Beyond Contraction: Geometry-Faithful Supervised Dimensionality Reduction for Data Visualization

Xu Zhu **Yilin Ye** **Shixia Liu***
BNRist, Tsinghua University

Abstract

Supervised dimensionality reduction (DR) is widely used in visualization to reveal task-specific structure in high-dimensional data. However, in regression settings with continuous supervision, existing methods often improve signal readability at the cost of severe neighborhood distortion, limiting the reliability of the resulting visualization. To understand this trade-off, we analyze the induced readout mapping from the low-dimensional embedding to the supervision signal, using two finite-sample proxies for its bi-Lipschitz behavior: contraction (co-Lipschitz) and smoothness (Lipschitz). We show that enforcing contraction can impose a sharp geometric distortion floor in many existing methods, whereas regularizing smoothness alone is sufficient to improve signal readability. In particular, near a local geometry optimum, regularizing smoothness yields first-order regularity gains with only second-order geometry loss. Guided by this insight, we propose ISDR (***I***nterpretable ***S***upervised ***DR***), a nonparametric method that preserves geometric structure while enforcing readout smoothness. Experiments on HPDv3 and Tabula Muris show that ISDR improves the geometry–signal Pareto frontier, reducing local signal variation while retaining faithful neighborhoods compared with representative supervised DR baselines. Our code is available at `https://github.com/Circular-Aflame/ISDR`.

1 Introduction

Dimensionality reduction (DR) methods are widely used in data visualization to reveal structure in high-dimensional data [1, 2, 3, 4]. In many applications, each data point is associated with a continuous supervision signal, such as gene expression levels in single-cell transcriptomics [5] and model confidence scores in LLM-as-a-Judge evaluation [6]. Commonly used unsupervised DR methods, such as t-SNE [7] and UMAP [8], do not explicitly consider these signals to guide the projection. To address this limitation, supervised DR methods explicitly integrate the supervision signal into the projection [9, 10]. However, these methods face a recurring trade-off: stronger supervision can improve signal readability, but often distorts the geometric structure needed for interpretation [11]. The first two rows of Figure 1 illustrate this trade-off: increasing the supervision weight λ makes signal patterns more distinct, but flattens semantically meaningful local clusters, reducing the layout's fidelity to the high-dimensional data geometry.

Understanding how supervision affects DR layouts is essential for explaining this trade-off and designing more robust methods. Existing supervised DR methods differ in how and when they incorporate the supervision signal. For example, supervised UMAP [8] combines the high-dimensional data graph with the one-dimensional supervision signal graph before projection, while FDA [12] and recent methods such as AKRMap [13] apply supervision for classification or regression after projection. However, a unified framework for interpreting these methods is still lacking, making it difficult to identify their shared strengths and limitations.

*Corresponding author: `shixia@tsinghua.edu.cn`.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

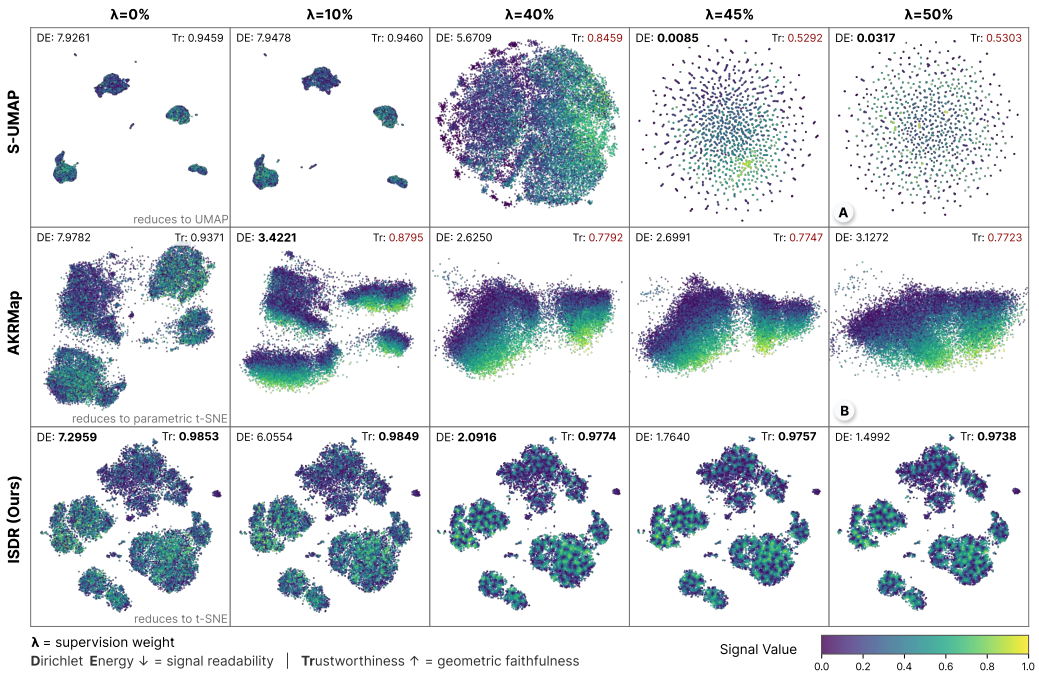

Figure 1: Comparison of ISDR with two representative supervised DR methods on the LLM-as-a-Judge case under varying supervision weights λ. The supervision signal is the model's confidence in its judgments. Increasing λ makes the signal pattern more distinct but distorts the geometric layout, illustrating the trade-off between geometric faithfulness and signal readability (Section 6.2). The columns sample λ more densely near the S-UMAP collapse thresholds.

To provide such a unified view, we formulate the two objectives of supervised DR as *geometric faithfulness* and *signal readability*, and analyze their trade-off using distortion theory [14, 15, 16]. Our analysis centers on the bi-Lipschitz conditions of a readout function that maps the projected two-dimensional layout to the supervision signal. These conditions distinguish two requirements: *contraction*, which requires points with similar signal values to be spatially close, and *smoothness*, which requires nearby projected points to have similar signal values. Using finite-sample proxies, we show that contraction is distinct from smoothness and can impose a sharp lower bound on geometric distortion. Empirically, several representative supervised DR methods increasingly favor contraction as supervision strengthens. By contrast, near a local optimum of the geometry objective, directly regularizing readout smoothness on the final layout can control local signal variation while incurring only a second-order increase in the geometry objective.

Motivated by this analysis, we propose ISDR (***I***nterpretable ***S***upervised ***DR***), a nonparametric method that directly optimizes the final low-dimensional layout by combining geometry preservation with readout regularity. We evaluate ISDR on an LLM-as-a-Judge benchmark and a single-cell atlas benchmark. The Pareto curves on both benchmarks show the same pattern: ISDR reduces local signal variation while preserving neighborhoods, whereas several baselines achieve such reductions under stronger supervision only at the cost of substantial geometric distortion.

The contributions of our work are threefold:

- We formulate supervised DR for data visualization as a bi-objective problem that balances geometric faithfulness and signal readability via the readout induced by DR embedding.
- We analyze this trade-off through finite-sample proxies for the two bi-Lipschitz requirements (*contraction* and *smoothness*). We also show that smoothness regularization controls local signal variation, while contraction can impose a sharp lower bound on geometric distortion.
- We develop ISDR, a nonparametric method for optimizing the final layout based on this analysis, and demonstrate an improved empirical trade-off between geometric faithfulness and signal readability on HPDv3 and Tabula Muris.

2 Related Work

2.1 Supervised Dimensionality Reduction

In statistical regression, supervised DR methods, such as sufficient dimension reduction [17, 18, 19, 20, 21], kernel dimension reduction [22, 23, 24], and supervised PCA [25, 26], seek low-dimensional representations for downstream prediction. We study supervised DR for data visualization, where the goal is to reflect the supervision signal in a 2D or 3D layout while preserving the geometric structure needed for interpretation [27]. We group existing methods into two categories based on how they incorporate supervision: *relation-driven* and *target-driven* methods.

Relation-driven methods use the supervision signal to modify pairwise relations by adjusting affinities or selecting positive and negative pairs. Many relation-driven methods extend manifold learning [11, 28] by incorporating supervision into pairwise distances or affinities. For example, supervised UMAP [8] intersects affinities derived from the supervision signal with those derived from the input data. RF-PHATE [29] replaces the original PHATE affinities [30] with proximities learned by a random forest using the signal. Other efforts include contrastive methods such as CEBRA [31], which use supervision to sample positive and negative pairs, and Neighborhood Components Analysis methods [32, 33], which learn soft neighborhood probabilities. By modifying pairwise relationships before or during layout optimization, these methods can bring points with similar signals closer together and separate points with dissimilar signals.

Target-driven methods incorporate supervision through an explicit objective on the low-dimensional representation. FDA [34] and its related discriminant methods [35, 12] encourage separation between classes and compactness within each class. Centroid Encoder [36] and SLCE [37] use class centers as reconstruction targets. Supervised NMF variants [38, 39] augment the factorization objective with classification or regression losses. RF-AE [9] and TP-SE [40] refine autoencoder representations using auxiliary supervision or classifiers. Perception-driven DR [41] adapts simulated annealing to maximize visual class separation, while AKRMap [13] combines a parametric t-SNE objective with a Nadaraya–Watson regression loss. These objectives can encourage points with similar supervision signals to lie closer together in the layout.

In both categories, stronger supervision can bring points with similar signal values together globally and distort geometry. ISDR instead regularizes readout smoothness on the final layout, reducing local signal variation while preserving neighborhoods.

2.2 Theoretical Foundations for Dimensionality Reduction

Metric embedding theory provides a foundation for analyzing dimensionality reduction by studying how mappings between metric spaces distort distances [42, 14, 15, 43]. Classical theories such as the Johnson–Lindenstrauss lemma [44] and Bourgain's theorem [45] use bi-Lipschitz conditions to provide worst-case distortion guarantees. Subsequent work examines average-case distortion in machine learning practice [46, 16] and distortion measures beyond the standard bi-Lipschitz setting [47]. Other studies analyze the behavior of t-SNE and UMAP directly [48, 49, 50]. While the theory of unsupervised DR provides a foundation for understanding geometric distortion, theoretical work on supervised DR is more limited and has focused mainly on classification. Hajderanj et al. [11] show that dissimilarities defined by supervision can lead to geometric distortion in manifold-based DR. Vural and Guillemot [51] relate the generalization of classifiers in supervised manifold learning to the Lipschitz regularity of the interpolation function.

In this work, we analyze the trade-off between geometric faithfulness and signal readability by applying distortion theory to the readout that maps the layout to the supervision signal.

3 Problem Formulation

We first define the DR embedding and its geometric distortion. Then we introduce the induced readout and the finite-sample proxies used to analyze the trade-off between geometric faithfulness and signal readability.

3.1 Dimensionality Reduction Setup

Given a dataset $\mathcal{D} = \{(x_i, s_i)\}_{i=1}^N$ of high-dimensional inputs $x_i \in \mathbb{R}^D$ and scalar signals $s_i \in \mathbb{R}$, a DR embedding is a mapping $f : \mathcal{X} \subseteq \mathbb{R}^D \to \mathcal{Y} \subseteq \mathbb{R}^d$, where $\mathcal{Y}$ is the low-dimensional Euclidean space. We write $y_i = f(x_i)$ and $Y = (y_1, \ldots, y_N)^\top \in \mathbb{R}^{N \times d}$ for the sample layout in $\mathcal{Y}$, typically $d = 2$ for visualization.

The pairwise expansion ratio [16], defined by $\rho_f(i, j) = \|y_i - y_j\|/d_{\mathcal{X}}(x_i, x_j)$, records how the embedding rescales the (i, j)-th distance and underlies the geometry analysis below. To avoid degenerate cases, we assume that both the sampled inputs and their layout points are pairwise distinct. We measure geometric faithfulness with the scale-invariant σ-distortion [46, 16] introduced for machine learning literature.

Definition 3.1 (σ-distortion [46, 16])**.** Let $\binom{[N]}{2}$ denote the set of unordered pairs of distinct indices in $[N] = \{1, \ldots, N\}$, and let $\bar{\rho}_f = \frac{2}{N(N-1)} \sum_{(i,j)\in\binom{[N]}{2}} \rho_f(i, j)$ be the mean pairwise expansion ratio. The σ-distortion of f is

$$\Phi_\sigma(f) = \frac{1}{\binom{N}{2}} \sum_{(i,j)\in\binom{[N]}{2}} \left(\frac{\rho_f(i, j)}{\bar{\rho}_f} - 1 \right)^2. \tag{1}$$

3.2 Induced Readout of Supervision Signal

The embedding induces another mapping from the low-dimensional layout to the supervision signal.

Definition 3.2 (Induced readout)**.** Given a dataset $\mathcal{D} = \{(x_i, s_i)\}_{i=1}^N$ and a layout $Y = (y_1, \ldots, y_N)^\top$ with $y_i = f(x_i)$, an induced readout is any scalar field $h : \mathbb{R}^d \to \mathbb{R}$ satisfying $h(y_i) = s_i$ for all i.

We operationalize signal readability through local signal variation: lower variation means that nearby layout points carry similar signal values, making scalar patterns locally coherent [52, 53, 54].

Definition 3.3 (Local signal variation)**.** For $r > 0$, let $\mathcal{N}_r(Y) = \{(i, j) : 1 \le i < j \le N, \|y_i - y_j\| \le r\}$. The local signal variation at scale r is

$$M_{\mathrm{sm}}(Y; r) = \frac{1}{|\mathcal{N}_r(Y)|} \sum_{(i,j)\in\mathcal{N}_r(Y)} |s_i - s_j|^2. \tag{2}$$

The scale-normalized proxy is $M_{\mathrm{sv}}(Y; \tau) = M_{\mathrm{sm}}(Y; \tau\ell_k(Y))$, where $\ell_k(Y) = \frac{1}{N} \sum_{i=1}^N d_k(y_i)$ and $d_k(y_i)$ is the distance from y_i to its k-th nearest neighbor in the layout.

Together with geometric distortion, local signal variation defines the geometry–signal trade-off.

3.3 Readout Analysis Proxies

Metric embedding theory often uses bi-Lipschitz embeddings as a desirable model for preserving pairwise distances [15, 43]. We first examine the finite-sample implications of bi-Lipschitz conditions for the induced readout, then define layout-level proxies. For thresholds $0 < \epsilon_{\mathrm{near}} < \epsilon_{\mathrm{far}}$ and $0 < r_{\mathrm{near}} < r_{\mathrm{far}}$, the implication form is

$$|s_i - s_j| \le \epsilon_{\mathrm{near}} \Longrightarrow \|y_i - y_j\| \le r_{\mathrm{near}}, \tag{3}$$

$$|s_i - s_j| > \epsilon_{\mathrm{far}} \Longrightarrow \|y_i - y_j\| > r_{\mathrm{far}}. \tag{4}$$

The first implication says that signal-near pairs are made layout-near, corresponding to the co-Lipschitz direction. The second says that signal-far pairs cannot remain layout-near; equivalently, nearby layout points should not have large signal differences, corresponding to the Lipschitz direction.

