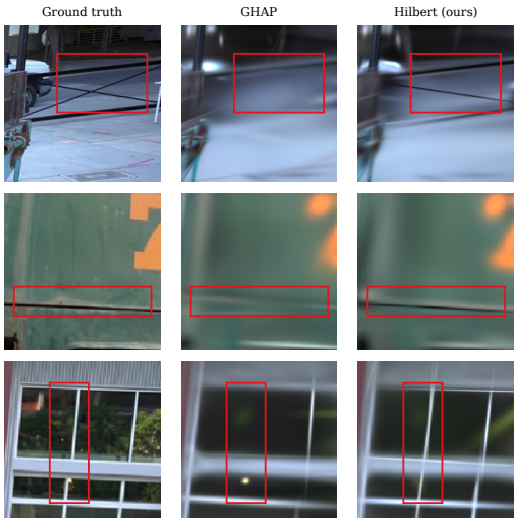


Figure 5: **Matched local details after common fine-tuning.** Red boxes mark the railing, panel seam, and window frame. Rows show Truck (1%), Train (1%), and Truck (0.1%), from top to bottom.

Hilbert improves initial PSNR and SSIM in all six scene–retention pairs and LPIPS in five. Fine-tuning restores much of the lost rendering quality for both methods, and their final full-view scores are close. Hilbert retains a small advantage in PSNR and SSIM, with PSNR tied at the displayed precision for Truck at 1% retention. Figure 5 shows differences within selected image regions after fine-tuning. The Hilbert reconstructions show a more distinct railing, panel seam, and window frame in these regions.

At 10% retention, Hilbert compresses Train and Truck 2.59× and 1.98× faster, respectively, with similar or slightly better final quality (Table 2). Fine-tuning dominates total time, reducing the corresponding overall savings to 3.2% and 1.7%. At 0.1%, Hilbert takes longer overall and gives slightly worse final LPIPS on Truck.

7 Discussion and Conclusion

Preparing a Gaussian mixture once allows HMW to reuse its ordered representation across comparisons. The resulting metric upper-bounds MW, with the excess separating the effects of fixed factor alignment and ordered matching. The covariance examples distinguish HMW from sliced distances, while timings show the savings from sparse matching and reuse. Ordered grouping also supplies initial representatives for density reduction and, through its spatial extension, 3DGS scene compression. Ordering stability warrants further study because small parameter changes can swap components and cause abrupt distance changes. A useful extension would reduce this sensitivity while retaining efficient reuse. For compression, adapting group boundaries to local covariance variation may improve the initial representatives.

AI USE STATEMENT

Generative AI assisted literature and source inspection, mathematical formulation and proof drafting, prose revision, and verification code. The authors are responsible for checking the final claims, proofs, experiments and citations, and for completing the submission disclosure.

REPRODUCIBILITY STATEMENT

Section 3 specifies the construction, and Appendix A gives all proofs. Appendices B and C record implementation conventions, diagnostic checks, application protocols, timing boundaries, and analytic density-error evaluation.

REFERENCES

Ravindra K. Ahuja, Thomas L. Magnanti, and James B. Orlin. *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, Englewood Cliffs, NJ, 1993. ISBN 978-0-13-617549-0.

E. Alpaydin and C. Kaynak. Optical recognition of handwritten digits. UCI Machine Learning Repository, 1998. URL https://doi.org/10.24432/C50P49.

Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.neurips.cc/paper_files/paper/2017/hash/491442df5f88c6aa018e86dac21d3606-Abstract.html.

Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019. doi: 10.1016/j.exmath.2018.01.002.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015. doi: 10.1007/s10851-014-0506-3.

Yongxin Chen, Tryphon T. Georgiou, and Allen Tannenbaum. Optimal transport for Gaussian mixture models. *IEEE Access*, 7:6269–6278, 2019. doi: 10.1109/ACCESS.2018.2889838.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013. URL https://papers.neurips.cc/paper_files/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

Julie Delon and Agnès Desolneux. A Wasserstein-type distance in the space of Gaussian mixture models. *SIAM Journal on Imaging Sciences*, 13(2):936–970, 2020. doi: 10.1137/19M1301047.

Ishan Deshpande, Yuan-Ting Hu, Ruoyu Sun, Ayis Pyrros, Nasir Siddiqui, Sanmi Koyejo, Zhizhen Zhao, David Forsyth, and Alexander G. Schwing. Max-sliced Wasserstein distance and its use for GANs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10648–10656, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Deshpande_Max-Sliced_Wasserstein_Distance_and_Its_Use_for_GANs_CVPR_2019_paper.html.

D. C. Dowson and B. V. Landau. The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, 1982. doi: 10.1016/0047-259X(82)90077-X.

Geneviève Dusson, Virginie Ehrlacher, and Nathalie Nouaime. A Wasserstein-type metric for generic mixture models, including location-scatter and group invariant measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 32:19, 2026. doi: 10.1051/cocv/2026004.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2681–2690, 2019. URL https://proceedings.mlr.press/v89/feydy19a.html.

Matthias Gelbrich. On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Mathematische Nachrichten*, 147(1):185–203, 1990. doi: 10.1002/mana.19901470121.

Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/2a27b8144ac02f67687f76782a3b5d8f-Abstract.html.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139:1–139:14, 2023. doi: 10.1145/3592433. URL https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and Temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4):78:1–78:13, 2017. doi: 10.1145/3072959.3073599. URL https://www.tanksandtemples.org/.

Soheil Kolouri, Gustavo K. Rohde, and Heiko Hoffmann. Sliced Wasserstein distance for learning Gaussian mixture models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3427–3436, 2018. doi: 10.1109/CVPR.2018.00361.

Soheil Kolouri, Kimia Nadjahi, Umut Simsekli, Roland Badeau, and Gustavo Rohde. Generalized sliced Wasserstein distances. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/f0935e4cd5920aa6c7c996a5ee53a70f-Abstract.html.

Mengyu Li, Jun Yu, Tao Li, and Cheng Meng. Importance sparsification for Sinkhorn algorithm. *Journal of Machine Learning Research*, 24(247):1–44, 2023. URL https://jmlr.org/papers/v24/22-1311.html.

Tao Li, Cheng Meng, Hongteng Xu, and Jun Yu. Hilbert curve projection distance for distribution comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(7):4993–5007, 2024. doi: 10.1109/TPAMI.2024.3363780.

Xinran Liu, Yikun Bai, Rocío Díaz Martín, Kaiwen Shi, Ashkan Shahbazi, Bennett A. Landman, Catie Chang, and Soheil Kolouri. Linear spherical sliced optimal transport: A fast metric for comparing spherical data. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/cc10fd7fe3e0d2af517cedf99fbc0651-Abstract-Conference.html.

Khai Nguyen and Nhat Ho. Sliced Wasserstein estimation with control variates. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/08f628998ca37c9df8c6a0df3570db86-Abstract-Conference.html.

Khai Nguyen, Nhat Ho, Tung Pham, and Hung Bui. Distributional sliced-Wasserstein and applications to generative modeling. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=QYjO70ACDK.

Khai Nguyen, Shujian Zhang, Tam Le, and Nhat Ho. Sliced Wasserstein with random-path projecting directions. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 37879–37899, 2024. URL https://proceedings.mlr.press/v235/nguyen24l.html.

François-Pierre Paty and Marco Cuturi. Subspace robust Wasserstein distances. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5072–5081, 2019. URL https://proceedings.mlr.press/v97/paty19a.html.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019. doi: 10.1561/2200000073.

Moritz Piening and Robert Beinert. Slicing the Gaussian mixture Wasserstein distance. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=yPBtJ4JPwi.

Mark Rowland, Jiri Hron, Yunhao Tang, Krzysztof Choromanski, Tamas Sarlos, and Adrian Weller. Orthogonal estimation of Wasserstein distances. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 186–195, 2019. URL https://proceedings.mlr.press/v89/rowland19a.html.

Meyer Scetbon, Marco Cuturi, and Gabriel Peyré. Low-rank Sinkhorn factorization. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9344–9354, 2021. URL https://proceedings.mlr.press/v139/scetbon21a.html.

Tao Wang, Mengyu Li, Geduo Zeng, Cheng Meng, and Qiong Zhang. Gaussian herding across pens: An optimal transport perspective on global Gaussian reduction for 3DGS. In *Advances in Neural Information Processing Systems*, volume 38, pp. 157898–157923, 2025. doi: 10.52202/085713-5277. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/e79574cc3355e831cc276c845605ed72-Abstract-Conference.html.

Wei Wang, Dejan Slepčev, Saurav Basu, John A. Ozolek, and Gustavo K. Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International Journal of Computer Vision*, 101(2):254–269, 2013. doi: 10.1007/s11263-012-0566-z.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Qiong Zhang, Archer Gong Zhang, and Jiahua Chen. Gaussian mixture reduction with composite transportation divergence. *IEEE Transactions on Information Theory*, 70(7):5191–5212, 2024. doi: 10.1109/TIT.2023.3323346.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

A COMPLETE PROOFS

We first derive the projected covariance response, then establish the ordered-plan metric properties, approximation bounds, and density-reduction result. All statements use exact arithmetic.

A.1 THE SLICED LOCAL RESPONSE

For independent uniform $\theta \in \mathbb{S}^{d-1}$ and $v \in \mathbb{S}^1$, set

$$z_\theta(G) = \left(\theta^\mathsf{T}\mu, \sqrt{\theta^\mathsf{T}\Sigma\theta}\right) \in \mathbb{R}^2, \qquad f_{\theta,v}(G) = v^\mathsf{T} z_\theta(G).$$

With normalized spherical probability measures, the population squared distances are

$$D^2_{\mathrm{MSW}} = \min_{\pi\in\Pi(a,b)} \sum_{ij} \pi_{ij}\mathbb{E}_\theta\|z_\theta(G_i) - z_\theta(F_j)\|^2,$$

$$D^2_{\mathrm{DSMW}} = \mathbb{E}_{\theta,v} W_2^2(f_{\theta,v\#}\mathcal{P}, f_{\theta,v\#}\mathcal{Q}).$$

Here $f_\#\mathcal{P}$ denotes the pushforward of $\mathcal{P}$ through f, and W_2 is the usual 2-Wasserstein distance. MSW uses a common component coupling across directions. DSMW sorts the scalar projections and matches cumulative weights separately for each direction pair. The expectations define the distances with all directions integrated out, so their differences from MW can be studied independently of numerical sampling error (Piening & Beinert, 2025). The following expansion quantifies how these distances respond to a small change in the covariance root.

Proposition 2 (Local covariance response of population slicing)**.** *Fix $d \geq 2$, $\sigma > 0$, and a symmetric matrix $S \in \mathbb{R}^{d\times d}$. Let $G_\star = \mathcal{N}(0, \sigma^2 I)$ and $G_\epsilon = \mathcal{N}(0, (\sigma I + \epsilon S)^2)$, with $|\epsilon|$ small enough that $\sigma I + \epsilon S \succ 0$. Set $\mathcal{P}_\star = \delta_{G_\star}$ and $\mathcal{Q}_\epsilon = \delta_{G_\epsilon}$. Then*

$$d^2_{\mathrm{BW}}(G_\star, G_\epsilon) = \epsilon^2\|S\|_F^2,$$

$$D^2_{\mathrm{MSW}}(\mathcal{P}_\star, \mathcal{Q}_\epsilon) = \epsilon^2 \frac{(\operatorname{tr} S)^2 + 2\|S\|_F^2}{d(d+2)} + O(|\epsilon|^3), \qquad D^2_{\mathrm{DSMW}} = \tfrac{1}{2} D^2_{\mathrm{MSW}}.$$

The remainder constant may depend on d, σ, S.

The leading coefficients measure squared distance per unit squared perturbation size. For BW this coefficient is $\|S\|_F^2$, whereas slicing also depends on $\operatorname{tr} S$. Decomposing S into its trace and traceless parts yields the unequal scale and shape weights in (1).

Proof of Proposition 2. Because the reference root is σI and the perturbed principal root is $\sigma I + \epsilon S$, the BW covariance term is exactly $\epsilon^2\|S\|_F^2$. For a unit θ, put $h = \theta^\mathsf{T} S\theta$ and $k = \theta^\mathsf{T} S^2\theta$. The projected standard deviation is

$$\sqrt{\sigma^2 + 2\sigma\epsilon h + \epsilon^2 k} = \sigma + \epsilon h + \frac{\epsilon^2}{2\sigma}(k - h^2) + O(|\epsilon|^3).$$

For fixed S, σ, this expansion is uniform on the sphere for sufficiently small $|\epsilon|$. Squaring its difference from σ and integrating gives $D^2_{\mathrm{MSW}} = \epsilon^2\mathbb{E}h^2 + O(|\epsilon|^3)$. Rotational symmetry gives

$$\mathbb{E}[\theta_i\theta_j\theta_k\theta_l] = \frac{\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}}{d(d+2)}.$$

Here δ_{ij} is the Kronecker delta. Multiplying by $S_{ij}S_{kl}$ and summing over the four indices yields $\mathbb{E}h^2 = ((\operatorname{tr} S)^2 + 2\|S\|_F^2)/(d(d+2))$. For singleton measures the second projection has no transport optimization, and $\mathbb{E}vv^\mathsf{T} = I_2/2$ implies the exact DSMW factor of one half. □

A.2 A MIXTURE COMPARISON CONSEQUENCE

Corollary 3 (Reversal of mixture comparisons)**.** *Let $d \geq 2$, $\sigma > 0$, and let a, b be positive probability vectors with fixed means $\mu_i, \nu_j \in \mathbb{R}^d$. For a symmetric $S \in \mathbb{R}^{d\times d}$, define*

$$\mathcal{P} = \sum_i a_i \delta_{\mathcal{N}(\mu_i, \sigma^2 I)}, \qquad \mathcal{Q}_{\epsilon,S} = \sum_j b_j \delta_{\mathcal{N}(\nu_j, (\sigma I + \epsilon S)^2)}.$$

Let $S_0 \in \mathbb{R}^{d\times d}$ be symmetric with $\operatorname{tr} S_0 = 0$ and $\|S_0\|_F = 1$, and choose c such that $1 < c^2 < (d+2)/2$. For sufficiently small nonzero ϵ,

$$D(\mathcal{P}, \mathcal{Q}_{\epsilon,cS_0}) > D(\mathcal{P}, \mathcal{Q}_{\epsilon,I/\sqrt{d}}),$$
$$D_\bullet(\mathcal{P}, \mathcal{Q}_{\epsilon,cS_0}) < D_\bullet(\mathcal{P}, \mathcal{Q}_{\epsilon,I/\sqrt{d}}), \qquad \bullet \in \{\mathrm{MSW}, \mathrm{DSMW}\}.$$

The comparison is between a uniform expansion and a larger traceless perturbation from the same reference mixture. The means stay fixed, so the reversed inequalities come entirely from how the three distances weight covariance changes.

Proof. Let $P_\mu = \sum_i a_i \delta_{\mu_i}$, $Q_\nu = \sum_j b_j \delta_{\nu_j}$, $C_\mu = W_2^2(P_\mu, Q_\nu)$, and

$$V_\epsilon(S) = \mathbb{E}_\theta \left(\sqrt{\theta^\mathsf{T}(\sigma I + \epsilon S)^2 \theta} - \sigma \right)^2.$$

The covariance cost is constant across all cross-component pairs, so

$$D^2 = C_\mu + \epsilon^2 \|S\|_F^2, \qquad D^2_{\mathrm{MSW}} = C_\mu/d + V_\epsilon(S).$$

For DSMW, fix θ and write $P_\theta = \theta_\# P_\mu$, $Q_\theta = \theta_\# Q_\nu$. Let their means be m_P, m_Q, put $T_\theta = W_2^2(P_\theta, Q_\theta)$, and set $\Delta s = \sigma - \sqrt{\theta^\mathsf{T}(\sigma I + \epsilon S)^2\theta}$. For $v = (v_1, v_2)$, expanding the squared cost gives

$$W_2^2(f_{\theta,v\#}\mathcal{P}, f_{\theta,v\#}\mathcal{Q}_{\epsilon,S}) = v_1^2 T_\theta + 2v_1 v_2 \Delta s (m_P - m_Q) + v_2^2 (\Delta s)^2.$$

The cross term depends only on the marginals, so it does not change the minimizing coupling. Scaling both projected mean distributions by the same v_1 multiplies their optimal squared cost by v_1^2, including negative v_1. The identity also holds at $v_1 = 0$. For uniform $v \in \mathbb{S}^1$, $\mathbb{E}(v_1 v_2) = 0$ and $\mathbb{E}v_1^2 = \mathbb{E}v_2^2 = 1/2$. Thus averaging eliminates the cross term without an equal-barycenter assumption and gives

$$D^2_{\mathrm{DSMW}} = \tfrac{1}{2}\mathbb{E}_\theta W_2^2(\theta_\# P_\mu, \theta_\# Q_\nu) + \tfrac{1}{2} V_\epsilon(S).$$

The radial term is exactly $V_\epsilon(I/\sqrt{d}) = \epsilon^2/d$. Proposition 2 gives $V_\epsilon(cS_0) = 2c^2\epsilon^2/[d(d+2)] + O(|\epsilon|^3)$. Thus the squared MW difference is $(c^2-1)\epsilon^2 > 0$, whereas the squared MSW difference is $\epsilon^2(2c^2/(d+2) - 1)/d + O(|\epsilon|^3) < 0$ for small nonzero ϵ. The DSMW difference is half the latter. Taking square roots preserves these strict comparisons. □

A.3 Weighted plans and factor pricing

Lemma 4 (Feasibility and sparsity of the weighted NW scan)**.** *For any two permutations of positive probability vectors a, b, the interval overlaps in* (4) *define a member of $\Pi(a,b)$ with at most $K+M-1$ positive entries.*

Each overlap is one feasible transport edge, so the plan can be constructed without evaluating a dense cost matrix.

Proof. The target intervals partition $(0,1]$, so $\sum_s |I_r \cap J_s| = |I_r| = a_{\alpha_r}$; the column statement is identical. Every positive overlap exhausts at least one current residual mass. The last overlap exhausts both remaining masses, hence at most $K + M - 1$ positive overlaps occur. The union of the interval breakpoints gives the same bound directly. □

Theorem 5 (MW upper bound)**.** *For $G, F \in \mathcal{G}_d^+$, $c_{\mathrm{BW}}(G,F) \le c_\Phi(G,F)$. Every $\pi \in \Pi(a,b)$ satisfies*

$$D^2 \le \sum_{ij} \pi_{ij} c_{\mathrm{BW}}(G_i, F_j) \le \sum_{ij} \pi_{ij} c_\Phi(G_i, F_j).$$

In particular, $D \le U$.

A feasible coupling under fixed factors therefore provides an upper bound on the optimal Gaussian and mixture costs.

Proof of Theorem 5. The random pair $(\mu + L_\Sigma Z, \nu + L_\Lambda Z)$ has the desired Gaussian marginals. Its expected squared displacement is $\|\mu - \nu\|^2 + \|L_\Sigma - L_\Lambda\|_F^2 = c_\Phi(G, F)$. The Gaussian Wasserstein optimum cannot exceed the cost of this coupling. For an outer feasible plan, multiply the ground inequality by π_{ij} and sum. The BW outer optimum is no larger than its value at that plan. Lemma 4 supplies feasibility for HMW. □

Lemma 6 (Exact factor-alignment gap)**.** *For $L = L_\Sigma$, $R = L_\Lambda$ and $M = L^\mathsf{T} R$,*

$$\Delta_{\mathrm{chol}}(G, F) = c_\Phi(G, F) - c_{\mathrm{BW}}(G, F) = 2(\|M\|_* - \operatorname{tr} M) \geq 0.$$

Here $\|M\|_*$ is the nuclear norm, the sum of its singular values. The gap is zero precisely when M is symmetric positive definite. In particular it vanishes for diagonal covariances.

The difference $\|M\|_* - \operatorname{tr} M$ is the improvement in the cross term obtained by optimizing the relative factor alignment instead of fixing $O = I$.