Taken literally, these implications impose a strong prediction property: the layout must encode the signal with both forward and inverse control. However, exact bi-Lipschitzness is not an appropriate objective for a readout to a one-dimensional signal. For example, an exact co-Lipschitz readout would force all samples with the same signal value to occupy the same layout point; strongly supervised layouts can visibly approach this collapse, as in S-UMAP (Figure 1A). We therefore define finite-sample proxies for *smoothness* and *contraction* on the realized layout. For the Lipschitz direction,

the smoothness proxy is the local signal-variation quantity $M_{\mathrm{sv}}(Y;\tau)$ from Definition 3.3. For the co-Lipschitz direction, we define the following contraction proxy. Formal Lipschitz/co-Lipschitz definitions and proxy relations are collected in Appendix A.2.

Definition 3.4 (Contraction proxy)**.** Fix a signal-difference band $0 \le \delta \le \epsilon_s$ and let

$$\mathcal{E}_{\delta,\epsilon_s} = \{(i,j) : 1 \le i < j \le N,\ \delta \le |s_i - s_j| \le \epsilon_s\}. \tag{5}$$

The contraction proxy is

$$M_{\mathrm{ctr}}(f;\delta,\epsilon_s) = \frac{1}{|\mathcal{E}_{\delta,\epsilon_s}|} \sum_{(i,j)\in\mathcal{E}_{\delta,\epsilon_s}} \frac{\rho_f(i,j)}{\bar{\rho}_f}. \tag{6}$$

$M_{\mathrm{ctr}} < 1$ means that pairs in this signal-difference band are, on average, compressed relative to the global mean expansion ratio.

We use these two proxies to analyze the geometry–signal trade-off theoretically.

4 Theory of the Geometry–Signal Trade-off

Definitions 3.1 and 3.3 give the finite-sample quantities used below: geometric distortion and local signal variation. Many supervised DR objectives introduce signal-neighbor attraction, class compactness, or predictive surrogate pressure toward contraction [28, 11] (see Appendix A.10). We show that contraction is not a readability criterion and can impose a geometric cost, then analyze a direct final-layout smoothness regularizer.

Main conclusions.

C1 (Contraction is not readability and can force distortion). The contraction proxy is not a readability proxy: without additional structural assumptions, it does not control, and is not controlled by, local signal variation. Moreover, average compression of a nontrivial signal-difference band forces a sharp lower bound on σ-distortion.

C2 (Direct smoothness route with controlled geometry loss exists). Near a strict local minimum of the geometry objective, there exists a smooth branch along which direct final-layout smoothness regularization gives first-order regularity gain with only a second-order increase in that objective.

4.1 Distortion Floor from Contraction

The contraction proxy M_{ctr} is measured on a signal-difference band, whereas the smoothness proxy M_{sv} is measured on the realized local graph. Separating constructions in Appendix A.8 show that a small value of either proxy need not imply a small value of the other. The geometric cost of contraction is more direct: if a nontrivial signal-difference band is compressed on average, the normalized expansion ratios cannot all remain close to one. Let

$$p_{\delta,\epsilon_s} := \frac{|\mathcal{E}_{\delta,\epsilon_s}|}{\binom{N}{2}}. \tag{7}$$

Proposition 4.1 (Distortion floor from the contraction proxy)**.** *For every embedding map f and every signal-difference band $0 \le \delta \le \epsilon_s$ with $0 < p_{\delta,\epsilon_s} < 1$,*

$$\Phi_\sigma(f) \ge \frac{p_{\delta,\epsilon_s}}{1 - p_{\delta,\epsilon_s}} \left(1 - M_{\mathrm{ctr}}(f;\delta,\epsilon_s)\right)^2. \tag{8}$$

The bound is sharp (Appendix A.4) and informative when the band $[\delta,\epsilon_s]$ carries nonnegligible pair mass and M_{ctr} is materially below 1. Together, these results establish C1: contraction is not itself a readability criterion, and pushing it can impose a geometric cost.

4.2 Existence of a Smoothness Route

We next study a final-layout *Nadaraya–Watson readout regularizer*, written as L_{Reg}. The regularizer uses a fixed-support leave-one-out kernel readout adapted from Nadaraya–Watson regression [55, 13, 56] (see Appendix A.9 for its relation to regression).

For each point i, let $S_i \subseteq \{1, \dots, N\} \setminus \{i\}$ be a fixed leave-one-out support of size k, taken from the current-layout k-nearest neighbors and rebuilt periodically during optimization. With the Student-t kernel $K_t(y, y') = (1 + \|y - y'\|^2)^{-1}$, the readout and regularizer are

$$\hat{s}_i(y) = \frac{\sum_{j \in S_i} K_t(y, y_j)\, s_j}{\sum_{j \in S_i} K_t(y, y_j)}, \qquad L_{\mathrm{Reg}}(Y) = \frac{1}{N} \sum_{i=1}^{N} \big(s_i - \hat{s}_i(y_i)\big)^2. \tag{9}$$

Let $\Omega_\tau(Y)$ denote the support-overlap residual defined in Appendix A.7; it is small when neighboring points use similar readout supports and weights.

Proposition 4.2 (Final-layout NW regularizer upper-bounds local signal variation)**.** *Under the kernel and score boundedness assumptions in Appendix A.6, there exist constants $c_1, c_2, c_3 > 0$ depending on the kernel/support bounds and k, but independent of N, such that for every $\tau > 0$ with $\mathcal{N}_{\tau \ell_k(Y)}(Y) \neq \varnothing$,*

$$M_{\mathrm{sv}}(Y; \tau) \le c_1 \tau^2 \ell_k(Y)^2 + c_2 \Gamma_\tau(Y) L_{\mathrm{Reg}}(Y) + c_3 S_{\max}^2 \Omega_\tau(Y). \tag{10}$$

Here $\Gamma_\tau(Y)$ is the scale-invariant local crowding factor from Appendix A.6.

Thus L_{Reg} controls the local signal variation that defines signal readability, up to explicit local-scale, crowding, and support-overlap factors. Appendix A.5 transfers the bound on M_{sv} to the reported DE metric, which averages signal variation over directed layout kNN edges. We next show that this regularity control can be obtained near a geometry optimum with controlled geometry loss.

Proposition 4.3 (First-order regularity gain and second-order geometry cost)**.** *Let $G(Y)$ be a smooth layout-level geometry objective and consider $G(Y) + \lambda L_{\mathrm{Reg}}(Y)$. If G and L_{Reg} are C^2 after gauge fixing (see Appendix A.8), then near any strict local minimum of G with positive-definite reduced Hessian and nonzero readout-regularity gradient, there is a smooth branch of nearby local minima for small $\lambda > 0$ along which L_{Reg} improves at first order while G worsens only at second order.*

Together, these results establish C2: near a local optimum of the geometry objective, the NW regularizer bounds local signal variation while incurring only a second-order increase in that objective.

5 Method

We propose ISDR (***I***nterpretable ***S***upervised ***DR***). Here, we call a layout *interpretable* if it is both geometrically faithful and signal-readable. Guided by the smoothness analysis in Section 4, ISDR combines a t-SNE [7] geometry objective with the NW readout regularizer.

Algorithm 1 ISDR: Supervised DR with Final-Layout NW Regularizer

Require: data X, signals s, weight λ, perplexity p, support size $k = \lfloor 3p \rfloor$, rebuild interval R, refinement steps T, learning rate η

$s \leftarrow (s - \bar{s})/\sigma_s$ ▷ z-score signals
Compute high-D affinity matrix P from X
$Y \leftarrow$ Barnes–Hut t-SNE warm start ▷ geometry-only init
for $t = 1, \dots, T$ **do**
  if $t \bmod R = 1$ **then**
    Rebuild leave-one-out support $\{S_i\}$ from current kNN of Y
  end if
  $g \leftarrow (1 - \lambda)\, \nabla_Y \mathrm{KL}(P \,\|\, Q(Y)) + \lambda\, \nabla_Y L_{\mathrm{Reg}}(Y; \{S_i\})$
  $Y \leftarrow Y - \eta\, g$
end for
return Y

Let $Y = (y_1, \ldots, y_N)^\top \in \mathbb{R}^{N\times 2}$ denote the final layout. ISDR minimizes

$$\mathcal{L}_{\mathrm{ISDR}}(Y) = (1-\lambda)\,\mathrm{KL}\big(P \,\big\|\, Q(Y)\big) + \lambda\, L_{\mathrm{Reg}}(Y), \qquad \lambda \in [0,1], \tag{11}$$

where the first term is the t-SNE geometry objective, which measures the KL divergence between the high-dimensional affinities P and the low-dimensional Student-t affinities $Q(Y)$. The second is the NW readout regularizer, and λ balances these two terms. Algorithm 1 summarizes the optimization process, with full implementation details in Appendix B.

Design rationale. We choose t-SNE for its strong local neighborhood preservation. Signal z-scoring (line 1) makes a fixed λ comparable across datasets. The geometry-only t-SNE warm start (line 3) initializes Y near a geometry-objective optimum, consistent with the setting of Proposition 4.3. Periodic support rebuilding (line 6) keeps the leave-one-out kNN neighborhoods aligned with the evolving layout while reducing the instability and cost of rebuilding them at every step. Between rebuilds, the supports remain fixed, as assumed in Proposition 4.2. Differences between nearby points' supports are captured by the support-overlap residual in Appendix A.7.

6 Evaluation

6.1 Experimental Setup

Datasets. We use two visualization benchmarks with continuous scalar signals. **HPDv3** (*LLM-as-a-Judge* [57, 6, 58]) consists of a fixed 20K subset of image generation comparisons [59]. Each point is represented by the D=4096 hidden state from Qwen3-VL-8B-Instruct [60]. The signal is the judge's logit margin, used as a measure of confidence [61, 62]. **Tabula Muris** (*single-cell biology*) contains the full cross-organ FACS atlas [5] (44,949 cells). Each cell is represented by 50 principal components of log-normalized expression for highly variable genes. The signal is a globally z-scored dissociation score computed from the published dissociation-associated gene set [63].

Metrics. We assess geometric faithfulness primarily by trustworthiness [64], complemented by continuity and kNN overlap [9, 65]. We measure signal readability using the Dirichlet energy [66] (DE) of the normalized signal on the directed layout kNN graph. Lower DE indicates less signal variation between neighboring points. All reported DE values are multiplied by 100. We also report Φ_σ and M_{ctr} to link the empirical results to the theoretical analysis. We additionally report Label Recovery, measured by Macro-F1 (%) on labels excluded from supervision (Appendix C.7).

Baselines. We compare ISDR with unsupervised anchors (PCA [67], t-SNE [7], UMAP [8], PaCMAP [50]), relation-driven supervised DR (S-UMAP [8], RF-PHATE [29], CEBRA [31]), and target-driven supervised DR (TP-SE [40], RF-AE [9], AKRMap [13]). For methods with adjustable supervision weights, we compare the full trade-off curves between DE and trustworthiness. Fixed methods are represented by single points. Results are averaged over five random seeds, with error bars and $\pm$ values indicating standard deviations. Appendix B provides implementation details, and

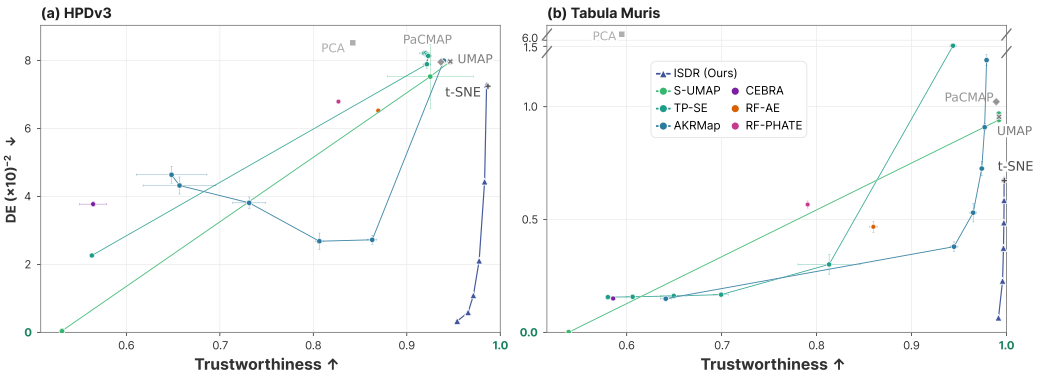

Figure 2: Trade-offs between geometric faithfulness and signal readability on HPDv3 and Tabula Muris. Lower DE and higher trustworthiness are better. Curves include all tested supervision weights, while points represent fixed baselines.

Table 1: Representative operating points from Figure 2. Swept methods use the knee point (Appendix C.3); fixed baselines contribute their single point.

Dataset	Method	Trust ↑	Cont. ↑	Overlap ↑	DE ↓	Φ_σ ↓	M_{ctr}	Label Recovery ↑
HPDv3	t-SNE	0.987±0.000	0.978±0.000	0.323±0.000	7.24±0.01	0.200±0.000	0.994±0.000	51.13±0.43
	S-UMAP	0.925±0.041	0.967±0.031	0.082±0.026	7.53±0.84	0.310±0.073	0.961±0.033	43.50±0.69
	CEBRA	0.564±0.013	0.667±0.008	0.002±0.000	3.77±0.06	0.384±0.008	0.734±0.006	34.15±0.39
	RF-PHATE	0.827±0.001	0.816±0.001	0.022±0.000	6.79±0.02	0.626±0.007	0.947±0.001	45.59±0.74
	RF-AE	0.869±0.001	0.815±0.008	0.028±0.000	6.53±0.05	0.389±0.010	0.956±0.010	37.55±0.72
	TP-SE	0.922±0.004	0.964±0.005	0.038±0.002	7.89±0.11	0.247±0.050	0.954±0.015	35.50±0.21
	AKRMap	0.863±0.004	0.952±0.001	0.019±0.001	2.72±0.12	0.233±0.027	0.940±0.008	44.96±2.35
	ISDR	0.977±0.000	0.978±0.000	0.252±0.001	2.11±0.01	0.208±0.000	0.989±0.001	50.68±0.68
Tabula Muris	t-SNE	0.998±0.000	0.996±0.000	0.380±0.000	0.67±0.00	0.217±0.014	0.986±0.027	80.96±0.19
	S-UMAP	0.992±0.000	0.996±0.000	0.171±0.001	0.94±0.00	0.281±0.007	0.737±0.005	74.53±0.18
	CEBRA	0.586±0.002	0.848±0.000	0.005±0.000	0.15±0.00	0.544±0.017	0.167±0.003	9.25±0.46
	RF-PHATE	0.791±0.002	0.936±0.001	0.038±0.000	0.57±0.01	0.326±0.004	0.689±0.004	37.48±0.46
	RF-AE	0.860±0.004	0.944±0.002	0.047±0.000	0.47±0.02	0.294±0.004	0.810±0.007	44.05±0.22
	TP-SE	0.813±0.029	0.922±0.018	0.022±0.003	0.30±0.04	0.648±0.091	0.450±0.075	33.57±4.18
	AKRMap	0.945±0.004	0.990±0.001	0.054±0.002	0.38±0.02	0.274±0.017	0.852±0.010	54.60±0.96
	ISDR	0.996±0.000	0.995±0.000	0.319±0.001	0.23±0.00	0.218±0.014	0.984±0.027	80.02±0.30

Appendix C reports full sweeps and linear comparisons. Additional regression and classification results are reported in Appendices E and F, respectively.

6.2 Geometry–Signal Pareto Results

Figure 2 reports the primary quantitative result, read as a trade-off curve: a method improves on the baselines only by attaining lower DE at comparable trustworthiness, not at a single favorable operating point. Table 1 summarizes representative operating points; complete sweeps appear in Appendix C, and Figures 1 and 6 visualize the layouts. Geometry-only methods (i.e., t-SNE, UMAP, and PaCMAP) serve as geometry anchors because the relationship between signal and data geometry varies across datasets.

Supervised baselines suffer geometric distortion without consistent readability gains (C1). Swept methods lose trustworthiness as λ grows, while single-point baselines generally have lower trustworthiness than geometry anchors. RF-PHATE and RF-AE sacrifice substantial trustworthiness for modest DE reductions, suggesting that their random-forest step is not well matched to these continuous signals. CEBRA reaches lower DE only with a substantial trustworthiness drop, as do TP-SE, S-UMAP, and AKRMap at high λ.

Readability improvement does not imply contraction (C1). Geometric distortion Φ_σ and the contraction proxy M_{ctr} help distinguish the mechanisms behind these Pareto points. In the full grid (Table 7), the swept baselines drive M_{ctr} down as λ grows, especially in the low-trust low-DE regime, indicating that these methods increasingly favor contraction. By contrast, ISDR reduces DE while leaving M_{ctr} nearly unchanged, showing that readability can improve without signal-neighbor compression. The contraction effect is visible in Figures 1 and 6.

ISDR improves readability robustly while preserving geometric faithfulness (C2). On both datasets (Figure 2), ISDR keeps high trustworthiness across the full λ grid while reducing DE consistently. At comparable DE, the supervised baselines have lower trustworthiness than ISDR: S-UMAP and TP-SE show a steep trustworthiness drop as DE decreases, while AKRMap on HPDv3 is non-monotone, with DE rising again at high λ. Figure 1 confirms this qualitatively: ISDR preserves the overall layout structure while refining local signal variation. Combined with the stable M_{ctr} values, this supports the direct smoothness route: final-layout smoothness can improve scalar-field readability with controlled geometry loss.

6.3 Initialization Ablation Studies

We conduct three ablation studies to evaluate whether geometry initialization is necessary, whether the geometry-only stage is robust to raw initializers, and whether initialization alone explains the performance advantage (Table 2).

Necessity of geometry-first initialization. To assess the role of geometry initialization, we apply the same ISDR refinement to t-SNE, PCA, Gaussian-random, and UMAP layouts. Table 2(a) shows that

Table 2: Initialization ablations on HPDv3. Panels (a,b) use $\lambda = 0.4$ with standard and fixed-weight refinement, respectively. Full settings and results with standard deviations are in Appendix C.9.

(a) Choice of refinement start.

Start	Trust ↑	DE ↓
t-SNE	**0.9775**	**2.11**
PCA	0.9620	5.42
Random	0.9625	5.34
UMAP	0.9649	4.87

(b) Robustness to t-SNE initialization.

Stage	Metric	PCA	Random	UMAP
After t-SNE	Trust ↑	0.9869	0.9869	0.9868
	DE ↓	7.243	7.253	7.240
After ISDR	Trust ↑	0.9816	0.9816	0.9816
	DE ↓	2.598	2.591	2.597

(c) Comparison under shared t-SNE initialization.

Metric	Method	$\lambda = 0$	0.2	0.4	0.6	0.8	1.0
Trust ↑	ISDR	0.9854	0.9831	0.9776	0.9711	0.9659	0.9539
	S-UMAP	0.9484	0.9477	0.9290	0.6401	0.6402	0.6392
DE ↓	ISDR	7.281	4.445	2.111	1.088	0.590	0.328
	S-UMAP	7.929	7.935	7.532	1.714	1.718	1.699

refinement from t-SNE yields the highest trustworthiness and lowest DE. UMAP outperforms PCA and random starts on both metrics, but still yields substantially higher DE than t-SNE. These results are consistent with the theoretical motivation for starting refinement near a geometry optimum.

Robustness to raw initialization. To assess the robustness of the geometry-only stage, we initialize t-SNE with PCA, Gaussian-random, or UMAP coordinates, followed by fixed-weight ISDR refinement. Across the three raw initializers, trustworthiness and DE remain nearly identical both after geometry-only t-SNE and after ISDR refinement (Tables 2(b) and 12(b)). These results indicate that a suitable geometric starting point can be obtained without carefully tuning the raw initializer.

Advantage beyond initialization. To test whether initialization alone explains the performance gap, we give ISDR and S-UMAP identical t-SNE starting coordinates for each paired seed. Table 2(c) shows that S-UMAP remains dominated across supervision weights. Thus, initialization alone does not explain the performance gap. Together with the first ablation, this result supports the effectiveness of ISDR's geometry-faithful initialization and smoothness refinement as a unified strategy.

6.4 Case Study

To examine whether the resulting maps support sample-level visual analysis, a researcher on our team with expertise in LLM-as-a-Judge evaluation used ISDR to analyze judge behavior on HPDv3 before adopting the judge for subsequent research. The expert explored judge confidence using a simple visualization system that supports zooming, region selection, and sample inspection (see Appendix D for details). Figure 3 shows the overview map together with the regions discussed below.

Data geometry patterns. Globally, the expert identified two coherent confidence clusters and checked the samples in these regions. Region A is a low-confidence island whose prompts consist only of the literal phrase "The prompt in English translates to:". Checking the original dataset showed that this pattern is already present in HPDv3, so the low confidence is expected: such prompts carry little semantic content for judging the generated images. Region B is a high-confidence cluster in which one candidate image is blank or nearly blank, indicating a failure in image generation. This makes the comparison largely depend on the non-blank image, resulting in high judge confidence. Geometric faithfulness preserves the grouping of related samples, allowing the expert to trace shared confidence patterns to common causes. Representative inspected samples and artifact checks are provided in Appendices D.3 and D.4.

Local confidence transitions. At a finer scale, the expert inspected a region whose samples each contain one real and one generated image, with the prompt written based on the real image. Across the local transition, judge confidence is generally low when the generated image is selected (Figure 3C), and higher when the real image is selected (Figure 3D). This suggests uncertainty when both images

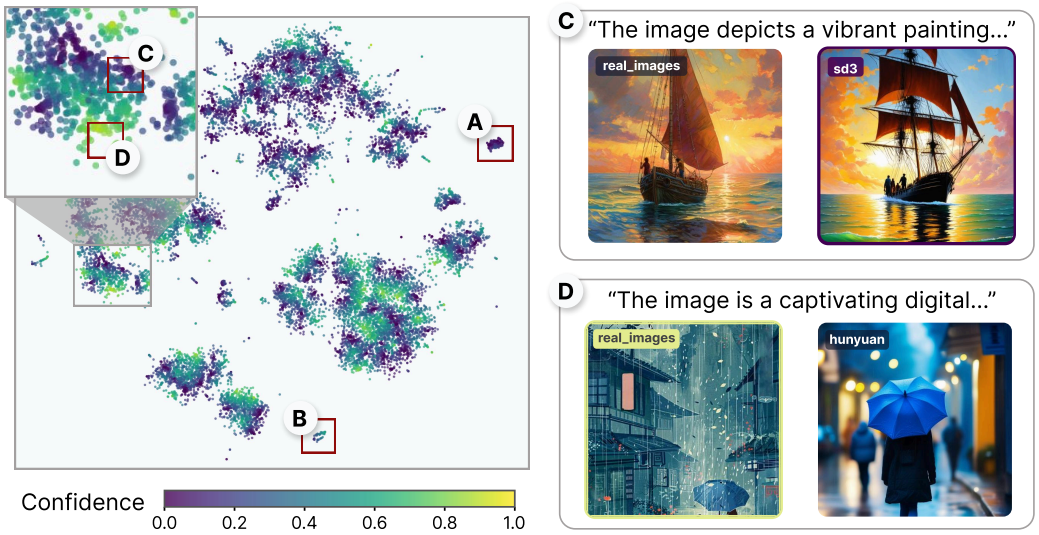

Figure 3: Case-study overview on HPDv3. Regions A and B highlight confidence clusters associated with prompt artifacts and blank-image failures, respectively; C and D show a local transition between generated- and real-image preferences.

are well aligned with the prompt, even if one is preferred. The gradual color transition across this coherent neighborhood made the pattern legible.

Together, the case study illustrates how geometric faithfulness supports the interpretation of spatial structures, while readout smoothness makes local signal transitions visible.

7 Conclusion

In this work, we formulate supervised DR for visualization as a bi-objective problem balancing geometric faithfulness and signal readability. Drawing on the bi-Lipschitz conditions of the induced readout, we use finite-sample proxies to distinguish contraction from smoothness. Enforcing contraction can impose a sharp lower bound on geometric distortion without guaranteeing signal readability. Near a local optimum of the geometry objective, smoothness regularization can yield first-order gains in readout regularity with only a second-order increase in that objective. On HPDv3 and Tabula Muris, ISDR improves the trade-off by reducing local signal variation while preserving neighborhoods. Our framework assumes scalar signals, and the nonparametric ISDR requires recomputation to embed new points. Future directions include tensor-valued signals and their links to multi-view DR (see Appendix G), parametric variants, and distortion analysis beyond the bi-Lipschitz setting.

Acknowledgments and Disclosure of Funding

The authors would like to thank Yukai Guo and Jiangning Zhu for their helpful discussions. This work was supported by the National Natural Science Foundation of China under grant 62636008 and the Capital's Funds for Health Improvement and Research (Code: 2026-2G-2076). The authors have no additional competing interests to declare.

References

[1] Luis Gustavo Nonato and Michaël Aupetit. Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. *IEEE Transactions on Visualization and Computer Graphics*, 25(8):2650–2673, 2019. doi: 10.1109/TVCG.2018.2846735.

[2] Hyeon Jeon, Hyunwook Lee, Yun-Hsin Kuo, Taehyun Yang, Daniel Archambault, Sungahn Ko, Takanori Fujiwara, Kwan-Liu Ma, and Jinwook Seo. Unveiling high-dimensional backstage:

A survey for reliable visual analytics with dimensionality reduction. In *ACM CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2025. doi: 10.1145/3706598.3713551.

[3] Jiazhi Xia, Yuchen Zhang, Jie Song, Yang Chen, Yunhai Wang, and Shixia Liu. Revisiting dimensionality reduction techniques for visual cluster analysis: An empirical study. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):529–539, 2022. doi: 10.1109/TVCG.2021.3114694.

[4] Mengchen Liu, Jiaxin Shi, Zhen Li, Chongxuan Li, Jun Zhu, and Shixia Liu. Towards better analysis of deep convolutional neural networks. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):91–100, 2017.

[5] Nicholas Schaum, Jim Karkanias, Norma F. Neff, Andrew P. May, Stephen R. Quake, Tony Wyss-Coray, Spyros Darmanis, et al. Single-cell transcriptomics of 20 mouse organs creates a tabula muris. *Nature*, 562(7727):367–372, 2018. doi: 10.1038/s41586-018-0590-4.

[6] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623, 2023.

[7] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[8] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[9] Adrien Aumon, Shuang Ni, Myriam Lizotte, Guy Wolf, Kevin R. Moon, and Jake Slater Rhodes. Random forest autoencoders for guided representation learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[10] Joshua T. Vogelstein, Eric W. Bridgeford, Minh Tang, Da Zheng, Christopher Douville, Randal Burns, and Mauro Maggioni. Supervised dimensionality reduction for big data. *Nature Communications*, 12(1):2872, 2021. doi: 10.1038/s41467-021-23102-2.

[11] Laureta Hajderanj, Daqing Chen, and Isakh Weheliye. The impact of supervised manifold learning on structure preserving and classification error: A theoretical study. *IEEE Access*, 9:43909–43922, 2021. doi: 10.1109/ACCESS.2021.3066259.

[12] Cem Örnek and Elif Vural. Nonlinear supervised dimensionality reduction via smooth regular embeddings. *Pattern Recognition*, 87:55–66, 2019. doi: 10.1016/j.patcog.2018.10.006.

[13] Yilin Ye, Junchao Huang, Xingchen Zeng, Jiazhi Xia, and Wei Zeng. AKRMap: Adaptive kernel regression for trustworthy visualization of cross-modal embeddings. In *International Conference on Machine Learning*, 2025.

[14] Armin Eftekhari and Michael B. Wakin. What happens to a manifold under a bi-lipschitz map? *Discrete & Computational Geometry*, 57(3):641–673, 2017. doi: 10.1007/s00454-016-9847-6.

[15] Ittai Abraham, Yair Bartal, and Ofer Neimany. Advances in metric embedding theory. In *ACM Symposium on Theory of Computing*, pages 271–286. Association for Computing Machinery, 2006. doi: 10.1145/1132516.1132557.

[16] Yair Bartal, Nova Fandina, and Ofer Neiman. Dimensionality reduction: theoretical perspective on practical measures. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019.

[17] Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991. doi: 10.1080/01621459.1991.10475035.

[18] R. Dennis Cook. Principal components, sufficient dimension reduction, and envelopes. *Annual Review of Statistics and Its Application*, 5:533–559, 2018. doi: 10.1146/annurev-statistics-031017-100257.

[19] Siqi Liang, Yan Sun, and Faming Liang. Nonlinear sufficient dimension reduction with a stochastic neural network. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27360–27373, 2022.

[20] Qin Wang and Edmund Osei. Boosted sliced regression for dimension reduction in binary classification. *Computational Statistics & Data Analysis*, 218:108342, 2026. doi: 10.1016/j.csda.2026.108342.

[21] Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray, and Mark Crowley. Sufficient dimension reduction for high-dimensional regression and low-dimensional embedding: Tutorial and survey. *arXiv preprint arXiv:2110.09620*, 2021.

[22] Kenji Fukumizu, Francis Bach, and Michael Jordan. Kernel dimensionality reduction for supervised learning. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16, 2003.