Proof. The squared Gaussian covariance distance has the orthogonal factor-alignment form $\min_{O^\mathsf{T} O = I} \|L - RO\|_F^2$. The cross term maximizes $\operatorname{tr}(L^\mathsf{T} RO)$, whose maximum is $\|L^\mathsf{T} R\|_*$ by a singular value decomposition. Its singular values equal those of $\Sigma^{1/2}\Lambda^{1/2}$ because covariance factors differ from principal roots by right orthogonal factors. Subtracting the value at $O = I$ gives the formula. Write $M = U \operatorname{diag}(s_i) V^\mathsf{T}$. Here $s_i > 0$ by positive definiteness. Equality $\operatorname{tr} M = \sum_i s_i$ requires every diagonal entry of $V^\mathsf{T} U$ to equal one, and hence $U = V$. Since M is invertible, this is equivalent to its being symmetric positive definite. Diagonal Cholesky factors satisfy the condition. □

A.4 The metric theorem

Proof of Theorem 1. Intersecting the intervals in the two ordered mass representations yields

$$\int_0^1 \|Z_\mathcal{P}(u) - Z_\mathcal{Q}(u)\|^2 \, du = \sum_{r,s} |I_r \cap J_s| \|\Phi(G_{\alpha_r}) - \Phi(F_{\beta_s})\|^2 = U^2.$$

Finite support makes all these functions square-integrable. Nonnegativity and symmetry follow from the norm. Canonical ordering ensures that equal atom measures yield the same function almost everywhere, independently of how their lists were presented. Conversely, $U = 0$ makes the two step functions equal almost everywhere. Evaluating $Z_\mathcal{P}$ at a uniform point in $(0, 1]$ returns $\Phi(G_i)$ with probability a_i, so its distribution is $\Phi_\# \mathcal{P}$. The same observation for $\mathcal{Q}$ gives $\Phi_\# \mathcal{P} = \Phi_\# \mathcal{Q}$. The feature map is injective because its coordinates recover μ and L, hence $\Sigma = LL^\mathsf{T}$. Thus $\mathcal{P} = \mathcal{Q}$. For three measures, each has a fixed function under the same rule. Minkowski's inequality gives

$$\|Z_\mathcal{P} - Z_\mathcal{R}\|_2 \leq \|Z_\mathcal{P} - Z_\mathcal{Q}\|_2 + \|Z_\mathcal{Q} - Z_\mathcal{R}\|_2.$$

The common mass interval therefore gives the triangle inequality across separately constructed orders. □

The BW edge metric. Let $\Gamma_\mathcal{P}$ be the Gaussian-valued step function determined by the same canonical order and mass intervals as $Z_\mathcal{P}$. Then

$$V(\mathcal{P}, \mathcal{Q}) = \left[\int_0^1 d_{\mathrm{BW}}^2\big(\Gamma_\mathcal{P}(u), \Gamma_\mathcal{Q}(u)\big) \, du \right]^{1/2}.$$

Pointwise triangle inequalities for d_{BW} and Minkowski prove the triangle inequality for V. If $V = 0$, the atom-valued functions agree almost everywhere, and their pushforwards of uniform mass recover the same atom measure. Symmetry and nonnegativity follow immediately. Thus V is a metric under the same canonical-order convention as U.

A.5 Local Cholesky response

Proposition 7 (Local response of Cholesky pricing)**.** *Under the assumptions of Proposition 2,*

$$c_\Phi(G_\star, G_\epsilon) = \epsilon^2 \left(\sum_i S_{ii}^2 + 4 \sum_{i>j} S_{ij}^2 \right) + O(|\epsilon|^3). \tag{7}$$

For $S \neq 0$, its limiting ratio to $d^2_{\mathrm{BW}}(G_\star, G_\epsilon)$ is in $[1, 2]$. For diagonal S, the costs are exactly equal whenever $\sigma I + \epsilon S \succ 0$.

BW has quadratic coefficient $\sum_i S_{ii}^2 + 2\sum_{i>j} S_{ij}^2$. Thus Cholesky costs give the same coefficient to diagonal perturbations and twice the BW coefficient to off-diagonal perturbations. The ratio concerns the limit as $\epsilon \to 0$ around this isotropic reference.

Table 3 compares Proposition 7 with the radial calibration in (1) and the principal-root cost c_Ψ defined below.

Table 3: Coefficients of ϵ^2 for the isotropic reference in Proposition 2, with $\|S\|_F = 1$. MSW is calibrated to agree with BW in the radial direction. Singleton DSMW has the same calibrated column after multiplication by $2d$.

Perturbation S	BW	dD^2_{MSW}	c_Φ	c_Ψ
Radial, $I/\sqrt{d}$	1	1	1	1
Diagonal traceless	1	$2/(d+2)$	1	1
Symmetric with zero diagonal	1	$2/(d+2)$	2	1

Define $\Delta(G, F) = c_\Phi(G, F) - c_{\mathrm{BW}}(G, F)$. This nonnegative gap is zero precisely when $L_\Sigma^\mathsf{T} L_\Lambda$ is symmetric positive definite, including diagonal covariances. Lemma 6 gives its nuclear-norm expression and proof.

Proof of Proposition 7. The Cholesky map is smooth in a neighborhood of the SPD matrix $\sigma^2 I$. Write $L_\epsilon = \sigma I + \epsilon T + O(\epsilon^2)$ with lower-triangular T. Differentiating $L_\epsilon L_\epsilon^\mathsf{T} = (\sigma I + \epsilon S)^2$ at zero gives $T + T^\mathsf{T} = 2S$. Consequently $T_{ii} = S_{ii}$, $T_{ij} = 2S_{ij}$ for $i > j$, and $T_{ij} = 0$ for $i < j$. Squaring $\|L_\epsilon - \sigma I\|_F$ proves (7). Since

$$\|S\|_F^2 = \sum_i S_{ii}^2 + 2\sum_{i>j} S_{ij}^2,$$

the coefficient lies between $\|S\|_F^2$ and $2\|S\|_F^2$. For diagonal S, the positive diagonal matrix $\sigma I + \epsilon S$ is itself the Cholesky factor, giving exact equality with BW. For a traceless diagonal perturbation the limiting ratio is one; rotating it to a purely off-diagonal symmetric perturbation can make the ratio two, exhibiting the coordinate dependence of Cholesky pricing. □

Principal-root comparison. Principal-root pricing replaces Cholesky factors by positive square roots,

$$c_\Psi(G, F) = \|\mu - \nu\|^2 + \|\Sigma^{1/2} - \Lambda^{1/2}\|_F^2.$$

It upper-bounds BW and is exact in the isotropic model of Proposition 2. Factor choice affects both the pairwise cost and the sorting coordinates. Define $\Psi(G) = (\mu, \mathrm{svec}(\Sigma^{1/2})) \in \mathbb{R}^p$, where svec lists the diagonal once and each distinct off-diagonal entry multiplied by $\sqrt{2}$. Then $\|\Psi(G) - \Psi(F)\|^2 = c_\Psi(G, F)$. Coupling through the same standard Gaussian noise proves the BW upper bound. In the isotropic perturbation model, the positive principal roots are σI and $\sigma I + \epsilon S$, giving $c_\Psi = \epsilon^2 \|S\|_F^2 = c_{\mathrm{BW}}$ exactly.

A.6 Conditional approximation

The coupling contribution B depends on the BW costs after arranging components in their canonical orders. A Monge cost matrix favors matching smaller ranks to smaller ranks, so the NW plan is optimal. The next result controls the added cost when the cost matrix is close to a Monge matrix in every entry.

Theorem 8 (Accuracy under near-Monge costs)**.** *Let $C_{rs} = c_{\mathrm{BW}}(G_{\alpha_r}, F_{\beta_s})$ and suppose $\|C - C_0\|_{\max} \leq \varepsilon$ for a Monge matrix $C_0 \in \mathbb{R}^{K \times M}$, where $\varepsilon \geq 0$ and $\|\cdot\|_{\max}$ is the largest absolute entry. Thus $(C_0)_{rs} + (C_0)_{r's'} \leq (C_0)_{rs'} + (C_0)_{r's}$ for $r < r'$ and $s < s'$. Then*

$$0 \leq U^2 - D^2 \leq A + 2\varepsilon, \qquad 0 \leq U - D \leq \sqrt{A + 2\varepsilon}.$$

If $\Delta_{ij} \leq \delta$ on the support of π^{H}, then $A \leq \delta$.

The factor term A remains unchanged, while the deviation from Monge costs contributes at most 2ε to the squared-distance gap.

Lemma 9 (Near-Monge optimality of the NW plan)**.** *Let $C, C_0 \in \mathbb{R}^{K\times M}$ with $\|C - C_0\|_{\max} \le \varepsilon$, where C_0 is Monge in the two given orders. For positive probability marginals, their NW plan satisfies*

$$0 \le \langle C, \pi^{\mathrm{H}}\rangle - \min_{\pi\in\Pi(a,b)} \langle C, \pi\rangle \le 2\varepsilon.$$

Changing every entry by at most ε changes any feasible plan's total cost by at most ε, since the plan has mass one. Applying this once to the NW plan and once to an optimal plan gives the factor two.

Proof. First, NW is optimal for C_0. Start from an optimal plan. If its upper-left entry is smaller than the minimum of the first row and column masses, there are positive entries to its right and below it. Transfer the smaller of those masses from these two entries to the upper-left and opposite rectangle corners. Marginals are preserved and the Monge inequality prevents a cost increase. Repeating exhausts a row or column; induction on the remaining matrix produces NW. Each transfer either fills the required corner mass or exhausts an entry in its row or column, so this construction terminates in finitely many steps. Let π_* be optimal for C. The unit total mass of both plans gives

$$\langle C, \pi^{\mathrm{H}}\rangle \le \langle C_0, \pi^{\mathrm{H}}\rangle + \varepsilon \le \langle C_0, \pi_*\rangle + \varepsilon \le \langle C, \pi_*\rangle + 2\varepsilon.$$

The lower bound follows from optimality under C. □

Proof of Theorem 8. Subtracting the optimal BW cost from the feature cost gives (6). Theorem 5 makes every Δ_{ij} nonnegative. The BW cost of the feasible Hilbert plan is at least the optimal cost D^2, making B nonnegative as well. Lemma 9 bounds that second term by 2ε. Since $\sum_{ij} \pi^{\mathrm{H}}_{ij} = 1$, bounding Δ_{ij} by δ on the support of this plan gives $A \le \delta$. Finally $(U - D)^2 \le U^2 - D^2$ for $U \ge D \ge 0$, yielding the distance bound. □

Corollary 10 (An exact subclass and perturbations preserving the orders)**.** *Let $\mathcal{P}^0 = \sum_i a_i \delta_{G_i^0}$ and $\mathcal{Q}^0 = \sum_j b_j \delta_{F_j^0}$ have diagonal SPD covariances and features on a common affine line,*

$$\Phi(G_i^0) = z + t_i v, \qquad \Phi(F_j^0) = z + s_j v, \qquad z, v \in \mathbb{R}^p, \quad v \ne 0.$$

Suppose the two returned orders are both nondecreasing, or both nonincreasing, in t_i and s_j. Then $U(\mathcal{P}^0, \mathcal{Q}^0) = D(\mathcal{P}^0, \mathcal{Q}^0)$. For perturbed SPD atoms G_i, F_j, preserve the positive weights, support cardinalities, atom correspondences, and returned order of each list. Define

$$\eta_P^2 = \sum_i a_i \|\Phi(G_i) - \Phi(G_i^0)\|^2, \qquad \eta_Q^2 = \sum_j b_j \|\Phi(F_j) - \Phi(F_j^0)\|^2.$$

Then

$$0 \le U(\mathcal{P}, \mathcal{Q}) - D(\mathcal{P}, \mathcal{Q}) \le 2(\eta_P + \eta_Q).$$

On the reference line, matching ranks minimizes the squared Euclidean cost, and diagonal covariances make that cost equal to BW. The quantities η_P and η_Q measure the root mean squared feature displacement from this exact configuration. Keeping the weights and orders fixed preserves the mass intervals, allowing these displacements to control the HMW–MW gap.