[23] Kenji Fukumizu, Francis R. Bach, and Michael I. Jordan. Kernel dimension reduction in regression. *The Annals of Statistics*, 37(4):1871–1905, 2009. doi: 10.1214/08-AOS637.

[24] Chieh Wu, Jared Miller, Yale Chang, Mario Sznaier, and Jennifer Dy. Solving interpretable kernel dimensionality reduction. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, 2019.

[25] Elnaz Barshan, Ali Ghodsi, Zohreh Azimifar, and Mansoor Zolghadri Jahromi. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition*, 44(7):1357–1371, 2011. doi: 10.1016/j.patcog.2010.12.015.

[26] Eric Bair, Trevor Hastie, Debashis Paul, and Robert Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137, 2006. doi: 10.1198/016214505000000628.

[27] Duan Li, Jun Yuan, Xinyuan Guo, Xiting Wang, Yang Liu, Weikai Yang, and Shixia Liu. NCP: Neighborhood-preserving non-uniform circle packing method for visualization. *Computational Visual Media*, 2026. to appear.

[28] Guoqing Chao, Yuan Luo, and Weiping Ding. Recent advances in supervised dimension reduction: A survey. *Machine Learning and Knowledge Extraction*, 1(1):341–358, 2019. doi: 10.3390/make1010020.

[29] Jake S. Rhodes, Adrien Aumon, Sacha Morin, Marc Girard, Catherine Larochelle, Elsa Brunet-Ratnasingham, Amélie Pagliuzza, Lorie Marchitto, Wei Zhang, Adele Cutler, Francois Grand'Maison, Anhong Zhou, Andrés Finzi, Nicolas Chomont, Daniel E. Kaufmann, Stephanie Zandee, Alexandre Prat, Guy Wolf, and Kevin R. Moon. Gaining biological insights through supervised data visualization. *bioRxiv*, 2024. doi: 10.1101/2023.11.22.568384.

[30] Kevin R. Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B. Burkhardt, William S. Chen, Kristina Yim, Antonia van den Elzen, Matthew J. Hirn, Ronald R. Coifman, Natalia B. Ivanova, Guy Wolf, and Smita Krishnaswamy. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 37(12):1482–1492, 2019. doi: 10.1038/s41587-019-0336-3.

[31] Steffen Schneider, Jin Hwa Lee, and Mackenzie Weygandt Mathis. Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, 617(7960):360–368, 2023.

[32] Jacob Goldberger, Geoffrey E. Hinton, Sam Roweis, and Russ R. Salakhutdinov. Neighbourhood components analysis. *Advances in Neural Information Processing Systems*, 17, 2004.

[33] Ruslan Salakhutdinov and Geoff Hinton. Learning a nonlinear embedding by preserving class neighbourhood structure. In Marina Meila and Xiaotong Shen, editors, *International Conference on Artificial Intelligence and Statistics*, volume 2, pages 412–419. PMLR, 2007.

[34] Ronald A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[35] Yuan Wang, Yunde Jia, Changbo Hu, and Matthew Turk. Fisher non-negative matrix factorization for learning local features. In *Asian Conference on Computer Vision*, pages 27–30, 2004.

[36] Tomojit Ghosh and Michael Kirby. Supervised dimensionality reduction and visualization using centroid-encoder. *Journal of Machine Learning Research*, 23(20):1–34, 2022.

[37] Tomojit Ghosh and Michael Kirby. Linear centroid encoder for supervised principal component analysis. *Pattern Recognition*, 155:110634, 2024. doi: 10.1016/j.patcog.2024.110634.

[38] Mithun Das Gupta and Jing Xiao. Non-negative matrix factorization as a feature selection tool for maximum margin classifiers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2841–2848, 2011. doi: 10.1109/CVPR.2011.5995492.

[39] Liping Jing, Chao Zhang, and Michael K. Ng. Snmfca: Supervised nmf-based image classification and annotation. *IEEE Transactions on Image Processing*, 21(11):4508–4521, 2012. doi: 10.1109/TIP.2012.2206040.

[40] Ioannis A. Nellas, Sotiris K. Tasoulis, Spiros V. Georgakopoulos, and Vassilis P. Plagianakos. Two phase cooperative learning for supervised dimensionality reduction. *Pattern Recognition*, 144:109871, 2023. doi: 10.1016/j.patcog.2023.109871.

[41] Yunhai Wang, Kang Feng, Xiaowei Chu, Jian Zhang, Chi-Wing Fu, Michael Sedlmair, Xiaohui Yu, and Baoquan Chen. A perception-driven approach to supervised dimensionality reduction for visualization. *IEEE Transactions on Visualization and Computer Graphics*, 24(5):1828–1840, 2018. doi: 10.1109/TVCG.2017.2701829.

[42] Lawrence Cayton and Sanjoy Dasgupta. Robust euclidean embedding. In *International Conference on Machine Learning*, pages 169–176. Association for Computing Machinery, 2006. doi: 10.1145/1143844.1143866.

[43] Jiří Matoušek. Bi-lipschitz embeddings into low-dimensional euclidean spaces. *Commentationes Mathematicae Universitatis Carolinae*, 31(3):589–600, 1990.

[44] William B. Johnson, Joram Lindenstrauss, et al. Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1, 1984.

[45] J. Bourgain. On lipschitz embedding of finite metric spaces in hilbert space. *Israel Journal of Mathematics*, 52(1):46–52, 1985. doi: 10.1007/BF02776078.

[46] Leena Chennuru Vankadara and Ulrike von Luxburg. Measures of distortion for machine learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, 2018.

[47] Ainesh Bakshi, Vincent Cohen-Addad, Rajesh Jayaram, Samuel B. Hopkins, and Silvio Lattanzi. Metric embeddings beyond bi-lipschitz distortion via sherali-adams. In Nika Haghtalab and Ankur Moitra, editors, *Conference on Learning Theory*, volume 291, pages 264–279. PMLR, 2025.

[48] T. Tony Cai and Rong Ma. Theoretical foundations of t-sne for visualizing high-dimensional clustered data. *Journal of Machine Learning Research*, 23(301):1–54, 2022.

[49] Sebastian Damrich, Jan Niklas Böhm, Fred A. Hamprecht, and Dmitry Kobak. From t-sne to umap with contrastive learning. In *International Conference on Learning Representations*, 2023.

[50] Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *Journal of Machine Learning Research*, 22(201):1–73, 2021.

[51] Elif Vural and Christine Guillemot. A study of the classification of low-dimensional data with supervised manifold learning. *Journal of Machine Learning Research*, 18(157):1–55, 2018.

[52] Colin Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann, 2019.

[53] Anne-Flore Cabouat, Tingying He, Petra Isenberg, and Tobias Isenberg. PREVis: Perceived readability evaluation for visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):1083–1093, 2025. doi: 10.1109/TVCG.2024.3456318.

[54] Liang Zhou and Charles D. Hansen. A survey of colormaps in visualization. *IEEE Transactions on Visualization and Computer Graphics*, 22(8):2051–2069, 2016. doi: 10.1109/TVCG.2015.2489649.

[55] E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1): 141–142, 1964. doi: 10.1137/1109020.

[56] Taha Hussein Ali, Heyam Abd Al-Majeed Hayawi, and Delshad Shaker Ismael Botani. Estimation of the bandwidth parameter in nadaraya-watson kernel non-parametric regression based on universal threshold level. *Communications in Statistics - Simulation and Computation*, 52(4): 1476–1489, 2023. doi: 10.1080/03610918.2021.1884719.

[57] Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. LLMScore: Unveiling the power of large language models in text-to-image synthesis evaluation. In *Advances in Neural Information Processing Systems*, volume 36, pages 23075–23093, 2023.

[58] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment: opportunities and challenges of LLM-as-a-Judge. In *Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, 2025.

[59] Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. HPSv3: Towards wide-spectrum human preference score. In *IEEE/CVF International Conference on Computer Vision*, pages 15086–15095, 2025.

[60] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[61] Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[62] Yongjin Yang, Haneul Yoo, and Hwaran Lee. MAQA: Evaluating uncertainty quantification in LLMs regarding data uncertainty. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5861–5878. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.325.

[63] Susanne C van den Brink, Fanny Sage, Ábel Vértesy, Bastiaan Spanjaard, Josi Peterson-Maduro, Chloé S Baron, Catherine Robin, and Alexander van Oudenaarden. Single-cell sequencing reveals dissociation-induced gene expression in tissue subpopulations. *Nature Methods*, 14(10): 935–936, 2017. doi: 10.1038/nmeth.4437.

[64] Jarkko Venna and Samuel Kaski. Local multidimensional scaling. *Neural Networks*, 19(6): 889–899, 2006. doi: 10.1016/j.neunet.2006.05.014.

[65] Mateus Espadoto, Rafael M. Martins, Andreas Kerren, Nina S. T. Hirata, and Alexandru C. Telea. Toward a quantitative survey of dimension reduction techniques. *IEEE Transactions on Visualization and Computer Graphics*, 27(3):2153–2173, 2021. doi: 10.1109/TVCG.2019.2944182.

[66] Kaixiong Zhou, Xiao Huang, Daochen Zha, Rui Chen, Li Li, Soo-Hyun Choi, and Xia Hu. Dirichlet energy constrained learning for deep graph neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 21834–21846, 2021.

[67] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi: 10.1080/14786440109462720.

[68] Haiyang Huang, Yingfan Wang, and Cynthia Rudin. Navigating the effect of parametrization for dimensionality reduction. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 12977–13019, 2024.

[69] R. Kelley Pace and Ronald Barry. Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3):291–297, 1997. doi: 10.1016/S0167-7152(96)00140-X.

[70] Kam Hamidieh. A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346–354, 2018. doi: 10.1016/j.commatsci.2018.07.052.

[71] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

[72] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

[73] Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. MedMNIST v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023. doi: 10.1038/s41597-022-01721-8.

[74] Peng Lai, Jianjie Zheng, Sijie Cheng, Yun Chen, Peng Li, Yang Liu, and Guanhua Chen. Beyond the surface: Enhancing LLM-as-a-Judge alignment with human via internal representations. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

[75] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-Judge: Assessing multimodal LLM-as-a-Judge with vision-language benchmark. In *International Conference on Machine Learning*, volume 235, pages 6562–6595, 2024.

[76] Brian Montambault, Gabriel Appleby, Jen Rogers, Camelia D. Brumar, Mingwei Li, and Remco Chang. Dimbridge: Interactive explanation of visual patterns in dimensionality reductions with predicate logic. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):207–217, 2025. doi: 10.1109/TVCG.2024.3456391.

[77] Aeri Cho, Hyeon Jeon, Kiroong Choe, Seokhyeon Park, and Jinwook Seo. Toward more explainable nonlinear dimensionality reduction: A feature-driven interaction approach. *IEEE Transactions on Visualization and Computer Graphics*, 32(2):1835–1849, 2026. doi: 10.1109/TVCG.2025.3622114.

[78] Khairi Reda, Pratik Nalawade, and Kate Ansah-Koi. Graphical perception of continuous quantitative maps: the effects of spatial frequency and colormap design. In *ACM CHI Conference on Human Factors in Computing Systems*, pages 1–12. Association for Computing Machinery, 2018. doi: 10.1145/3173574.3173846.

[79] Yong Luo, Dacheng Tao, Kotagiri Ramamohanarao, Chao Xu, and Yonggang Wen. Tensor canonical correlation analysis for multi-view dimension reduction. *IEEE Transactions on Knowledge and Data Engineering*, 27(11):3111–3124, 2015. doi: 10.1109/TKDE.2015.2445757.

A Theory Supplement

This appendix proves the results in Section 4, relates the finite-sample proxies to readout conditions and the reported DE metric, and describes contraction-directed terms in the baselines.

A.1 Finite-Sample Notation

Let $T_N = \binom{[N]}{2}$ be the unordered sample-pair set. For pairwise distinct inputs, write

$$\rho_f(i,j) = \frac{\|y_i - y_j\|}{d_{\mathcal{X}}(x_i, x_j)}, \qquad \bar{\rho}_f = \frac{1}{|T_N|} \sum_{(i,j)\in T_N} \rho_f(i,j). \tag{12}$$

The finite-sample σ-distortion is

$$\Phi_\sigma(f) = \frac{1}{|T_N|} \sum_{(i,j)\in T_N} \left(\frac{\rho_f(i,j)}{\bar{\rho}_f} - 1 \right)^2. \tag{13}$$

For a layout radius $r > 0$, define

$$\mathcal{N}_r(Y) = \{(i,j) \in T_N : \|y_i - y_j\| \le r\}, \qquad M_{\mathrm{sm}}(Y;r) = \frac{1}{|\mathcal{N}_r(Y)|} \sum_{(i,j)\in\mathcal{N}_r(Y)} |s_i - s_j|^2. \tag{14}$$

Let $d_k(y_i)$ be the distance from y_i to its k-th nearest layout neighbor and

$$\ell_k(Y) = \frac{1}{N} \sum_{i=1}^{N} d_k(y_i), \qquad M_{\mathrm{sv}}(Y;\tau) = M_{\mathrm{sm}}(Y;\tau\ell_k(Y)). \tag{15}$$

We abbreviate $r_\tau(Y) = \tau\ell_k(Y)$ and $\mathcal{N}_\tau(Y) = \mathcal{N}_{r_\tau(Y)}(Y)$. For a signal-difference band $0 \le \delta \le \epsilon_s$, let

$$\mathcal{E}_{\delta,\epsilon_s} = \{(i,j) \in T_N : \delta \le |s_i - s_j| \le \epsilon_s\}. \tag{16}$$

When this set is nonempty, the contraction proxy is

$$M_{\mathrm{ctr}}(f;\delta,\epsilon_s) = \frac{1}{|\mathcal{E}_{\delta,\epsilon_s}|} \sum_{(i,j)\in\mathcal{E}_{\delta,\epsilon_s}} \frac{\rho_f(i,j)}{\bar{\rho}_f}. \tag{17}$$

All finite-sample averages in this appendix are interpreted only when their denominator is nonzero.

A.2 Readout Conditions and Finite-Sample Proxies

The following conditions are imposed on a set A of layout points, typically the finite sample. We use co-Lipschitz for the inverse distance inequality below.

Definition A.1 (Readout co-Lipschitz)**.** A readout h is co-Lipschitz on A with constant $c > 0$ if

$$c\|y - y'\| \le |h(y) - h(y')|, \qquad \forall y, y' \in A. \tag{18}$$

Definition A.2 (Readout Lipschitz)**.** A readout h is Lipschitz on A with constant $L > 0$ if

$$|h(y) - h(y')| \le L\|y - y'\|, \qquad \forall y, y' \in A. \tag{19}$$

The Lipschitz condition is the forward smoothness implication: nearby layout points have nearby readout values. The co-Lipschitz condition is the inverse contraction implication: nearby readout values must be mapped nearby in the layout. For pairs in A, the following implications are equivalent to the distance inequalities when they hold for every positive threshold:

$$|h(y) - h(y')| \le L\|y - y'\| \iff \|y - y'\| \le r \Rightarrow |h(y) - h(y')| \le Lr, \tag{20}$$

$$\iff |h(y) - h(y')| > \epsilon \Rightarrow \|y - y'\| > \epsilon/L. \tag{21}$$

Likewise, for any $c > 0$,

$$c\|y - y'\| \le |h(y) - h(y')| \iff |h(y) - h(y')| \le \epsilon \Rightarrow \|y - y'\| \le \epsilon/c, \tag{22}$$

$$\iff \|y - y'\| > r \Rightarrow |h(y) - h(y')| > cr. \tag{23}$$

A condition at a single threshold is a finite-scale relaxation of these inequalities.

A.3 Proxy Relations

Let $h(y_i) = s_i$ on the sample layout. Write $\tilde{\rho}_{ij} = \rho_f(i,j)/\bar{\rho}_f$.

Proposition A.1 (Finite-sample proxy relations)**.** *Assume the relevant pair sets are nonempty. If $\|y_i - y_j\| \le r_\tau(Y)$ implies $|s_i - s_j| \le \epsilon$ for all sampled pairs, then*

$$M_{\mathrm{sv}}(Y;\tau) \le \epsilon^2. \tag{24}$$

If h is L-Lipschitz on all local pairs in $\mathcal{N}_\tau(Y)$, then

$$M_{\mathrm{sv}}(Y;\tau) \le L^2\tau^2\ell_k(Y)^2. \tag{25}$$

Conversely, for every $\epsilon > 0$,

$$\frac{|\{(i,j) \in \mathcal{N}_\tau(Y) : |s_i - s_j| > \epsilon\}|}{|\mathcal{N}_\tau(Y)|} \le \frac{M_{\mathrm{sv}}(Y;\tau)}{\epsilon^2}. \tag{26}$$

For the contraction band $\mathcal{E} = \mathcal{E}_{\delta,\epsilon_s}$, define

$$d_{\min}(\mathcal{E}) = \min_{(i,j)\in\mathcal{E}} d_{\mathcal{X}}(x_i, x_j). \tag{27}$$

If every pair in $\mathcal{E}$ satisfies $\|y_i - y_j\| \le r_c$, then

$$M_{\mathrm{ctr}}(f;\delta,\epsilon_s) \le \frac{r_c}{\bar{\rho}_f|\mathcal{E}|} \sum_{(i,j)\in\mathcal{E}} \frac{1}{d_{\mathcal{X}}(x_i,x_j)} \le \frac{r_c}{\bar{\rho}_f d_{\min}(\mathcal{E})}. \tag{28}$$

Conversely, for every $a > 0$,

$$\frac{|\{(i,j) \in \mathcal{E} : \tilde{\rho}_{ij} > a\}|}{|\mathcal{E}|} \le \frac{M_{\mathrm{ctr}}(f;\delta,\epsilon_s)}{a}. \tag{29}$$

Proof. Average the pointwise smoothness bounds and apply Markov's inequality to $|s_i - s_j|^2$. For contraction, average $\tilde{\rho}_{ij} \le r_c/[\bar{\rho}_f d_{\mathcal{X}}(x_i,x_j)]$, then use $d_{\mathcal{X}}(x_i,x_j) \ge d_{\min}(\mathcal{E})$. The final tail bound is Markov's inequality for $\tilde{\rho}_{ij} \ge 0$. □

A.4 Average Contraction and the Sharp Distortion Floor

The contraction proxy is the band-wise mean of $\tilde{\rho}$, whose global mean is one. Proposition 4.1 gives

$$\Phi_\sigma(f) \ge \frac{p_{\delta,\epsilon_s}}{1 - p_{\delta,\epsilon_s}}\big(1 - M_{\mathrm{ctr}}(f;\delta,\epsilon_s)\big)^2. \tag{30}$$

The bound is sharp as an inequality for the two conditional moments, with equality characterized in the proof below. It gives an absolute distortion floor rather than a comparison with a geometry-only layout.

A.5 Bridging the Local SV Proxy and Dirichlet Energy

The experimental smoothness metric (Appendix B.5) is the Dirichlet energy on the directed layout kNN graph using min–max normalized scores $\tilde{s}_i = (s_i - s_{\min})/(s_{\max} - s_{\min})$:

$$\mathrm{DE}(Y,s;k) = \frac{1}{Nk} \sum_{(i,j)\in E_k^{\rightarrow}(Y)} (\tilde{s}_i - \tilde{s}_j)^2, \qquad E_k^{\rightarrow}(Y) = \{(i,j) : j \in \mathrm{kNN}_Y(i)\}. \tag{31}$$

Note $|E_k^{\rightarrow}(Y)| = Nk$, and each unordered pair $\{i,j\}$ contributes to the sum at most twice (once for each direction in kNN). The theoretical proxy $M_{\mathrm{sv}}(Y;\tau)$ uses the τ-scale ball graph $\mathcal{N}_\tau(Y)$ on raw scores. Let $\Delta s = s_{\max} - s_{\min} > 0$ and $r^*(Y) = \max_i d_k(y_i)$. Define the layout-density factor

$$\beta(Y;\tau) = \frac{2\,|\mathcal{N}_\tau(Y)|}{Nk\,\Delta s^2}. \tag{32}$$

Lemma A.2 (M_{sv} controls Dirichlet energy)**.** *Assume $\Delta s > 0$. For every $\tau \ge r^*(Y)/\ell_k(Y)$,*

$$\mathrm{DE}(Y,s;k) \le \beta(Y;\tau)\, M_{\mathrm{sv}}(Y;\tau). \tag{33}$$

Proof. Each ordered edge $(i,j) \in E_k^{\rightarrow}(Y)$ satisfies $\|y_i - y_j\| \le d_k(y_i) \le r^*(Y) \le \tau \ell_k(Y)$, so the underlying unordered pair $\{i,j\}$ lies in $\mathcal{N}_\tau(Y)$. Each unordered pair contributes to the directed sum at most twice, hence

$$\sum_{(i,j)\in E_k^{\rightarrow}(Y)} (s_i - s_j)^2 \le 2 \sum_{\{i,j\}\in\mathcal{N}_\tau(Y)} (s_i - s_j)^2 = 2\,|\mathcal{N}_\tau(Y)|\, M_{\mathrm{sv}}(Y;\tau). \tag{34}$$

Normalizing $(\tilde{s}_i - \tilde{s}_j)^2 = (s_i - s_j)^2/\Delta s^2$ and dividing by Nk gives the claim. □

The factor β is bounded across a sample sequence if $|\mathcal{N}_\tau(Y)| = O(Nk)$ and Δs has a uniform positive lower bound. Applying the lemma at $\tau = 1$ requires all kNN edges to lie within radius $\ell_k(Y)$.

A.6 Assumptions for the Final-Layout kNN Readout

For a fixed nonempty leave-one-out support S_i, recall the readout from Section 4:

$$\hat{s}_i(y) = \frac{\sum_{j\in S_i} K_t(y,y_j) s_j}{\sum_{j\in S_i} K_t(y,y_j)}. \tag{35}$$

Here $K_t(y,y') = (1 + \|y - y'\|^2)^{-1} > 0$, $|s_i| \le S_{\max}$, and $\Delta s = s_{\max} - s_{\min}$. Positivity makes the denominator nonzero for every query. Set $g_j(y) = \nabla_y \log K_t(y,y_j)$, so $\|g_j(y)\| \le 1$ and

$$\nabla \hat{s}_i(y) = \sum_j \pi_i^y(j)\big(s_j - \hat{s}_i(y)\big) g_j(y). \tag{36}$$

Cauchy–Schwarz and $\mathrm{Var}_{\pi_i^y}(s) \le (\Delta s)^2/4$ give

$$\|\nabla \hat{s}_i(y)\| \le \Delta s/2 =: L_{\mathrm{Reg}}^{(k)} \le S_{\max}. \tag{37}$$

The constant is independent of both k and N, and is attained at query 0 with support points $-1, +1$ carrying the extreme scores. The normalized weights π_i^y are defined below.

For $r > 0$, let

$$D_r(Y) = \max_i |\{j \ne i : (\min\{i,j\}, \max\{i,j\}) \in \mathcal{N}_r(Y)\}|. \tag{38}$$

The local crowding factor used in Proposition 4.2 is

$$\Gamma_\tau(Y) = \frac{N D_{\tau\ell_k(Y)}(Y)}{|\mathcal{N}_\tau(Y)|}, \tag{39}$$

whenever the denominator is nonzero.

A.7 Support Overlap and the General Bound

For a query point z, define the normalized kernel-weight distribution of readout i by

$$\pi_i^z(\ell) = \frac{K_t(z,y_\ell)\mathbf{1}\{\ell \in S_i\}}{\sum_{a\in S_i} K_t(z,y_a)}, \qquad \ell \in [N]. \tag{40}$$

The local support-overlap residual is

$$\Omega_\tau(Y) = \frac{1}{|\mathcal{N}_\tau(Y)|} \sum_{(i,j)\in\mathcal{N}_\tau(Y)} \mathrm{TV}\big(\pi_i^{y_j}, \pi_j^{y_j}\big)^2, \tag{41}$$

where $\mathrm{TV}(p,q) = \frac{1}{2}\sum_\ell |p_\ell - q_\ell|$. Small Ω_τ means neighboring readouts use similar normalized weights at a common query.

For a local edge (i,j), define the *support-mismatch residual*

$$\Delta_{ij}^{(k)} = \hat{s}_i(y_j) - \hat{s}_j(y_j). \tag{42}$$

This term vanishes when the two normalized weight distributions coincide. Centering the scores at their range midpoint gives

$$|\Delta_{ij}^{(k)}| \le \Delta s \, \mathrm{TV}\big(\pi_i^{y_j}, \pi_j^{y_j}\big) \le 2S_{\max} \, \mathrm{TV}\big(\pi_i^{y_j}, \pi_j^{y_j}\big) \, . \tag{43}$$

The proof below combines these bounds to obtain

$$M_{\mathrm{sv}}(Y;\tau) \le c_1 \tau^2 \ell_k(Y)^2 + c_2 \Gamma_\tau(Y) L_{\mathrm{Reg}}(Y) + c_3 S_{\max}^2 \Omega_\tau(Y), \tag{44}$$

with $c_1 = 4(L_{\mathrm{Reg}}^{(k)})^2$, $c_2 = 4$, and $c_3 = 16$. If the distance gap between the k-th and $(k+1)$-st neighbors is positive at every point of Y_0, the supports remain fixed in a neighborhood, as required by the local branch argument.

Corollary A.3 (NW regularizer envelope controls the reported Dirichlet energy)**.** *Assume $\Delta s = s_{\max} - s_{\min} > 0$ and $\tau \ge r^*(Y)/\ell_k(Y)$, where $r^*(Y) = \max_i d_k(y_i)$. Under the assumptions above,*

$$\mathrm{DE}(Y, s; k) \le \beta(Y;\tau) \left[c_1 \tau^2 \ell_k(Y)^2 + c_2 \Gamma_\tau(Y) L_{\mathrm{Reg}}(Y) + c_3 S_{\max}^2 \Omega_\tau(Y) \right] . \tag{45}$$

Proof. Combine Lemma A.2 with the bound in Equation (44). □

A.8 Proofs for the Main Results

Proposition A.4 (Small contraction does not force small local signal variation)**.** *Fix $0 \le \delta < \epsilon_s$ and $\tau > 0$. For any fixed $k \ge 1$, there exists a sequence of Euclidean samples with uniformly bounded signals and planar layouts with $\mathcal{E}_{\delta,\epsilon_s} \neq \varnothing$ such that*

$$M_{\mathrm{ctr}}(f_N; \delta, \epsilon_s) \to 0, \qquad \liminf_{N\to\infty} M_{\mathrm{sv}}(Y_N;\tau) > 0. \tag{46}$$

Proof. Set $a = (\delta + \epsilon_s)/2$ and $H > \epsilon_s + a$. Take $N_m = 4m + 1$ inputs forming a regular simplex of side length one. Place $4m$ layout points at $(j/(4m), 0)$, $j = 1, \dots, 4m$, with scores cycling through $(0, a, H, H + a)$. Place the last point at $(m^3, 0)$ with score $3H$. All band pairs lie in the unit-length cluster, and at least $2m^2$ pairs have score difference a. The outlier makes $\bar{\rho}_f = \Theta(m^2)$, so every band ratio is $O(m^{-2})$ and $M_{\mathrm{ctr}} \to 0$. For fixed k, the outlier also makes $\ell_k(Y) = \Theta(m^2)$. Thus, for all sufficiently large m, the radius $\tau \ell_k(Y)$ includes every cluster pair and excludes the outlier. The local average is consequently the mean squared score difference in the cluster, which tends to $2 \operatorname{Var}(0, a, H, H + a) = (H^2 + a^2)/2 > 0$. This witness keeps the signal range bounded and the band fraction bounded away from zero. It also shows why a mean neighborhood scale can be sensitive to layout outliers. □

Proposition A.5 (Small smoothness does not force small contraction)**.** *Fix $0 \le \delta < \epsilon_s$ and $\tau > 0$. For any fixed $k \ge 1$, there exists a sequence of planar samples with bounded scalar signals such that*

$$M_{\mathrm{sv}}(Y_N;\tau) \to 0, \qquad \liminf_{N\to\infty} M_{\mathrm{ctr}}(f_N; \delta, \epsilon_s) > 0. \tag{47}$$

Proof. Take the identity layout on m equally spaced points in $[0, 1] \times \{0\}$, together with the point $(m^{-3}, 0)$, and assign $s_i = A(x_i)_1$ with $A > \epsilon_s$. All expansion ratios equal one, and the band is nonempty for large m, so $M_{\mathrm{ctr}} = 1$. For any fixed $k \ge 1$, $\ell_k(Y) = \Theta(m^{-1})$. The additional point and the origin are therefore local neighbors for every fixed $\tau > 0$ once m is large, ensuring $\mathcal{N}_\tau(Y) \neq \varnothing$. The scores are A-Lipschitz, so

$$M_{\mathrm{sv}}(Y;\tau) \le A^2 \tau^2 \ell_k(Y)^2 = O(m^{-2}) \to 0. \tag{48}$$

□

Corollary A.6 (No universal inversion from smoothness to contraction)**.** *There is no function ψ with $\psi(t) \to 0$ as $t \downarrow 0$ such that*

$$M_{\mathrm{ctr}}(f; \delta, \epsilon_s) \le \psi(M_{\mathrm{sv}}(Y;\tau)) \tag{49}$$

holds for every planar sample and embedding.