Proof. For diagonal covariances, factor pricing is exact and the reference cost is $C_{ij}^0 = \|v\|^2 (t_i - s_j)^2$. For similarly ordered indices,

$$C_{ij}^0 + C_{i'j'}^0 - C_{ij'}^0 - C_{i'j}^0 = -2\|v\|^2 (t_i - t_{i'})(s_j - s_{j'}) \le 0.$$

Thus NW is optimal by Lemma 9 with $\varepsilon = 0$, and $U^0 = D^0$. Fixed orders and weights give the same cumulative mass intervals before and after perturbation. Therefore $\|Z_{\mathcal{P}} - Z_{\mathcal{P}^0}\|_2 = \eta_P$ and similarly for Q, so the reverse triangle inequality in L^2 yields $|U - U^0| \le \eta_P + \eta_Q$. MW is the 2-Wasserstein metric over Gaussian atoms with ground metric d_{BW}. Coupling each perturbed atom to its reference atom and using Theorem 5 gives $\mathrm{MW}_{\mathrm{BW}}(\mathcal{P}, \mathcal{P}^0) \le \eta_P$ and similarly for Q. Its triangle inequality then gives $|D - D^0| \le \eta_P + \eta_Q$. Combining these inequalities with $U^0 = D^0$ proves the upper bound; Theorem 5 proves nonnegativity. The result applies when the returned orders have the stated directions and remain fixed under perturbation. □

A.7 DISCONTINUITY OF A SINGLE CANONICAL ORDERING

Remark 11 (Order changes and strict topology). For every $d \geq 1$, a construction that lists each atom once in a deterministic, comparison-independent order cannot make the resulting L^2 distance continuous everywhere with respect to MW on the full Gaussian parameter space. On a compact family of equal-weight two-atom mixtures, there are pairs with $D_n \to 0$ and $U_n \to 2r$ for a fixed $r > 0$. Thus no finite global bound $U \leq CD$, or uniform bound $U \leq \omega(D)$ with $\omega(r) \to 0$, holds for this construction. The topology induced by U is strictly finer than the MW topology.

The construction moves two distinct atoms around a closed path that returns to the same unordered mixture. An ordering must return the same first atom at the end, which forces it to switch between the two moving atoms somewhere along the path. Near such a switch, optimal matching follows the nearby atoms, while matching by rank pairs atoms that remain far apart.

Proof. For $d \geq 2$, let $m_t = (\cos t, \sin t, 0, \ldots, 0) \in \mathbb{R}^d$ and define

$$\mathcal{P}_t = \tfrac{1}{2}\delta_{\mathcal{N}(m_t, I_d)} + \tfrac{1}{2}\delta_{\mathcal{N}(-m_t, I_d)}, \qquad t \in [0, \pi].$$

The map $t \mapsto \mathcal{P}_t$ is continuous in MW and $\mathcal{P}_0 = \mathcal{P}_\pi$. Let $b_t = \xi_t m_t$, $\xi_t \in \{-1, 1\}$, be the mean of the first atom selected by the canonical order. Since the two weights are equal and the covariance features coincide, the interval representation gives

$$U(\mathcal{P}_t, \mathcal{P}_s) = \|b_t - b_s\|.$$

If the ordering representation were continuous everywhere in MW, then b_t would be continuous in t. The sign $\xi_t = b_t^\mathsf{T} m_t$ would be a continuous function from $[0, \pi]$ to $\{-1, 1\}$ and hence constant. This would give $b_\pi = -b_0$, contrary to canonicality and $\mathcal{P}_0 = \mathcal{P}_\pi$.

For an explicit sequence, take $n \geq 3$ and the grid $t_j = j\pi/n$. Canonicality requires $\xi_{t_n} = -\xi_{t_0}$, so at least one adjacent pair t_j, t_{j+1} has opposite signs. The optimal matching uses the closer pair of antipodal means, whereas the ordered matching uses the opposite pair. Their distances are

$$D_n = 2\sin\Big(\frac{\pi}{2n}\Big) \longrightarrow 0, \qquad U_n = 2\cos\Big(\frac{\pi}{2n}\Big) \longrightarrow 2.$$

Hence $U_n/D_n \to \infty$, which rules out a finite global multiplicative bound and any uniform modulus tending to zero. The entire construction lies in an MW-compact family. Take a subsequence of the constructed pairs with $\mathcal{P}_n \to \mathcal{P}_\star$ in MW. Since $D(\mathcal{P}_n, \mathcal{Q}_n) \to 0$, also $\mathcal{Q}_n \to \mathcal{P}_\star$. The triangle inequality for U gives

$$\max\{U(\mathcal{P}_n, \mathcal{P}_\star), U(\mathcal{Q}_n, \mathcal{P}_\star)\} \geq U(\mathcal{P}_n, \mathcal{Q}_n)/2 \longrightarrow 1.$$

Selecting a member attaining this maximum produces a sequence converging to $\mathcal{P}_\star$ in MW but not in HMW. Combined with $D \leq U$, this proves that the HMW topology is strictly finer for $d \geq 2$.

For $d = 1$, fix $\sigma_0 > r > 0$ and use

$$G_t^\pm = \mathcal{N}\big(\pm r\cos t, (\sigma_0 \pm r\sin t)^2\big), \qquad \mathcal{P}_t = \tfrac{1}{2}\delta_{G_t^+} + \tfrac{1}{2}\delta_{G_t^-}.$$

The mean–standard-deviation features are $(0, \sigma_0) \pm r(\cos t, \sin t)$. The variances stay positive, and the two atoms remain distinct. Univariate Gaussian BW distance equals Euclidean distance in these coordinates, so the same sign argument yields

$$D_n = 2r\sin(\pi/(2n)) \longrightarrow 0, \qquad U_n = 2r\cos(\pi/(2n)) \longrightarrow 2r.$$

The compactness and subsequence argument above applies unchanged. In both constructions, $V = U$, so the same topological conclusion holds for V. The obstruction already appears when two Gaussian parameters can vary; it does not apply to a scalar location family with fixed variance and monotone mean ordering. □

A.8 FIXED-GROUP REDUCTION

We use the density and grouping notation of Appendix C.1. The source density is $f = \sum_i a_i\varphi_i$, and each nonempty group C_j is represented by a Gaussian density ψ_j with weight $b_j = \sum_{i \in C_j} a_i$. The reduced density is $g = \sum_j b_j\psi_j$. Moment matching initializes the representatives by (9); refinement minimizes the componentwise surrogate $J = \sum_j \sum_{i \in C_j} a_i\|\varphi_i - \psi_j\|_{L^2}^2$ in (10). The results below connect this surrogate to the full density error and then to the initial feature distortion.

Lemma 12 (Moment matching preserves moments and positivity)**.** *For positive source weights and SPD covariances,* (9) *gives* $b_j > 0$*,* $\Lambda_j^{(0)} \succ 0$*, and a reduced mixture with the same mean and raw second moment as the source mixture.*

Moment matching therefore gives a valid reduced mixture before refinement.

Proof. Each covariance is a positive weighted sum of SPD matrices plus positive-semidefinite scatter terms. Also $\sum_j b_j \nu_j^{(0)} = \sum_i a_i \mu_i$. Expanding the scatter around $\nu_j^{(0)}$ gives

$$b_j(\Lambda_j^{(0)} + \nu_j^{(0)}\nu_j^{(0)\mathsf{T}}) = \sum_{i \in C_j} a_i(\Sigma_i + \mu_i \mu_i^{\mathsf{T}}).$$

Summing proves second-moment preservation at initialization. Subsequent ISE refinement can change these moments. □

Lemma 13 (Assignment with free representative weights)**.** *For fixed representatives and finite costs* c_{ij}*,*

$$\min_{\pi \geq 0,\ \pi \mathbf{1} = a} \sum_{ij} \pi_{ij} c_{ij} = \sum_i a_i \min_j c_{ij}.$$

Assigning every row to a minimizer and setting $b_j = \sum_i \pi_{ij}$ *attains the minimum.*

Free representative weights let the assignment separate across source components, as in the CTD update.

Proof. Every row is a nonnegative allocation of total mass a_i, whose cost is at least $a_i \min_j c_{ij}$. Concentrating it on a minimizer attains this bound, independently for each row. □

Lemma 14 (Density ISE bound and finite-step descent)**.** *For a fixed grouping, let* $\pi_{ij} = a_i \mathbf{1}_{\{i \in C_j\}}$ *and* $b_j = \sum_i \pi_{ij}$*. With SPD Gaussian densities,*

$$J - \|f - g\|_{L^2}^2 = \mathbb{E}_{(i,j)\sim\pi} \|(\varphi_i - \psi_j) - (f - g)\|_{L^2}^2 \geq 0.$$

If a joint representative update returns the valid trial with smallest J*, including the initialization among its candidates, then the total* J *does not increase. Nonincreasing updates to every* J_j *are also sufficient. This conclusion holds for any finite update budget.*

The full density difference averages the componentwise differences, whereas J averages their squared norms. Their difference is a nonnegative variance, so J upper-bounds the density error.

Proof. All SPD Gaussian densities belong to L^2. The random Hilbert-space element $V = \varphi_i - \psi_j$ has expectation $f - g$ and squared-norm expectation J. Expanding $\mathbb{E}\|V - \mathbb{E}V\|^2 = \mathbb{E}\|V\|^2 - \|\mathbb{E}V\|^2$ proves the identity. Retaining the initialization as a candidate directly gives the joint-update claim. Since $J = \sum_j b_j J_j$ with fixed nonnegative b_j, nonincreasing group objectives are an alternative sufficient condition. The descent property concerns the surrogate J. The variance slack can vary between iterates, so the full-density error need not decrease. □

Lemma 15 (Gaussian density regularity in Cholesky coordinates)**.** *Let* $G = \mathcal{N}(\mu, \Sigma)$ *and* $F = \mathcal{N}(\nu, \Lambda)$ *satisfy* $\lambda I \preceq \Sigma, \Lambda \preceq \bar{\lambda} I$ *for* $0 < \lambda \leq \bar{\lambda}$*. Their densities obey*

$$\|\varphi_G - \varphi_F\|_{L^2}^2 \leq C_{d,\lambda,\bar{\lambda}}\, c_\Phi(G, F), \qquad C_{d,\lambda,\bar{\lambda}} = (4\pi\lambda)^{-d/2} \max\left\{ \frac{1}{2\lambda}, \frac{(d+2)\bar{\lambda}}{4\lambda^2} \right\}. \tag{8}$$

The eigenvalue bounds keep the Gaussians from becoming arbitrarily narrow or diffuse. Under these bounds, a small Euclidean displacement of the mean and Cholesky features implies a small L^2 change in density.