Proof. Apply the assumed inequality to the sequence in Proposition A.5. The right-hand side tends to zero, while the left-hand side stays bounded away from zero. □

Proof of Proposition 4.1. Write $p = |\mathcal{E}|/|T_N| \in (0,1)$, $c = E_{\mathcal{E}}[\tilde{\rho}] = M_{\mathrm{ctr}}$, and $b = E_{\mathcal{E}^c}[\tilde{\rho}]$. Since $E_{T_N}[\tilde{\rho}] = 1$, $b = (1-pc)/(1-p)$. The law of total variance gives the exact identity

$$\Phi_\sigma(f) = p\,\mathrm{Var}_{\mathcal{E}}(\tilde{\rho}) + (1-p)\,\mathrm{Var}_{\mathcal{E}^c}(\tilde{\rho}) + \frac{p}{1-p}(1-c)^2. \tag{50}$$

Dropping the two nonnegative variances proves Equation (30), with equality precisely when the ratios are constant within each group. □

Lemma A.7 (A third-point witness for distortion propagation)**.** *Let $d_{ab} = d_{\mathcal{X}}(x_a, x_b) > 0$ and $R_{ab} = \tilde{\rho}_{ab}$. For any three distinct sample indices,*

$$(R_{ik}-1)^2 + (R_{jk}-1)^2 \geq \frac{\big(|d_{ik}-d_{jk}| - R_{ij}d_{ij}\big)_+^2}{d_{ik}^2 + d_{jk}^2}, \tag{51}$$

where $t_+ = \max\{t, 0\}$.

Proof. The reverse triangle inequality in the layout gives $|d_{ik}R_{ik} - d_{jk}R_{jk}| \leq d_{ij}R_{ij}$. Hence $|d_{ik} - d_{jk}| \leq d_{ij}R_{ij} + d_{ik}|R_{ik}-1| + d_{jk}|R_{jk}-1|$. Apply Cauchy–Schwarz to the last two terms. □

Thus contracting (i,j) also forces distortion on edges to any third point whose original distances to i and j differ sufficiently. Unlike the moment floor, this witness uses the original geometry and applies in any output dimension.

Proof of Proposition 4.2. Use the global fixed-support Lipschitz bound in Equation (37).

Let $e_i = s_i - \hat{s}_i(y_i)$. For every local edge, the identity

$$s_i - s_j = \big(\hat{s}_i(y_i) - \hat{s}_i(y_j)\big) + \Delta_{ij}^{(k)} + e_i - e_j \tag{52}$$

combined with $(a+b+c+d)^2 \leq 4(a^2+b^2+c^2+d^2)$ gives, for every local edge with $\|y_i - y_j\| \leq r$,

$$|s_i - s_j|^2 \leq 4(L_{\mathrm{Reg}}^{(k)})^2 r^2 + 4|\Delta_{ij}^{(k)}|^2 + 4|e_i|^2 + 4|e_j|^2. \tag{53}$$

Averaging over $\mathcal{N}_\tau(Y)$ with $r = \tau\ell_k(Y)$ and using

$$\sum_{(i,j)\in\mathcal{N}_\tau(Y)} (|e_i|^2 + |e_j|^2) \leq N D_r(Y) L_{\mathrm{Reg}}(Y) \tag{54}$$

and Equation (43) gives

$$M_{\mathrm{sv}}(Y;\tau) \leq (\Delta s)^2\big[\tau^2\ell_k(Y)^2 + 4\Omega_\tau(Y)\big] + 4\Gamma_\tau(Y)L_{\mathrm{Reg}}(Y). \tag{55}$$

This envelope is invariant under adding a constant to all scores and implies Equation (44) since $\Delta s \leq 2S_{\max}$. □

Gauge fixing and the reduced Hessian

The geometry objective $G(Y) = \mathrm{KL}(P \,\|\, Q(Y))$ is invariant under rigid motions of the layout: $G(YQ^\top + \mathbf{1}b^\top) = G(Y)$ for every $Q \in O(d)$ and $b \in \mathbb{R}^d$. At a critical layout Y_0, let $\mathcal{G} \subseteq \mathbb{R}^{N\times d}$ be the tangent space of its rigid-motion orbit. Invariance implies that $\nabla^2 G(Y_0)$ annihilates $\mathcal{G}$. The *reduced Hessian* is its restriction to the orthogonal complement,

$$H_G(Y_0) := \nabla^2 G(Y_0)\big|_{\mathcal{G}^\perp}. \tag{56}$$

Positive definiteness of $H_G(Y_0)$ is a nondegeneracy assumption in addition to strict local minimality. The kNN NW regularizer L_{Reg} is also gauge-invariant: $K_t(y,y')$ depends only on $\|y-y'\|$, and the support sets are determined by relative distances, so $\nabla L_{\mathrm{Reg}}(Y_0) \in \mathcal{G}^\perp$. We restrict both objectives to a local affine slice through Y_0 orthogonal to $\mathcal{G}$. Gradients below are taken on this slice.

Proof of Proposition 4.3. Work in gauge-fixed local coordinates on $\mathcal{G}^{\perp}$. Let Y_0 be a strict local minimum of G with positive-definite reduced Hessian $H_G(Y_0)$ and define

$$F(Y, \lambda) = \nabla G(Y) + \lambda \nabla L_{\mathrm{Reg}}(Y). \tag{57}$$

Since $D_Y F(Y_0, 0) = H_G(Y_0)$ is invertible, the implicit function theorem gives a C^1 stationary branch Y_λ for small λ. Differentiating $F(Y_\lambda, \lambda) = 0$ at $\lambda = 0$ gives

$$Y_0' = -H_G(Y_0)^{-1} \nabla L_{\mathrm{Reg}}(Y_0), \qquad Y_\lambda = Y_0 + \lambda Y_0' + o(\lambda). \tag{58}$$

A first-order Taylor expansion of L_{Reg} at Y_0 then yields

$$L_{\mathrm{Reg}}(Y_\lambda) = L_{\mathrm{Reg}}(Y_0) + \lambda \langle \nabla L_{\mathrm{Reg}}(Y_0), Y_0' \rangle + o(\lambda) \tag{59}$$

$$= L_{\mathrm{Reg}}(Y_0) - \lambda \langle \nabla L_{\mathrm{Reg}}(Y_0), H_G(Y_0)^{-1} \nabla L_{\mathrm{Reg}}(Y_0) \rangle + o(\lambda). \tag{60}$$

For G, a second-order Taylor expansion at the critical point Y_0 uses $\nabla G(Y_0) = 0$:

$$G(Y_\lambda) = G(Y_0) + \tfrac{1}{2}(Y_\lambda - Y_0)^\top H_G(Y_0)(Y_\lambda - Y_0) + o(\|Y_\lambda - Y_0\|^2) \tag{61}$$

$$= G(Y_0) + \tfrac{\lambda^2}{2} (Y_0')^\top H_G(Y_0)\, Y_0' + o(\lambda^2) \tag{62}$$

$$= G(Y_0) + \tfrac{\lambda^2}{2} \langle \nabla L_{\mathrm{Reg}}(Y_0), H_G(Y_0)^{-1} \nabla L_{\mathrm{Reg}}(Y_0) \rangle + o(\lambda^2), \tag{63}$$

where the last equality uses $H_G(Y_0)\, Y_0' = -\nabla L_{\mathrm{Reg}}(Y_0)$ from (58). The displayed inner product is positive when $\nabla L_{\mathrm{Reg}}(Y_0) \neq 0$ (since $H_G(Y_0)^{-1} \succ 0$). Positive definiteness of the reduced Hessian is an open condition, so for sufficiently small λ the stationary points Y_λ remain strict local minima. □

For the method objective in Section 5, division by $1-\lambda$ for $0 \leq \lambda < 1$ gives $G + [\lambda/(1-\lambda)] L_{\mathrm{Reg}}$, so the coefficient in Proposition 4.3 corresponds to $\lambda/(1-\lambda)$. The proposition controls $G = \mathrm{KL}(P\|Q)$ and L_{Reg}, rather than Φ_σ or DE directly. In particular, the reported unweighted kNN DE is constant while the layout neighbor graph is unchanged.

A.9 Why the NW Regularizer Differs from AKRMap

The design of the NW regularizer is inspired by AKRMap [13], but the two serve opposite purposes. As discussed in the related work, AKRMap uses the NW loss as a regression objective. Specifically, AKRMap's training loss uses random mini-batch supports, with mini-batches sampled globally rather than from a neighborhood in the layout. Thus, it does not directly supply the final-layout kNN residual L_{Reg} or control Ω_τ in Proposition 4.2. This observation, as well as the separate issue of geometric preservation in parametric DR [68], motivates our nonparametric implementation.

A.10 Implementation-Level Contraction Routes

Below, we theoretically analyze how the optimization objectives of several classical supervised DR methods affect contraction.

A.10.1 S-UMAP

The S-UMAP implementation calls `umap.UMAP` with `target_metric="l2"`, `target_weight=`λ, and scalar targets s_i. After PCA pre-reduction of X, umap-learn constructs a data fuzzy graph with memberships $u_{ij} \in [0,1]$ and a target fuzzy graph with memberships $v_{ij} \in [0,1]$ from the one-dimensional score array. For continuous targets, the code then applies `general_simplicial_set_intersection`. Ignoring only the small numerical floors used for absent sparse entries, the pre-reset intersected membership is

$$\tilde{w}_{ij}^{(\lambda)} = \begin{cases} u_{ij}\, v_{ij}^{\lambda/(1-\lambda)}, & 0 \leq \lambda < 1/2, \\ u_{ij}^{(1-\lambda)/\lambda}\, v_{ij}, & 1/2 \leq \lambda \leq 1. \end{cases} \tag{64}$$

The implementation then calls local-connectivity reset and optimizes the final sparse weights $w_{ij}^{(\lambda)}$ in the usual UMAP layout objective. Equation (64) shows the target route before this reset: for every

$\lambda > 0$, the pre-reset weight is nondecreasing in the score-graph membership v_{ij}, and at $\lambda = 1$ the pre-reset data-graph factor disappears.

Let the low-dimensional UMAP membership be

$$q_{ij}(Y) = \frac{1}{1 + a\|y_i - y_j\|^{2b}}, \qquad a, b > 0. \tag{65}$$

The positive-edge part of the optimized cross entropy is

$$L^{+}_{\text{S-UMAP}}(Y; \lambda) = \sum_{(i,j)} w_{ij}^{(\lambda)} \big[-\log q_{ij}(Y)\big]. \tag{66}$$

Negative sampling adds repulsive terms, but it does not change the sign of the positive-edge attraction in Proposition A.8.

Proposition A.8 (S-UMAP target weights induce positive-edge contraction)**.** *Fix a positive edge (i, j) with $w_{ij}^{(\lambda)} > 0$ and $r_{ij} = \|y_i - y_j\| > 0$. The negative gradient of this edge's contribution to Equation* (66) *points from y_i toward y_j, with magnitude proportional to $w_{ij}^{(\lambda)}$. Moreover, before the local-connectivity reset, increasing the score-graph membership v_{ij} weakly increases this positive-edge weight for every $\lambda > 0$.*

Proof. For $q(r) = (1 + ar^{2b})^{-1}$,

$$\frac{\partial}{\partial r}\big[-\log q(r)\big] = \frac{2abr^{2b-1}}{1 + ar^{2b}} > 0. \tag{67}$$

Therefore

$$\nabla_{y_i} w_{ij}^{(\lambda)} \big[-\log q_{ij}(Y)\big] = w_{ij}^{(\lambda)} \frac{2abr_{ij}^{2b-2}}{1 + ar_{ij}^{2b}}(y_i - y_j). \tag{68}$$

Negating this vector gives the attraction direction. The scalar multiplier is linear in $w_{ij}^{(\lambda)}$. For the implementation's pre-reset intersection, differentiating Equation (64) with respect to v_{ij} gives

$$\partial_{v_{ij}} \tilde{w}_{ij}^{(\lambda)} = \begin{cases} \frac{\lambda}{1-\lambda}\, u_{ij}\, v_{ij}^{\lambda/(1-\lambda)-1}, & 0 \le \lambda < 1/2, \\ u_{ij}^{(1-\lambda)/\lambda}, & 1/2 \le \lambda \le 1. \end{cases} \tag{69}$$

For $u_{ij}, v_{ij} > 0$, both expressions are positive when $\lambda > 0$. At $\lambda = 0$ the weight is independent of v_{ij}. Monotonicity extends to zero memberships by continuity, without assuming a finite derivative there. □

This result describes edge-wise attraction within the full objective, which also includes repulsive terms.

A.10.2 Other supervised baselines

Classical Fisher discriminant analysis gives an exact finite-sample contraction identity. For class c with index set I_c, size n_c, and centroid μ_c,

$$\sum_{i \in I_c} \|y_i - \mu_c\|^2 = \frac{1}{2n_c} \sum_{i,j \in I_c} \|y_i - y_j\|^2. \tag{70}$$

Thus within-class scatter is exactly a nonnegative pairwise contraction penalty on same-class pairs.

The TP-SE implementation first bins scalar scores into quantile labels, then trains an encoder g_θ, classifier c_θ, and decoder r_θ with the two-phase batch objective

$$\beta\|r_\theta(g_\theta(x_i)) - x_i\|^2 + (1 - \beta)\,\mathrm{CE}\big(c_\theta(g_\theta(x_i)), b(s_i)\big), \tag{71}$$

where the public sweep parameter satisfies $\lambda = 1 - \beta$. The supervised term favors correct bin prediction without requiring same-bin samples to be close. The reconstruction term additionally resists collapse.

RF-PHATE and RF-AE both start from a random-forest proximity learned from (X, s). With leaf index $\ell_t(x_i)$ in tree t, the basic proximity has the form

$$P_{ij} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{1}\{\ell_t(x_i) = \ell_t(x_j)\}. \tag{72}$$

Because the forest is trained to predict the scalar signal, high proximity is target-aware. The retained implementations use RF-GAP-corrected proximities. Equation (72) illustrates the underlying leaf-based mechanism. RF-PHATE embeds a diffusion geometry built from this proximity, and RF-AE trains a network toward diffusion coordinates derived from the same matrix. This introduces target-aware geometry, without directly bounding local signal variation on the final layout.

AKRMap optimizes a parametric t-SNE geometry loss plus a Nadaraya–Watson map loss,

$$L_{\mathrm{AKRMap}}(\theta) = (1 - \lambda) L_{\mathrm{tSNE}}(g_\theta(X)) + \lambda L_{\mathrm{map}}(g_\theta(X), s). \tag{73}$$

The readout ingredient is shared with ISDR, but the training supports are mini-batch dependent (Appendix A.9).

B Implementation Details

All main results use fixed data and five seeds; seeds affect method stochasticity, not sample selection. Main benchmark and direct-start initialization results use population SD; label recovery and the other initialization studies use sample SD.

B.1 Compute Resources

All experiments were run on an internal Linux server with two Intel Xeon Gold 6248R CPUs at 3.00 GHz, 96 logical CPU cores, 503 GiB RAM, and eight NVIDIA GeForce RTX 3090 GPUs with 24 GiB memory each. HPDv3 features and logit margins were cached once using one RTX 3090 and reused across sweeps. DR runs were primarily CPU-based, with GPU acceleration for some baselines. Per-method runtimes are reported in Appendix C.8.

B.2 Existing Assets and Licenses

Table 3 lists the external datasets and pretrained model used as benchmark inputs or feature-generation tools. HPDv3 image panels remain subject to the source dataset terms.

Table 3: Existing assets used in the experiments and their public license or terms.

Asset	Use in this paper	License	Public source
HPDv3 / HPSv3 [59]	HPDv3 prompt-image pair benchmark	MIT	`https://huggingface.co/datasets/MizzenAI/HPDv3`
Qwen3-VL-8B-Instruct	Existing VLM judge/feature extractor for HPDv3 representations and logit margins	Apache-2.0	`https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct`
Tabula Muris FACS [5]	Single-cell benchmark features and dissociation score	CC BY 4.0	`https://doi.org/10.6084/m9.figshare.5715040.v1`
California Housing [69]	Tabular regression benchmark; median house value	Not stated	`https://scikit-learn.org/stable/datasets/real_world.html#california-housing-dataset`
Superconductivity [70]	Tabular regression benchmark; critical temperature	CC BY 4.0	`https://archive.ics.uci.edu/dataset/464/superconductivty+data`
MNIST [71]	Traditional classification diagnostic	MIT	`https://huggingface.co/datasets/ylecun/mnist`
Fashion-MNIST [72]	Traditional classification diagnostic	MIT	`https://huggingface.co/datasets/zalando-datasets/fashion_mnist`
MedMNIST OrganCMNIST and PathMNIST [73]	Traditional classification diagnostics	CC BY 4.0 for the used subsets	`https://medmnist.com/`

B.3 ISDR Implementation

Readout regularity. ISDR uses the Student-t readout and objective in Section 5. Let p be the t-SNE perplexity and let $k = \lfloor 3p \rfloor$. The leave-one-out support set $S_i \subseteq \{1, \ldots, N\} \setminus \{i\}$ of size k

is set to the current-layout k nearest neighbors of i at each rebuild. Between rebuilds, the supports are fixed and L_{Reg} is smooth in Y. The implementation floors kernel sums at 10^{-12} for numerical stability. The theoretical bounds analyze the exact positive-kernel readout.

Initialization and refinement. The default initializer is a geometry-only Barnes–Hut t-SNE layout after shared PCA pre-reduction. The reported HPDv3 and Tabula Muris runs use perplexity 30, a t-SNE warm start with 1000 iterations, and a nonparametric refinement budget of 750 iterations. The refinement learning rate is 100, and the layout kNN support graph is rebuilt every 50 iterations. The released implementation makes the regularization warmup and t-SNE-style optimization schedules configurable. The regularization coefficient is linearly warmed up over 50 steps: $\lambda_t = \lambda \min(t/50, 1)$ for $t = 0, \ldots, 749$. The geometry force uses $12\times$ early exaggeration for the first 250 refinement steps, followed by ordinary t-SNE affinities; momentum is 0.5 during these 250 steps and 0.8 thereafter, with t-SNE adaptive coordinate gains. The last 500 steps therefore use the target objective in Section 5.

B.4 Time and Space Complexity

Let N be the number of visualized points, D' the PCA-reduced dimension, T the number of refinement steps, R the support-rebuild interval, and k the NW regularizer support size. For one ISDR operating point, using the usual Barnes–Hut approximation and expected $O(N \log N + Nk)$ queries for a spatial index in the two-dimensional layout, the cost is

$$C_{\mathrm{PCA}} + C_P + C_{\mathrm{init}} + T\left[O(N \log N) + O(Nk)\right] + \left\lceil \frac{T}{R} \right\rceil O(N \log N + Nk). \tag{74}$$

Here C_{PCA} is PCA preprocessing, C_P is construction of the high-dimensional sparse t-SNE affinity matrix, and C_{init} is the warm-start layout cost. The per-step terms are the Barnes–Hut force and fixed-support NW gradient. The final term accounts for support rebuilds. The preprocessing, affinity, and initialization costs can be shared across a weight sweep, with refinement repeated for each weight. The memory footprint after preprocessing is $O(ND')$ for working features, $O(N(q+k))$ for sparse affinities of degree q and readout supports, and $O(N)$ for scores and layout arrays, aside from method-specific buffers. With the default $q, k \approx 3p$, the sparse storage is $O(Nk)$.

B.5 Metrics

The main Pareto plane uses trustworthiness for geometry preservation and signal Dirichlet energy for scalar-field readability. Neighborhood metrics in the main sweeps use $k = 10$. Trustworthiness is

$$T(k) = 1 - \frac{2}{Nk(2N - 3k - 1)} \sum_i \sum_{j \in U_k(i)} (r_X(i,j) - k), \tag{75}$$

where $U_k(i)$ are layout neighbors of i that are not high-dimensional neighbors, and $r_X(i,j)$ is the high-dimensional rank of j from i. Continuity swaps the roles of high-dimensional and layout neighborhoods. The 10-NN overlap is the average fraction of shared high-dimensional and layout neighbors. Signal Dirichlet energy is computed on the directed layout kNN graph $E_k^{\rightarrow}(Y) = \{(i,j) : j \in \mathrm{kNN}_Y(i)\}$ (with $|E_k^{\rightarrow}(Y)| = Nk$) using min–max normalized scores $\tilde{s}_i = (s_i - s_{\min})/(s_{\max} - s_{\min})$:

$$\mathrm{DE}(Y, s; k) = \frac{1}{|E_k^{\rightarrow}(Y)|} \sum_{(i,j) \in E_k^{\rightarrow}(Y)} (\tilde{s}_i - \tilde{s}_j)^2. \tag{76}$$

Lower DE means smoother scalar variation across local edges of the realized map. Embedding tables annotate $100 \times$ DE and abbreviate trustworthiness as Tr. For constant signals, DE is defined as zero.

B.6 Method Settings

Input preprocessing protocol. All HPDv3 methods receive the same standardized 4096-dimensional layer-19 representation, which is also the reference for trustworthiness. PCA is applied within each method as specified in Table 4. It is skipped when the requested dimension is at least the input dimension, as for the 50-dimensional Tabula Muris features. For ISDR, scores are z-scored only inside the NW regularizer. Reported scalar-signal metrics use the original scores, with DE using the min–max normalization defined above. At a fixed layout, the exact NW residuals satisfy $L_{\mathrm{Reg}}^{\mathrm{raw}} = \max\{\sigma_s, 10^{-12}\}^2 L_{\mathrm{Reg}}^{z}$, matching the implementation's score-normalization scale, where σ_s is the score standard deviation. DE is invariant under this affine score scaling.

Table 4: Standard method settings used for HPDv3 and Tabula Muris. The standard grid for S-UMAP, TP-SE, AKRMap, and ISDR is $\lambda \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$.

Method	Sweep parameter	Fixed settings
PCA	none	Linear PCA to two dimensions.
t-SNE	none	PCA-50 preprocessing; perplexity 30; 1000 iterations; PCA initialization; Barnes–Hut implementation.
UMAP	none	PCA-50 preprocessing; 15 neighbors; minimum distance 0.1.
PaCMAP	none	PCA-50 preprocessing; 10 neighbors; library defaults otherwise.
S-UMAP	target weight λ	Same UMAP settings; scalar target metric for continuous labels.
CEBRA	fixed point	Temperature 1.0; output dimension 2; batch size 512; learning rate 3×10^{-4}; 2000 iterations; PCA-100 on HPDv3 and no additional reduction on the 50D Tabula Muris cache.
TP-SE	supervised weight λ	PCA-128 on HPDv3 and no additional reduction on the 50D Tabula Muris cache; latent dimension 2; 40 epochs; batch size 512; learning rate 10^{-3}; scores discretized into 10 quantile bins for the classifier branch.
RF-PHATE	fixed point	Random forest on standardized features without external PCA; 500 trees; 2000 PHATE landmarks; internal $n_{\mathrm{pca}} = 100$; PageRank teleportation $\beta = 0.9$.
RF-AE	fixed point	RF-PHATE transition rows as autoencoder inputs; random forest on standardized features without external PCA; 300 trees; 2000 landmarks; RF-PHATE internal $n_{\mathrm{pca}} = 100$; 50 epochs; batch size 256; learning rate 10^{-3}; reconstruction weight 10^{-2}.
AKRMap	map-loss weight λ	Parametric t-SNE plus NW map loss; 20 epochs; batch size 1000; learning rate 0.002; PCA-1000 on HPDv3 and no additional reduction on the 50D Tabula Muris cache.
ISDR	NW regularizer weight λ	PCA-50; geometry-only t-SNE initialization; support $k = 90$; initialization and refinement schedule in Appendix B.3.

C Full Results and Ablations

This appendix reports the full HPDv3 and Tabula Muris results, label recovery, supplementary layouts, runtime measurements, ablations, and a comparison with linear supervised DR.

C.1 Dataset Configurations

Table 5: Main benchmark configurations. Both are fixed-data visualization benchmarks. Repeated seeds affect method stochasticity but not sample selection.

Dataset	Configuration
HPDv3	$N = 20{,}000$ fixed visualization subset; Qwen3-VL-8B-Instruct layer-19 hidden state with $D = 4096$; signal is the top-two logit gap among the A/B/C judge options; seeds are $\{42, 123, 456, 789, 2026\}$.
Tabula Muris	$N = 44{,}949$ full FACS atlas; representation is 50-dimensional PCA of log-normalized HVG expression; scalar signal is the globally z-scored mean log-normalized expression over the matched dissociation-associated genes; same five seeds.

Table 5 summarizes the two main benchmarks. Each HPDv3 sample comprises a text prompt and two candidate images. The judge prompt and feature construction are described in Appendix D.1. The standard supervision grid is $\lambda \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ for S-UMAP, TP-SE, AKRMap, and ISDR. PCA, t-SNE, UMAP, PaCMAP, CEBRA, RF-PHATE, and RF-AE are single-point baselines in this protocol. The method-specific settings are listed in Table 4.

C.2 Tabula Muris Preprocessing Details

We align cell annotations and metadata from the figshare FACS release [5], deduplicate by cell, and align genes across the per-tissue count matrices. All 44,949 QC-passed cells are retained without a train/test split, so the same fixed sample is reused across seeds.

High-dimensional representation X. Per-cell raw counts are library-size normalized to counts-per-10K (CP10K) and then log-transformed via $\log(1 + x)$. We select the top 2,000 genes by raw variance of this log-normalized matrix and apply randomized PCA to 50 components. The resulting $44{,}949 \times 50$ matrix is the input to all dimensionality-reduction methods.

Scalar signal: dissociation score. We use the 140-gene list of dissociation-associated genes published by van den Brink et al. [63] and intersect it with the FACS gene index. The 138 matching genes are retained for the score. Within this matched-gene submatrix, raw counts are library-size normalized to counts-per-million (CPM) using each cell's full library total and then log-transformed via $\log(1 + x)$. The scalar signal is the per-cell mean of these matched-gene values, followed by a global z-score across all 44,949 cells. The signal is continuous, with no binning, clipping, or per-tissue normalization.

C.3 Representative Operating-Point Selection

Table 6 reports an anchor-inclusive compact summary under the same selection rule used for the main-text representative points. For swept methods, we retain the non-dominated points in the trustworthiness–DE plane, normalize both axes to $[0, 1]$, and select the Pareto knee, defined as the front point farthest from the segment joining the highest-trust and lowest-DE endpoints. Ties are broken by higher trustworthiness and then lower DE. If the front contains only one point or the endpoints coincide, that point is selected using the same tie rule. Fixed baselines use their single point.

Table 6: Anchor-inclusive representative operating-point summary over five seeds. Fixed baselines use their single point, and swept methods use the same Pareto-knee rule as Table 1. The full-grid tables below report every standard-grid point.

Dataset	Method	λ	Trust ↑	Cont. ↑	10-NN ↑	$10^2 \times$DE ↓
HPDv3	t-SNE	–	0.987±0.000	0.978±0.000	0.323±0.000	7.244±0.008
	UMAP	–	0.947±0.001	0.982±0.000	0.096±0.001	7.970±0.012
	S-UMAP	0.4	0.925±0.041	0.967±0.031	0.082±0.026	7.526±0.837
	AKRMap	0.2	0.863±0.004	0.952±0.001	0.019±0.001	2.724±0.116
	ISDR	0.4	0.977±0.000	0.978±0.000	0.252±0.001	2.108±0.008
Tabula Muris	t-SNE	–	0.998±0.000	0.996±0.000	0.380±0.000	0.672±0.003
	UMAP	–	0.992±0.000	0.995±0.000	0.179±0.001	0.955±0.012
	S-UMAP	0.4	0.992±0.000	0.996±0.000	0.171±0.001	0.941±0.005
	AKRMap	0.8	0.945±0.004	0.990±0.001	0.054±0.002	0.380±0.019
	ISDR	0.8	0.996±0.000	0.995±0.000	0.319±0.001	0.227±0.000

C.4 HPDv3 Full Standard Grid

Table 7 (HPDv3 panel) reports every method and every standard-grid point from the fixed-data HPDv3 five-seed sweep, providing the numerical values behind the HPDv3 panel of Figure 2. The theory-proxy columns use the finite-sample definitions from Section 3: Φ_σ is estimated on a fixed sample of 10^6 unordered pairs drawn uniformly with replacement, and the reported M_{ctr} is computed on the score-difference band $[0, 0.01\sigma_s]$ from the same pair sample. Here $M_{\mathrm{ctr}} < 1$ indicates average compression relative to the global mean expansion estimated from that sample. M_{sv} uses $k = 10$ and $\tau = 1$. Input distances at most 10^{-12} are excluded from expansion ratios.

C.5 Tabula Muris Full Standard Grid

Figure 4 visualizes the Tabula Muris supervision-weight paths for the swept supervised methods and ISDR, using the same single-seed embedding-table convention as Figure 1. Figure 5 complements it with the fixed unsupervised and single-point supervised baselines, following the HPDv3 supplementary method grouping in Figure 6.

Table 7 (Tabula Muris panel) reports every method and every standard-grid point from the full Tabula Muris FACS five-seed sweep, providing the numerical values behind the Tabula Muris panel of Figure 2. Theory-proxy definitions and sampling settings follow Appendix C.4.