Proof. For a Gaussian density $q = \varphi_{\mu,\Sigma}$, a mean direction h, and a symmetric covariance direction H, differentiation gives

$$Dq[h,H](x) = q(x)\left[h^\mathsf{T}\Sigma^{-1}(x-\mu) + \tfrac{1}{2}\big((x-\mu)^\mathsf{T}\Sigma^{-1}H\Sigma^{-1}(x-\mu) - \operatorname{tr}(\Sigma^{-1}H)\big)\right].$$

The normalized density proportional to q^2 is $\mathcal{N}(\mu, \Sigma/2)$ and $\|q\|_2^2 = (4\pi)^{-d/2}(\det\Sigma)^{-1/2}$. Writing $T = \Sigma^{-1/2}H\Sigma^{-1/2}$, Gaussian second and fourth moments, with odd cross terms zero, give

$$\|Dq[h,H]\|_2^2 = \|q\|_2^2\left[\tfrac{1}{2}h^\mathsf{T}\Sigma^{-1}h + \tfrac{1}{16}(\operatorname{tr}T)^2 + \tfrac{1}{8}\|T\|_F^2\right].$$

Along the straight parameter segment between G and F, covariances remain at least λI. Using $(\operatorname{tr}T)^2 \le d\|T\|_F^2$ and $\|T\|_F \le \|H\|_F/\lambda$, integration of this derivative in L^2 and Minkowski's inequality yield

$$\|\varphi_G - \varphi_F\|_2^2 \le (4\pi\lambda)^{-d/2}\left[\frac{\|\mu-\nu\|^2}{2\lambda} + \frac{d+2}{16\lambda^2}\|\Sigma-\Lambda\|_F^2\right].$$

This differentiation and integration are valid in L^2: the parameter segment is compact and uniformly SPD, so its density derivatives admit an integrable Gaussian envelope times a polynomial. For Cholesky factors L, R, the endpoint upper spectral bounds imply

$$\|\Sigma-\Lambda\|_F = \|(L-R)L^\mathsf{T} + R(L-R)^\mathsf{T}\|_F \le 2\sqrt{\bar\lambda}\,\|L-R\|_F.$$

Substituting and taking the larger coefficient proves (8). □

Proposition 16 (Density error from grouping distortion)**.** *Suppose $\lambda I \preceq \Sigma_i, \Lambda_j^{(0)} \preceq \bar\lambda I$ for all source and initial representative covariances, with $0 < \lambda \le \bar\lambda$. Let $E_0 = \sum_j \sum_{i\in C_j} a_i c_\Phi(G_i, F_j^{(0)})$. Any finite refinement preserving groups and weights and returning SPD representatives with $J_T \le J_0$ satisfies*

$$\|f - g_T\|_{L^2}^2 \le J_T \le J_0 \le C_{d,\lambda,\bar\lambda}E_0,$$

where (8) *gives the explicit constant.*

The quantity E_0 measures the weighted squared feature distance from each source component to its initial representative. A small E_0 therefore controls the final density error whenever refinement keeps J below its initial value. The covariance eigenvalue bounds apply to the source and initialization; the refined representatives need only remain SPD and satisfy the descent condition.

Proof of Proposition 16. Apply Lemma 15 to each source atom and its initial representative, multiply by a_i, and sum to obtain $J_0 \le C_{d,\lambda,\bar\lambda}E_0$. The accepted joint update satisfies $J_T \le J_0$, and Lemma 14 gives $\|f - g_T\|_2^2 \le J_T$. The final representatives' SPD covariances ensure that the ISE identity remains valid. □

B Implementation and diagnostic details

B.1 Canonical implementation and verification

Algorithm 1 implements the interval representation by advancing through the two ordered lists and pricing each positive overlap.

The canonical procedure removes zero weights, merges exactly identical atoms, normalizes positive mass, and lexicographically enumerates Cholesky features before recursive ordering. Duplicate detection uses exact equality of the original means and covariance entries. Weights are summed with compensated accumulation. After normalization, any floating-point residual is added to the largest weight, with ties resolved by its first position in the canonical list. Covariances with positive weight must be finite, symmetric, and SPD. Input covariances and feature scales are retained, and distinct atoms remain separate.

The feature layout lists the d mean coordinates first, followed by lower-triangular Cholesky entries in row order, including the diagonal. The backend is the unmodified official HCP `base.cpp` at commit `930ecc98` (Li et al., 2024). It starts from coordinate zero with all orientation flags false

Algorithm 1 Hilbert Mixture Wasserstein.

Require: Finite SPD Gaussian mixtures $\mathcal{P}, \mathcal{Q}$
1: Remove zero weights, merge exact duplicate atoms, and normalize weights
2: Compute $x_i = \Phi(G_i)$ and $y_j = \Phi(F_j)$ and enumerate lexicographically
3: $\alpha \leftarrow \mathcal{S}(x_1, \ldots, x_K)$; $\beta \leftarrow \mathcal{S}(y_1, \ldots, y_M)$
4: Set $r = s = 1$, initialize their residual masses, and set $\mathsf{cost} \leftarrow 0$
5: **while** both ordered lists have positive residual mass **do**
6: $\quad \omega \leftarrow \min\{\text{current source mass}, \text{current target mass}\}$
7: $\quad \mathsf{cost} \leftarrow \mathsf{cost} + \omega \|x_{\alpha_r} - y_{\beta_s}\|^2$
8: $\quad$ Subtract ω from both masses and advance exhausted positions to their next masses
9: **return** $U = \sqrt{\mathsf{cost}}$ and, optionally, the positive overlap edges

and uses unweighted median partitions through `std::nth_element`. Its comparator uses the selected coordinate alone. Lexicographic enumeration gives the backend the same input list for every relabeling of a mixture. With the backend build fixed, this makes its output independent of user-supplied labels. When partition coordinates tie, however, `std::nth_element` can arrange them differently across standard-library implementations. The canonical-order convention here fixes the backend build as part of the sorting rule. The recursion can terminate before visiting every feature coordinate. At $n = 128$, it inspects seven coordinates for $d \in \{3, 5, 10, 20\}$ and all five coordinates for $d = 2$. Because means come first, covariance changes cannot affect this ordering at $d = 10, 20$, although the coupling is still priced using every Cholesky coordinate. Replacing covariances while fixing means and weights confirms this behavior in all 60 checks per dimension.

Computational cost. Let N_0 be the supplied atom count and $N = K + M$ the count after canonicalization. Dense Cholesky factorization costs $O(Nd^3)$ time and $O(Nd^2)$ storage; preprocessing costs at most $O(pN_0 \log N_0)$. Balanced median recursion with linear-time selection uses $O(N \log N)$ scalar comparisons, with coordinate operations depending on p and the sorter. Given the orders, matching and cost evaluation require $O(Np)$ time and $O(N)$ plan storage. Algorithm 1 gives the mass sweep.

The weighted NW routine receives the canonical positive weights and the two permutations. The implementation checks the returned permutations, mass nonnegativity and marginal residuals, removes exact-zero edges, and prices the positive edges in the complete feature space. Returned indices refer to the canonical lists, which are included with the result. The canonical complete-call benchmarks include preprocessing, factorization, ordering, transport, and pricing. Table 4 distinguishes the timing series.

Regression checks cover input permutations, duplicate splitting and merging, zero weights, tied coordinates, marginal conservation, and exact MW comparisons on small instances in dimensions $1, 2, 3, 5, 10$. The maximum marginal residual over the twenty random-pair checks was 2.8×10^{-16}. Across the 161 points of each mechanism construction, computed HMW agrees with the reference curve to 2.8×10^{-16} in dimension two and to the recorded precision in dimension three.

Table 4: Timing series and reuse boundaries. Times are compared within each row. The server series uses separate inputs and software.

Series	Inputs and runtime	Timed operations and direction sampling
Figure 6	140 pairs; Python 3.9, NumPy	Canonical preparation, factors, orders, pricing, transport. No sliced methods.
Figure 3	Paired perturbations; separate server	Online preprocessing, matching and pricing. HMW omits duplicate merging and canonical enumeration. Sliced methods use 16 normalized Uniform$(0, 1)$ directions or direction pairs.
Table 1	Joint image GMMs; Python 3.10, NumPy	Cached comparisons after canonical preparation. MSW caches 16 spatial projections; DSMW caches 4/16/64/256 direction pairs.

B.2 DIRECT DISTANCE EXPERIMENT

The experiment contains 140 mixture pairs across seven (n, d) settings, two covariance regimes, and ten seeds per setting. We vary $n \in \{32, 128, 512\}$ at $d = 3$ and $d \in \{2, 3, 5, 10, 20\}$ at $n = 128$, counting the common $(128, 3)$ setting once. For $K = M = n$, source and target means are independent $\mathcal{N}(0, I_d)$ draws, weights are independently $\mathrm{Dirichlet}(2, \ldots, 2)$, and each covariance is $ZZ^\mathsf{T}/d + 0.25 I_d$, with independent standard Gaussian entries in Z. The dense and diagonal versions share means, weights, and diagonal covariance entries. All component weights are positive, all covariances are SPD, and no duplicate atoms occur.

MW uses the full BW matrix and an unregularized network-simplex solve through POT 0.9.4. For Cholesky factors L_i, L_j, its covariance cost is $\|L_i\|_F^2 + \|L_j\|_F^2 - 2\|L_i^\mathsf{T} L_j\|_*$, where $\|\cdot\|_*$ is the nuclear norm, evaluated by singular values. BW-edges uses this same formula on the HMW support; its edge indices and masses agree exactly with HMW. Each path includes validation, canonicalization, factorization, construction of its costs, and coupling computation. Diagonal inputs use the generic SPD routines in all methods.

Runs use an Intel Xeon Platinum 8280 CPU at a nominal 2.70 GHz, pinned to one core, with one numerical-library thread. Python is version 3.9.12 and NumPy 1.26.4. Three rounds independently randomize the order of all cases and methods, with an untimed complete call preceding each measurement. We take the median complete-call wall time for each pair, then report the median and interquartile range across ten pairs. Imports, input generation, and output serialization lie outside the timing boundary.

Independent 8×8 checks agree with a matrix-square-root BW formula to 1.2×10^{-14} and with a HiGHS transport solve to 4.5×10^{-16}. For all 140 pairs, the computed values satisfy $D^2 \le V^2 \le U^2$ within numerical precision and the positive support size is at most $2n - 1$. Table 5 reports the relative gap components and computation times for every setting.

Figure 6 isolates the two sources of computational saving and distance increase. BW-edges evaluates exact BW costs on the Hilbert coupling, giving the intermediate cost V^2. As n increases from 32 to 512 at $d = 3$, the mean relative squared excess $(U^2 - D^2)/D^2$ increases from 0.956 to 2.790. Coupling contributes 71%–89% of the excess across dense covariance settings. At $n = 512$, evaluating BW costs only on Hilbert matches gives an 81.1-fold speedup over MW, followed by a further factor of 1.45 from Cholesky pricing.

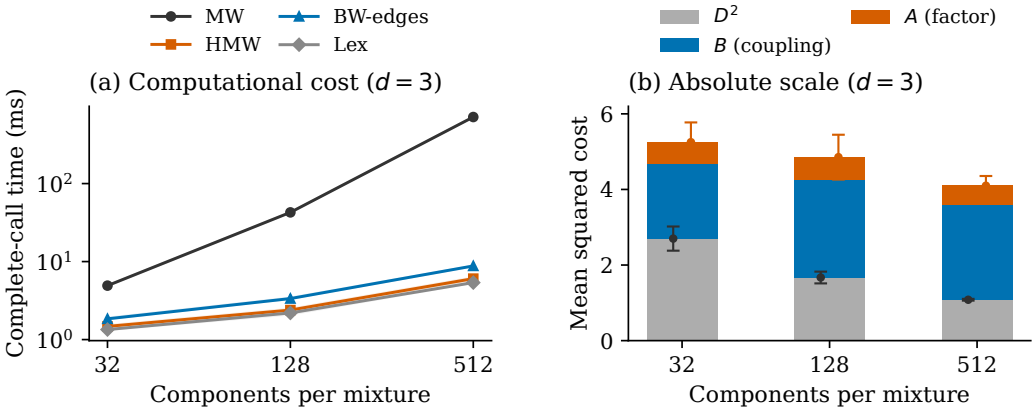


Figure 6: Computation and absolute cost at $d = 3$. Left: median complete-call time and interquartile range across ten mixture pairs. Right: mean D^2, coupling excess B, and factor excess A, whose stacked total is U^2. Error bars show one standard deviation of D^2 and U^2. Each pair's runtime is the median of three calls.

Table 5: Distance-gap decomposition and complete-call time across all random-mixture settings. Gap entries are means ± one standard deviation over ten pairs; HMW's relative squared excess is $A/D^2 + B/D^2$. Times are medians of pair-specific three-call medians in milliseconds.

n	d	A/D^2	B/D^2	MW (ms)	HMW (ms)
Dense covariances					
32	3	0.217 ± 0.042	0.739 ± 0.170	4.92	1.48
128	3	0.362 ± 0.051	1.550 ± 0.256	42.73	2.39
512	3	0.475 ± 0.034	2.315 ± 0.200	713.20	6.05
128	2	0.271 ± 0.101	2.183 ± 0.242	23.03	2.15
128	5	0.320 ± 0.018	1.040 ± 0.140	84.86	3.05
128	10	0.257 ± 0.010	0.812 ± 0.087	213.97	5.45
128	20	0.214 ± 0.005	0.526 ± 0.037	758.60	16.76
Diagonal covariances					
32	3	0.000 ± 0.000	0.864 ± 0.241	3.58	1.47
128	3	0.000 ± 0.000	2.038 ± 0.324	22.99	2.41
512	3	0.000 ± 0.000	3.437 ± 0.191	415.10	5.84
128	2	0.000 ± 0.000	2.792 ± 0.379	20.43	2.13
128	5	0.000 ± 0.000	1.436 ± 0.200	29.73	2.90
128	10	0.000 ± 0.000	1.106 ± 0.120	58.58	4.93
128	20	0.000 ± 0.000	0.682 ± 0.044	190.12	14.93

B.3 Image-mixture retrieval

We evaluate the joint position–intensity encoding used for repeated comparisons in Section 5.3. R@1 is the fraction of queries whose nearest gallery image has the same digit label, and P@5 is mean label agreement among the five nearest images. We rank by squared distance and break ties by gallery position.

B.3.1 Joint position–intensity representations

Each digit is an 8×8 image with row and column indices $r, c \in \{0, \dots, 7\}$ and intensity $v_{rc} \in [0, 16]$. The selected representation fits one Gaussian to each nonoverlapping 2×2 block C_j. For pixel observations $\boldsymbol{t}_{rc} = (c/7, r/7, sv_{rc}/16)^\mathsf{T} \in \mathbb{R}^3$, it uses weight $1/16$ and

$$\boldsymbol{\mu}_j = \tfrac{1}{4} \sum_{(r,c) \in C_j} \boldsymbol{t}_{rc}, \qquad \boldsymbol{\Sigma}_j = \tfrac{1}{4} \sum_{(r,c) \in C_j} (\boldsymbol{t}_{rc} - \boldsymbol{\mu}_j)(\boldsymbol{t}_{rc} - \boldsymbol{\mu}_j)^\mathsf{T} + 0.0025 \operatorname{diag}(1, 1, s^2).$$

The covariance records local spatial variation and position–intensity dependence. Table 1 uses $s = 2$ for every method.

Encoding selection and held-out evaluation. Before final evaluation, HMW and each DSMW budget select an encoding independently from 19 candidates. These comprise a mass-weighted 16-block spatial encoding; 64 fixed pixel Gaussians weighted by $v + \epsilon$ for $\epsilon \in \{0.01, 0.25, 1\}$; and three equal-weight families at $s \in \{0.25, 0.5, 1, 2, 4\}$. The latter families are 64 marked pixels in $\mathbb{R}^3$, the 16 joint Gaussians above, and 16 six-dimensional patch descriptors consisting of block centers and four scaled intensities. Marked-pixel means equal $\boldsymbol{t}_{rc}$; descriptor intensities follow row-major order. The spatial candidate moment-matches each block using pixel mass proportional to $v + 0.01$ and assigns the block's total mass as its mixture weight. The spatial encoding uses covariance ridge $0.0025 I_2$; marked pixels and descriptors have common covariance $0.0025 \operatorname{diag}(1, 1, s^2, \dots, s^2)$.

The original UCI partition of 1797 images supplies 100 development gallery images and 100 queries, selected by seed 2026092800 (Alpaydin & Kaynak, 1998). For each method, the two highest-R@1 encodings advance to validation, with ties broken by P@5, smaller component count, and declaration order. The original 3823-image partition supplies disjoint validation and test sets. Validation uses 200 gallery images and 100 queries, selected by seed 2026092810, and applies the same selection rule. All selected encodings are joint Gaussians, with $s = 2$ for HMW and $s = 0.25$ for each DSMW

budget. Labels determine the balanced splits and selection scores but do not enter the Gaussian parameters or distances.

From images absent from validation, seed 2026092820 selects 50 gallery and 30 query images per digit. The first ten gallery images per digit form the nested gallery of 100. Queries form three balanced cohorts of 100, with the same gallery of 500. Gallery and query images are disjoint, and exact pixel comparisons find no duplicates within or across the original partitions. DSMW direction pairs are nested prefixes of 256 normalized Gaussian draws, generated with seed $2026092830 + d$ in dimension d.

Table 6 summarizes retrieval on the larger gallery, using the independently selected encodings for HMW and DSMW. MW and Lex use the same encoding as HMW; Lex replaces Hilbert ordering by lexicographic ordering of the Cholesky features. The pixel reference compares normalized intensities $v + 0.01$. HMW gives 94.0% R@1, compared with 93.7%–95.3% for DSMW and 97.0% for Lex and the pixel reference. Using 10,000 paired query resamples within digit and seed 2026092860, the 95% intervals for HMW minus DSMW at $L = 4, 16, 64, 256$ are $[-2.7, 3.3]$, $[-3.7, 2.7]$, $[-4.3, 1.3]$, and $[-4.0, 2.0]$ percentage points, conditional on the gallery and selected encodings. These results associate useful retrieval with the joint encoding and show similar HMW and DSMW accuracy at their selected scales.

Table 6: Held-out retrieval from 500 gallery images. Entries are percentages, shown as mean $\pm$ sample standard deviation over three balanced query cohorts. HMW and DSMW use independently selected encodings; MW and Lex use HMW's encoding. The pixel reference uses normalized intensities.

Method	Intensity scale s	R@1	P@5
HMW	2	94.0 ± 1.7	86.3 ± 1.4
MW	2	90.0 ± 3.0	79.7 ± 1.2
Lex	2	97.0 ± 0.0	92.7 ± 1.6
DSMW, $L = 4$	0.25	93.7 ± 2.3	85.9 ± 0.2
DSMW, $L = 16$	0.25	94.3 ± 2.5	88.3 ± 0.6
DSMW, $L = 64$	0.25	95.3 ± 1.2	89.1 ± 0.3
DSMW, $L = 256$	0.25	95.0 ± 1.7	89.5 ± 0.1
Pixel reference	–	97.0 ± 1.0	92.4 ± 1.8

Prepared comparisons. HMW and DSMW use the equal-weight vectors described in Section 5.3. We evaluate their squared distances as $\|z\|^2 + \|w\|^2 - 2z^\top w$, caching the squared norms and obtaining all query–gallery inner products by a matrix–vector product. Single-image preparation includes the vector and its norm; packing the gallery matrix is a separate one-time operation. MW computes the complete BW matrix and solves unregularized transport for each mixture pair. MSW uses the first 16 spatial directions of the same saved sequence as DSMW. It caches projected means and standard deviations, scaled by $1/\sqrt{16}$, and their component-wise squared norms. Squared Euclidean distances between these component vectors give the direction-averaged Gaussian cost matrix, followed by one transport solve with the same POT settings as MW.

The cached HMW and DSMW calculations agree with the saved distances to numerical precision. An explicit projection-by-projection calculation and an equal-weight linear-assignment solver verify the MSW component costs and transport values on 30 pairs, with maximum absolute differences below 8×10^{-16}. All timings use double precision on one Intel Xeon Platinum 8280 core and one numerical-library thread, with Python 3.10, NumPy 1.24.4, and POT 0.9.5. The first query of each digit and cohort gives 30 timing cases per gallery size. Three rounds randomize case and method order after untimed warm-ups; preparation is timed on the 100-image gallery. Table 7 gives interquartile ranges for the same run summarized in Table 1. Image encoding, gallery packing, loading, serialization, and ranking are outside the distance-scan boundary.

Table 7: Variability of the prepared-comparison timings in Table 1. All methods use the same scale-2 encoding. Each cell gives the median above the interquartile range in milliseconds, over image- or query-specific three-round medians.

Method	Preparation	Scan 100	Scan 500
HMW	0.441 [0.436, 0.447]	0.011 [0.011, 0.012]	0.036 [0.034, 0.038]
MW	0.390 [0.382, 0.399]	112.528 [112.064, 113.220]	562.600 [560.330, 566.258]
MSW, $L = 16$	0.482 [0.466, 0.533]	31.811 [31.678, 31.897]	159.119 [158.583, 159.705]
DSMW, $L = 4$	0.483 [0.479, 0.488]	0.010 [0.010, 0.010]	0.018 [0.018, 0.019]
DSMW, $L = 16$	0.603 [0.598, 0.607]	0.014 [0.013, 0.015]	0.068 [0.066, 0.071]
DSMW, $L = 64$	0.715 [0.713, 0.718]	0.040 [0.038, 0.043]	0.246 [0.234, 0.264]
DSMW, $L = 256$	0.949 [0.945, 0.953]	0.185 [0.176, 0.204]	0.799 [0.784, 0.821]

B.4 Four-component construction

For Figure 2(b), the equal-weight mixtures are defined by

$$\mathcal{P} = \tfrac{1}{4}\sum_{i=1}^{4} \delta_{\mathcal{N}(\mu_i, I_3)}, \qquad \mathcal{Q}_t = \tfrac{1}{4}\sum_{j=1}^{4} \delta_{\mathcal{N}(\nu_j, [I_3 + \epsilon S(t)]^2)},$$

$$S(t) = \sqrt{1-t}\, S_r + c\sqrt{t}\, S_s, \qquad S_r = I_3/\sqrt{3}, \quad S_s = \mathrm{diag}(1, -1, 0)/\sqrt{2}.$$

Here $t \in [0, 1]$, $\epsilon = 0.04$, and $c = 1.25$; the covariance roots remain diagonal and SPD. The upper snapshots in Figure 2(b) show centered 3D ellipsoids at $t = 0, 0.5, 1$, using the display root $I_3 + 10\epsilon S(t)$. The ellipsoids are orthographic views of the transformed unit sphere, drawn with the same viewpoint and coordinate scale. Only the changes in standard deviations are magnified tenfold. The numbers below them give the unscaled covariance BW distance $\epsilon\|S(t)\|_F$, approximately 0.0400, 0.0453, and 0.0500. Mean locations and pairings are omitted from these shape views, whereas the distance curves use $\epsilon = 0.04$ and the original distinct mean supports. The exact mixture distance along this path is

$$\mathrm{MW}_{\mathrm{BW}}^2(\mathcal{P}, \mathcal{Q}_t) = C_\mu + \epsilon^2[(1-t) + c^2 t],$$

where C_μ is the optimal transport cost between the fixed mean sets.

The means are fixed in the first two coordinates, with zero third coordinate. The following displayed coordinates are rounded for readability; the released construction retains full precision:

$$\frac{1}{0.05}(\mu_i)_{i=1}^4 = \begin{pmatrix} 0.465044 & 0.419336 & 0 \\ -0.069712 & -0.044205 & 0 \\ -0.875084 & 0.114762 & 0 \\ 0.479752 & -0.489892 & 0 \end{pmatrix},$$

$$\frac{1}{0.05}(\nu_j)_{j=1}^4 = \begin{pmatrix} 0.513649 & 0.623127 & 0 \\ 0.182580 & 0.459004 & 0 \\ -0.823758 & -0.247748 & 0 \\ 0.127529 & -0.834383 & 0 \end{pmatrix}.$$

Both mean sets are centered at full precision. The identity permutation minimizes the equal-weight mean assignment cost, as verified by enumerating its $4!$ possibilities. The recorded Hilbert source order is $(2, 3, 1, 4)$ and the target order is $(4, 3, 1, 2)$, using one-based indices; they induce the same suboptimal mean coupling throughout the path. The covariance cost is constant across component pairs at each t, so both optimal MW and the fixed-order cost split into their respective mean costs plus the common covariance term.

Population sliced values use a deterministic Fibonacci-sphere rule with $80{,}000$ directions at 161 path locations. The second DSMW direction can be integrated analytically here. The covariance

contribution is a common scalar shift within each projected mixture; equal mixture barycenters make its cross term with projected mean transport vanish. Consequently DSMW squared equals one half of the projected mean W_2^2 average plus one half of the projected covariance term. MSW uses the exact mean term C_μ/d plus that covariance term. Numerical integration approximates these population quantities. The displayed MSW and DSMW distances are multiplied by $\sqrt{3}$ and $\sqrt{6}$, respectively, following the same dimension-dependent radial calibration as Figure 2(a).

B.5 Two-dimensional shape visualization

For Figure 2(a), both mixtures have weights $1/4$ and the same means

$$\mu_1 = \nu_1 = (-2.5, -2), \quad \mu_2 = \nu_2 = (-2, 2), \quad \mu_3 = \nu_3 = (2, -2), \quad \mu_4 = \nu_4 = (2.5, 2).$$

The source covariance is I_2, and every target component has covariance

$$\left[I_2 + 0.4 \left\{ \sqrt{1-t}\, I_2/\sqrt{2} + 1.25\sqrt{t}\, \operatorname{diag}(1,-1)/\sqrt{2} \right\} \right]^2 .$$

We evaluate 161 equally spaced values of $t \in [0, 1]$. The smallest principal-root eigenvalue along this path is greater than 0.646. The contours are the images of the unit circle under each component's covariance root, translated to its mean, with identical coordinate scales in both snapshots. They show the actual perturbation amplitude.

We use Cholesky features with unmodified HCP `base.cpp` at commit `930ecc98` (Li et al., 2024). Both returned orders are $(1, 2, 4, 3)$, giving mass $1/4$ to each corresponding atom pair. Canonical enumeration leaves the result unchanged in this example and in Figure 2(b). For an independent reference, we form the complete Gaussian BW cost matrix and solve the equal-weight assignment problem, checking all $4!$ permutations. Both MW and HMW agree with $0.4\sqrt{1 - t + 1.25^2 t}$ to numerical precision.

The population MSW covariance term is integrated over $8{,}192$ equally spaced directions on $\mathbb{S}^1$. Identical mean supports and the common covariance within each mixture give zero optimal mean cost and $D^2_{\mathrm{DSMW}} = D^2_{\mathrm{MSW}}/2$. We also check this identity by explicit second projections along 64 equally spaced directions and one-dimensional sorted transport at $t = 0, 0.25, 0.5, 0.75, 1$. Doubling the circular quadrature from $4{,}096$ to $8{,}192$ nodes changes the calibrated distance by less than 2×10^{-16}. The plotted factors $\sqrt{2}$ for MSW and 2 for DSMW follow the radial calibration in (1). All four curves start at 0.400. At $t = 1$, MW and HMW equal 0.500, while both calibrated sliced distances are approximately 0.352. This construction isolates covariance pricing with a verified diagonal component coupling.

B.6 Extended runtime scaling

Paired mixtures. Figure 3 uses source and target mixtures with $K = M = n$ components in $\mathbb{R}^d$. Source means are uniform on $[-2, 2]^d$, and weights follow $\mathrm{Dirichlet}(2, \ldots, 2)$. For each source covariance $\Sigma_i = Q_i \operatorname{diag}(\lambda_i) Q_i^{\mathsf{T}}$, we obtain Q_i by QR factorization of a standard Gaussian matrix, correcting column signs using the diagonal of the triangular factor. We sample log-eigenvalues independently from $[-\log(20)/2, \log(20)/2]$ and subtract their mean before exponentiation. Thus $\det \Sigma_i = 1$ and the source condition number is at most 20.

With perturbation level $\epsilon = 0.35$, target covariances are

$$\Lambda_i = R_i Q_i \operatorname{diag}(\lambda_i \odot e^{\delta_i}) Q_i^{\mathsf{T}} R_i^{\mathsf{T}}, \qquad \delta_i = 0.35\epsilon(\xi_i - \bar{\xi}_i \mathbf{1}), \qquad R_i = \exp\big((\pi\epsilon/6) S_i\big),$$

where $\xi_i \sim \mathcal{N}(0, I_d)$, $\bar{\xi}_i$ is its coordinate mean, and S_i is a random skew-symmetric matrix normalized to unit Frobenius norm. These perturbations preserve determinant one without imposing an additional condition-number cap on the target. Initial mean displacements are $0.30\epsilon\eta_i$, with $\eta_i \sim \mathcal{N}(0, I_d)$. We rescale all displacements by a common factor to match the average squared covariance perturbation,

$$\frac{1}{n} \sum_i \|\mu_i - \nu_i\|^2 = \frac{1}{n} \sum_i d^2_{\mathrm{BW}}\big(\mathcal{N}(0, \Sigma_i), \mathcal{N}(0, \Lambda_i)\big).$$

Target weights are proportional to $\max\{a_i \exp(0.45\epsilon\zeta_i), 10^{-5}\}$, with independent $\zeta_i \sim \mathcal{N}(0, 1)$. We randomly permute the target components and pass the same raw weights, means, and covariances to every method. The BW computations used to generate these inputs occur before timing and supply no precomputed factors to the methods.

Projection sampling. MSW and DSMW use the released implementation at commit `5e6f326` (Piening & Beinert, 2025) with a budget of 16 and no adaptive stopping. MSW normalizes vectors with independent Uniform$(0, 1)$ entries in $\mathbb{R}^d$. DSMW independently generates one such vector in $\mathbb{R}^d$ and one in $\mathbb{R}^2$ for each of its 16 direction pairs. These directions lie in the positive orthant and are not uniform on the full sphere. The projection seed is reset to 20260817 before each call, giving the same direction sequence at a fixed dimension. Seed resetting is outside the timed interval, while direction generation and normalization are inside it. This sampling convention applies to the runtime benchmark; Figure 2 uses the spherical integration rules described above.

Timed computation. Each measurement starts from the raw GMM arrays and ends with the returned distance. HMW includes input copying and validation, weight normalization, both Cholesky factorizations, full feature construction, official HCP ordering, weighted NW matching, and sparse cost evaluation. This implementation omits duplicate merging and canonical enumeration. MSW and DSMW include tensor conversion, direction sampling, projection, transport computation, and final aggregation. MSW averages the projected Gaussian cost matrices and solves exact EMD, while DSMW computes 16 one-dimensional transports. Exact MW constructs the full Gaussian BW cost matrix from the input covariances and solves exact EMD. All methods include the final square root.

Runs use double precision on a Dell PowerEdge T640 with Python 3.9.25, NumPy 2.0.2, and CPU PyTorch 2.8.0. We pin execution to one logical CPU and use one numerical-library thread. Method order rotates across repetitions, and garbage collection is disabled during timing. Data generation, imports, compilation, warm-up, input-mutation checks, and result storage are excluded. Warm-up uses $n = 32$, separately at each tested dimension.

Sizes and repetitions. Panel (a) fixes $d = 3$ and uses $n = 100, 1000, 10^4, 10^5$, with respectively $5, 3, 1, 1$ independently generated cases. HMW and DSMW use $31, 9, 5, 3$ repetitions per case, respectively. MSW and exact MW use seven repetitions at $n = 100$ and three at $n = 1000$. They are omitted at $n \geq 10^4$, where a single dense float64 cost matrix exceeds the preset 128 MiB threshold. This threshold controls admission of dense matrices rather than measuring peak process memory. Panel (b) fixes $n = 1000$ and uses $d = 5, 10, 15, 20, 25, 30$, with three cases per dimension. HMW and DSMW use seven repetitions per case, MSW three, and exact MW one. Exact MW is evaluated only at $d = 5$ under the preset dimension limit. Each plotted value is the median across case-level median times; no missing results are extrapolated.

C Application protocols and additional results

C.1 Density-reduction construction and evaluation

To approximate $f = \sum_{i=1}^{N} a_i \varphi_i$ by $g = \sum_{j=1}^{M} b_j \psi_j$, with Gaussian densities φ_i, ψ_j and $M \ll N$, we partition the Hilbert-ordered Cholesky features into M consecutive nonempty groups C_j of sizes differing by at most one. The group mass is $b_j = \sum_{i \in C_j} a_i$, and $w_{ij} = a_i / b_j$ is the source weight normalized within group j. Moment matching initializes the representative $F_j^{(0)} = \mathcal{N}(\nu_j^{(0)}, \Lambda_j^{(0)})$ by

$$\begin{aligned} \nu_j^{(0)} &= \sum_{i \in C_j} w_{ij} \mu_i, \\ \Lambda_j^{(0)} &= \sum_{i \in C_j} w_{ij} \left[\Sigma_i + (\mu_i - \nu_j^{(0)})(\mu_i - \nu_j^{(0)})^{\mathsf{T}} \right]. \end{aligned} \tag{9}$$

The scatter term accounts for variation of component means within each group. This initialization preserves the source mixture's mean and raw second moment and yields SPD representatives (Lemma 12).

For fixed groups and weights, the refinement objective is

$$J = \sum_j b_j J_j, \qquad J_j = \sum_{i \in C_j} w_{ij} \|\varphi_i - \psi_j\|_{L^2}^2. \tag{10}$$

Analytic Gaussian overlaps give both this objective and its gradients. Each term compares a source Gaussian density with its group's representative density. The CTD surrogate inequality (Zhang et al.,

2024) gives $\|f-g\|_{L^2}^2 \le J$. Proposition 16 combines this inequality with Gaussian density regularity to control the final error by the initial grouping distortion. For HMW+ISE5, groups and weights stay fixed. We parameterize each covariance by a lower-triangular factor with log diagonal and minimize the total surrogate J using analytic gradients and L-BFGS-B. The iteration limit is five, relative objective tolerance 10^{-12}, gradient tolerance 10^{-8}, and line-search limit twenty. Among evaluated candidates, including initialization, we retain the one with the smallest J and finite objective and gradient. The Cholesky parameterization ensures SPD covariances. Including the initial candidate guarantees that the returned surrogate value does not increase.

CTD-ISE uses the released hard-assignment implementation with zero regularization, tolerance 10^{-8}, and a limit of three assignment/support cycles. Each cycle reassigns source components, updates representative weights, and optimizes representative parameters. Its representative updates use the implementation's BFGS density barycenters; the inner gradient tolerance is 10^{-7}.

Analytic density evaluation. For Gaussian densities, the overlap is

$$\kappa(G,F) = \int \varphi_G(x)\varphi_F(x)\,dx = \mathcal{N}(\mu_G;\mu_F,\Sigma_G+\Sigma_F).$$

Consequently,

$$\|f-g\|_{L^2}^2 = \sum_{ii'} a_i a_{i'} \kappa(G_i,G_{i'}) + \sum_{jj'} b_j b_{j'} \kappa(F_j,F_{j'}) - 2\sum_{ij} a_i b_j \kappa(G_i,F_j).$$

We normalize this error as

$$\mathrm{nISE}(f,g) = \|f-g\|_{L^2}^2 / \|f\|_{L^2}^2.$$

The same overlap formula gives the denominator. The source, initial representatives, and both reduced mixtures are retained for reproducing the evaluation.

Additional density configuration. A $200 \to 20$ comparison uses 40 equally weighted atoms around each of $(-6,-5)$, $(-5,5)$, $(0,0)$, $(5,5)$, and $(6,-5)$. Means are uniform in radius-2.5 disks and covariances are $ZZ^{\mathsf{T}}/2 + 0.25 I_2$, with independent standard Gaussian entries in Z and seed 2026092503. Both methods start from ten atoms per Hilbert group and the same moment-matched representatives. HMW's five joint iterations reduce nISE from 0.2283 to 0.1770; CTD's three cycles give 0.0148 and 19 nonempty representatives, reaching the cycle limit before convergence. Thus the relative density fit depends on the mixture configuration.

Four-mode density example. Figure 4(a) uses 128 equally weighted atoms, with 32 means sampled around each center $(-3,-3)$, $(-3,3)$, $(3,3)$, and $(3,-3)$ using independent coordinate noise $\mathcal{N}(0,0.5^2)$ and seed 901. For the four groups, orientations are 0, $\pi/4$, $\pi/2$, and $3\pi/4$, respectively, with independent angular perturbations $\mathcal{N}(0,0.03^2)$. Each covariance is a rotated diagonal matrix with eigenvalues $\exp(0.1Z_1)$ and $\exp(0.1Z_2)/25$, where Z_1, Z_2 are independent standard normals. Canonical full Cholesky features give 8 consecutive groups of sixteen atoms and a common moment-matched initialization. HMW uses five joint iterations; CTD uses three assignment/support cycles under the tolerances above, returning 8 nonempty representatives. With its cycle limit increased to 1000, CTD declares convergence at cycle 4 and returns the same model. Analytic nISE is 0.02559 for HMW and 0.04413 for CTD, compared with 0.03261 at initialization.

The panel evaluates the source and both reduced models on a 500×500 grid over $[-7,7]^2$. Densities share one linear color scale, and each model's contours use common levels. Independent grid integration agrees with analytic nISE to within 0.02%; the ordering also matches the vendored C++ Hilbert recursion on this input. The lower-right CTD peak is more concentrated and taller than the source peak. HMW is closer to the source peak height in that region and has lower overall nISE for this example.

We inspected 264 two-dimensional configurations spanning cluster geometry, weight imbalance, overlap, and covariance anisotropy. Of these, 38 have lower HMW nISE under the stated finite budgets. The displayed case was selected for its visible shape difference and lower HMW nISE from a 48-setting anisotropy grid, of which 11 settings favor HMW.

C.2 One-dimensional scalar-order example

Figure 4(b) shows a selected exploratory $100 \to 10$ case with common source standard deviation 0.65. Ten atom means are sampled independently with standard deviation 0.55 around each center in $(-7.4, -5.9, -4.5, -2.8, -1.1, 0.7, 2.1, 3.8, 5.5, 7.3)$, using NumPy's default generator with seed 302. The corresponding cluster masses are $(0.055, 0.095, 0.125, 0.070, 0.150, 0.080, 0.105, 0.090, 0.140, 0.090)$; within each cluster, weights follow a symmetric Dirichlet distribution with parameter 2. Since all source variances are equal, this scalar-order example sorts atoms by their means and splits the ordered list into ten equal-sized groups for moment initialization. Overlapping generative clusters need not coincide with these consecutive groups. Both methods receive the same initialization. HMW uses five fixed-group joint ISE iterations; released CTD-ISE uses zero regularization, tolerance 10^{-8}, and at most 1000 assignment/support cycles. Analytic nISE is 1.9535×10^{-4} for HMW and 5.2566×10^{-4} for CTD.

The exploratory grid combines source standard deviations $\{0.35, 0.65, 1.0\}$ with mean spreads $\{0.25, 0.55\}$, and standard deviations $\{0.10, 0.20, 0.35\}$ with spreads $\{0.8, 1.2\}$. Each combination uses seeds $\{301, 302, 303\}$. Of these 36 settings, 10 favor HMW by nISE. The displayed case has the smallest HMW-to-CTD nISE ratio in this grid. The inset centers a width-1.1 interval at the plotted grid point where CTD's absolute density error exceeds HMW's by the largest amount.

C.3 3D Gaussian Splatting scene compression details

From ordered atoms to spatial compression. A 3DGS primitive additionally carries opacity and spherical-harmonic (SH) appearance. We compare a spatial Hilbert extension against the official GHAP compressor (Wang et al., 2025). For retention fraction $\rho \in (0, 1)$ and $M = \lfloor \rho N \rfloor \geq 1$, recursive HCP orders the centers in $\mathbb{R}^3$. Consecutive groups initialize M representatives; neighboring groups form blocks of at most 100 representatives. Within each block, weighted Lloyd iterations reassign source primitive i to representative j using GHAP's geometry cost

$$c_{\mathrm{GS}}(i, j) = \|\mu_i - \nu_j\|^2 + \|\Sigma_i - \Lambda_j\|_F^2.$$

They then update representative means and covariances by opacity-weighted averaging,

$$(\nu_j, \Lambda_j) = \sum_{i \in C_j} \omega_{ij}(\mu_i, \Sigma_i), \qquad \omega_{ij} = \alpha_i \Big/ \sum_{k \in C_j} \alpha_k.$$

Here α_i denotes source opacity. The covariance update omits the scatter term in (9). Each representative copies SH and opacity from the lowest-cost source in its final cluster. Equal-sized blocks run in GPU batches. The comparison includes each method's partitioning, initialization, empty-cluster handling, and execution. This spatial extension uses the ordering mechanism with a scene-specific cost, and its evaluation measures rendering quality.

Controlled comparison. We evaluate Train and Truck from Tanks and Temples on all 38 and 32 held-out views, at 10%, 1%, and 0.1% retention. Each method pair uses the same input, resolution, and RTX 3090. GHAP keeps its official implementation and KD-tree depth of ten. Both outputs receive the same 5000-iteration fine-tuning loop (4999 optimizer updates): fresh Adam, seed 0, common learning-rate interval 30001–35000, trainable geometry and appearance, and no densification or further compression. Truck starts from a 15000-step source checkpoint but uses this same fine-tuning interval in both arms. We report offline PSNR, SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018) over the full test split, both immediately after compression and after fine-tuning.

Implementation and timing conventions. Hilbert repairs empty clusters by moving the largest weighted-residual source from a non-singleton cluster. Blocks stop at unchanged labels, 100 updates, or objective change below $10^{-12} \max\{J^{(0)}_{\mathrm{block}}, 10^{-12}\}$. Batch sizes are 16 for $10\%/1\%$ retention and 2 for 0.1%. Compression uses seed 42, one warm-up, and three measured runs, including parameter preparation and output construction but excluding file I/O, rendering, and evaluation. Fine-tuning time includes process startup, loading, internal evaluation, and saving. Retention settings are measured in separate paired batches; reported stage-time sums combine separately measured compression and fine-tuning times. GHAP deletes empty clusters, explaining the slight count differences in Table 8.

Table 2 and Figure 5 in Section 6.2 report rendering quality; Table 8 below gives component counts and stage costs.

Table 8: **Component counts and stage costs.** M is the actual output count. Compression time t_c is a three-run median; fine-tuning time t_f is one process measurement. Bold marks the lower time in each pair.

Retained	Method	M	t_c (s)↓	t_f (s)↓
Train: $N = 1,026,508$				
10%	GHAP	102599	15.56	**96.89**
	Hilbert	102650	**6.01**	102.79
1%	GHAP	10261	18.75	**114.89**
	Hilbert	10265	**10.91**	115.50
0.1%	GHAP	1026	12.45	**111.01**
	Hilbert	1026	**12.03**	121.36
Truck: $N = 2,068,617$				
10%	GHAP	206642	24.77	**106.23**
	Hilbert	206861	**12.53**	116.27
1%	GHAP	20679	32.95	**110.91**
	Hilbert	20686	**21.44**	115.97
0.1%	GHAP	2068	25.29	**109.97**
	Hilbert	2068	**24.73**	112.65