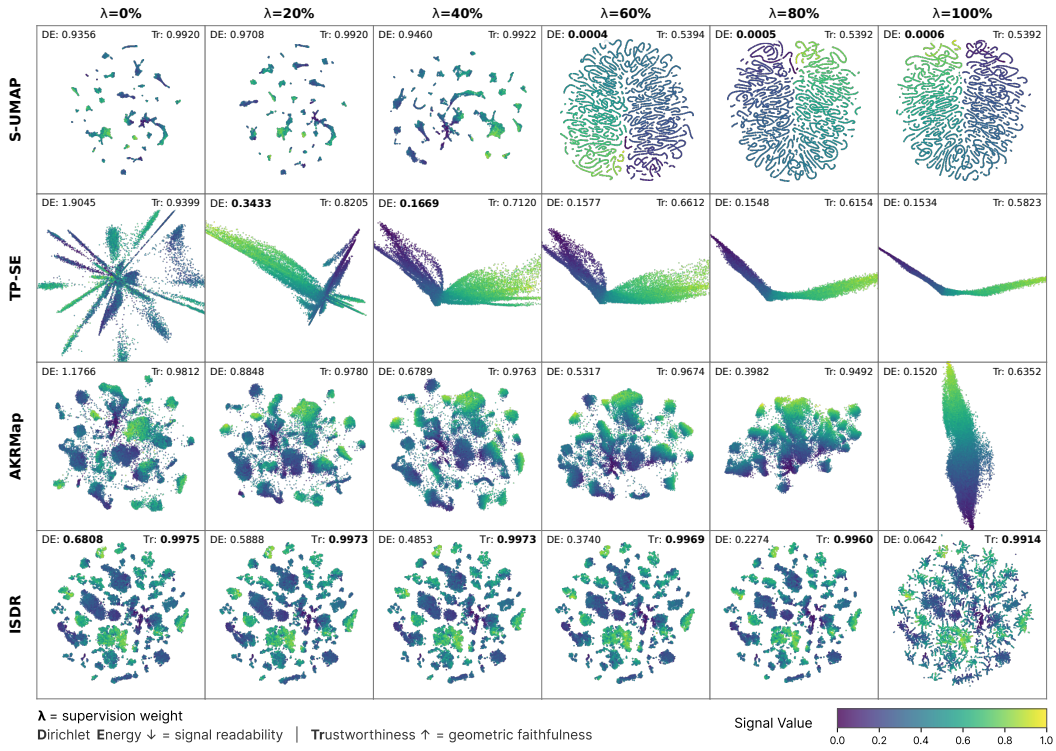

Figure 4: Tabula Muris embeddings for supervised sweep methods and ISDR. Rows show S-UMAP, TP-SE, AKRMap, and ISDR. Columns sweep the standard supervision grid $\lambda \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. Points are colored by the normalized dissociation score.

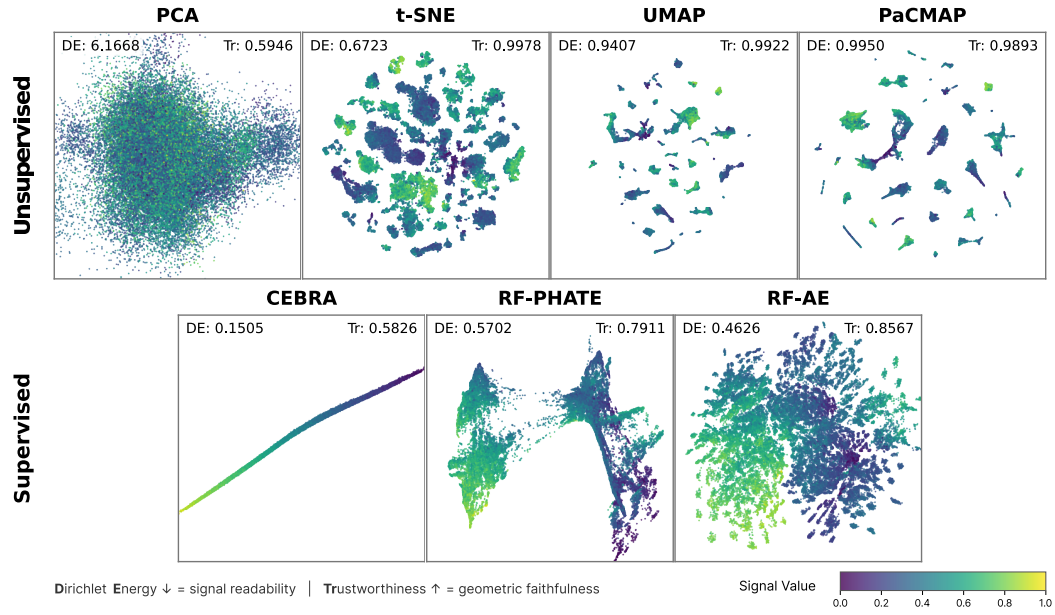

Figure 5: Tabula Muris fixed-baseline embeddings on seed 42. The first row shows PCA, t-SNE, UMAP, and PaCMAP. The second row shows CEBRA, RF-PHATE, and RF-AE. Colors show the normalized dissociation score.

Table 7: Full standard-grid results on the main benchmarks (HPDv3 and Tabula Muris). Values are mean ± standard deviation over five fixed-data seeds. Methods are grouped by row; swept methods report all six standard-grid points $\lambda \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$.

Method	λ	Trust ↑	Cont. ↑	10-NN ↑	$10^2 \times \mathrm{DE}$ ↓	Φ_σ ↓	M_{sv}	M_{ctr}
Dataset: HPDv3								
PCA	–	0.842 ± 0.000	0.960 ± 0.000	0.018 ± 0.000	8.523 ± 0.000	0.168 ± 0.000	35.948 ± 0.000	0.976 ± 0.000
t-SNE	–	0.987 ± 0.000	0.978 ± 0.000	0.323 ± 0.000	7.244 ± 0.008	0.200 ± 0.000	29.539 ± 0.045	0.994 ± 0.000
UMAP	–	0.947 ± 0.001	0.982 ± 0.000	0.096 ± 0.001	7.970 ± 0.012	0.366 ± 0.028	31.340 ± 0.155	0.970 ± 0.001
PaCMAP	–	0.936 ± 0.001	0.981 ± 0.000	0.091 ± 0.000	7.954 ± 0.016	0.263 ± 0.001	30.045 ± 0.052	0.974 ± 0.001
CEBRA	–	0.564 ± 0.013	0.667 ± 0.008	0.002 ± 0.000	3.773 ± 0.064	0.384 ± 0.008	15.148 ± 0.331	0.734 ± 0.006
RF-PHATE	–	0.827 ± 0.001	0.816 ± 0.001	0.022 ± 0.000	6.790 ± 0.018	0.626 ± 0.007	30.059 ± 0.168	0.947 ± 0.001
RF-AE	–	0.869 ± 0.001	0.815 ± 0.008	0.028 ± 0.000	6.525 ± 0.046	0.389 ± 0.010	26.797 ± 0.247	0.956 ± 0.010
S-UMAP	0.0	0.946 ± 0.000	0.982 ± 0.000	0.096 ± 0.001	7.965 ± 0.020	0.365 ± 0.020	31.238 ± 0.306	0.972 ± 0.003
	0.2	0.946 ± 0.000	0.982 ± 0.000	0.096 ± 0.001	7.961 ± 0.021	0.368 ± 0.025	31.400 ± 0.173	0.972 ± 0.001
	0.4	0.925 ± 0.041	0.967 ± 0.031	0.082 ± 0.026	7.526 ± 0.837	0.310 ± 0.073	29.936 ± 2.916	0.961 ± 0.033
	0.6	0.531 ± 0.001	0.499 ± 0.001	0.001 ± 0.000	0.049 ± 0.013	0.285 ± 0.000	0.000 ± 0.000	0.720 ± 0.015
	0.8	0.531 ± 0.001	0.499 ± 0.000	0.001 ± 0.000	0.053 ± 0.010	0.285 ± 0.001	0.000 ± 0.000	0.726 ± 0.006
	1.0	0.531 ± 0.000	0.498 ± 0.001	0.001 ± 0.000	0.034 ± 0.008	0.285 ± 0.000	0.000 ± 0.000	0.734 ± 0.024
TP-SE	0.0	0.919 ± 0.005	0.960 ± 0.005	0.037 ± 0.002	8.215 ± 0.051	0.206 ± 0.014	33.339 ± 1.755	0.982 ± 0.011
	0.2	0.920 ± 0.004	0.959 ± 0.007	0.037 ± 0.002	8.208 ± 0.078	0.213 ± 0.018	33.148 ± 1.461	0.982 ± 0.013
	0.4	0.922 ± 0.003	0.963 ± 0.005	0.038 ± 0.002	8.170 ± 0.065	0.207 ± 0.014	32.496 ± 1.049	0.974 ± 0.010
	0.6	0.923 ± 0.003	0.964 ± 0.003	0.038 ± 0.002	8.134 ± 0.046	0.223 ± 0.017	31.676 ± 1.892	0.963 ± 0.011
	0.8	0.922 ± 0.004	0.964 ± 0.005	0.038 ± 0.002	7.890 ± 0.109	0.247 ± 0.050	30.724 ± 1.541	0.954 ± 0.015
	1.0	0.563 ± 0.002	0.661 ± 0.004	0.004 ± 0.000	2.262 ± 0.053	0.642 ± 0.046	12.934 ± 0.466	0.681 ± 0.033
AKRMap	0.0	0.940 ± 0.002	0.982 ± 0.001	0.069 ± 0.001	7.994 ± 0.019	0.225 ± 0.011	31.433 ± 0.427	0.973 ± 0.005
	0.2	0.863 ± 0.004	0.952 ± 0.001	0.019 ± 0.001	2.724 ± 0.116	0.233 ± 0.027	11.136 ± 0.405	0.940 ± 0.008
	0.4	0.806 ± 0.004	0.916 ± 0.003	0.011 ± 0.000	2.682 ± 0.215	0.310 ± 0.025	11.442 ± 0.857	0.935 ± 0.009
	0.6	0.731 ± 0.016	0.843 ± 0.008	0.007 ± 0.000	3.811 ± 0.154	0.226 ± 0.009	15.855 ± 0.542	0.867 ± 0.007
	0.8	0.657 ± 0.035	0.760 ± 0.014	0.005 ± 0.000	4.321 ± 0.224	0.317 ± 0.011	16.894 ± 1.094	0.786 ± 0.017
	1.0	0.648 ± 0.036	0.712 ± 0.011	0.005 ± 0.001	4.637 ± 0.234	0.435 ± 0.017	18.352 ± 1.333	0.771 ± 0.017
ISDR	0.0	0.985 ± 0.000	0.978 ± 0.000	0.319 ± 0.001	7.286 ± 0.029	0.204 ± 0.000	29.729 ± 0.110	0.989 ± 0.000
	0.2	0.983 ± 0.000	0.978 ± 0.000	0.302 ± 0.000	4.433 ± 0.014	0.205 ± 0.000	18.804 ± 0.062	0.989 ± 0.000
	0.4	0.977 ± 0.000	0.978 ± 0.000	0.252 ± 0.001	2.108 ± 0.008	0.208 ± 0.000	8.842 ± 0.061	0.989 ± 0.001
	0.6	0.971 ± 0.000	0.978 ± 0.000	0.204 ± 0.001	1.087 ± 0.007	0.209 ± 0.000	4.543 ± 0.020	0.988 ± 0.001
	0.8	0.966 ± 0.001	0.976 ± 0.000	0.161 ± 0.001	0.588 ± 0.007	0.206 ± 0.000	2.298 ± 0.024	0.990 ± 0.001
	1.0	0.954 ± 0.000	0.974 ± 0.000	0.077 ± 0.000	0.331 ± 0.006	0.201 ± 0.000	0.980 ± 0.022	0.993 ± 0.000
Dataset: Tabula Muris								
PCA	–	0.595 ± 0.000	0.891 ± 0.000	0.007 ± 0.000	6.167 ± 0.000	0.336 ± 0.000	2.036 ± 0.000	1.047 ± 0.000
t-SNE	–	0.998 ± 0.000	0.996 ± 0.000	0.380 ± 0.000	0.672 ± 0.003	0.217 ± 0.014	0.181 ± 0.001	0.986 ± 0.027
UMAP	–	0.992 ± 0.000	0.995 ± 0.000	0.179 ± 0.001	0.955 ± 0.012	0.249 ± 0.004	0.260 ± 0.002	0.978 ± 0.028
PaCMAP	–	0.989 ± 0.000	0.994 ± 0.000	0.158 ± 0.001	1.022 ± 0.014	0.199 ± 0.002	0.265 ± 0.009	0.941 ± 0.004
CEBRA	–	0.586 ± 0.002	0.848 ± 0.000	0.005 ± 0.000	0.150 ± 0.001	0.544 ± 0.017	0.045 ± 0.000	0.167 ± 0.003
RF-PHATE	–	0.791 ± 0.002	0.936 ± 0.001	0.038 ± 0.000	0.567 ± 0.014	0.326 ± 0.004	0.155 ± 0.015	0.689 ± 0.004
RF-AE	–	0.860 ± 0.004	0.944 ± 0.002	0.047 ± 0.000	0.468 ± 0.022	0.294 ± 0.004	0.138 ± 0.009	0.810 ± 0.007
S-UMAP	0.0	0.992 ± 0.000	0.996 ± 0.000	0.172 ± 0.001	0.945 ± 0.007	0.255 ± 0.011	0.256 ± 0.004	0.961 ± 0.022
	0.2	0.992 ± 0.000	0.996 ± 0.000	0.172 ± 0.001	0.969 ± 0.005	0.258 ± 0.004	0.261 ± 0.002	0.984 ± 0.028
	0.4	0.992 ± 0.000	0.996 ± 0.000	0.171 ± 0.001	0.941 ± 0.005	0.281 ± 0.007	0.251 ± 0.002	0.737 ± 0.005
	0.6	0.539 ± 0.000	0.783 ± 0.000	0.002 ± 0.000	0.001 ± 0.000	0.258 ± 0.001	0.000 ± 0.000	0.198 ± 0.004
	0.8	0.539 ± 0.000	0.783 ± 0.000	0.002 ± 0.000	0.001 ± 0.000	0.257 ± 0.000	0.000 ± 0.000	0.194 ± 0.007
	1.0	0.539 ± 0.000	0.783 ± 0.001	0.001 ± 0.000	0.001 ± 0.000	0.257 ± 0.000	0.000 ± 0.000	0.200 ± 0.004
TP-SE	0.0	0.944 ± 0.003	0.977 ± 0.002	0.028 ± 0.001	1.910 ± 0.016	0.682 ± 0.039	0.389 ± 0.068	0.984 ± 0.065
	0.2	0.813 ± 0.029	0.922 ± 0.018	0.022 ± 0.003	0.301 ± 0.039	0.648 ± 0.091	0.081 ± 0.021	0.450 ± 0.075
	0.4	0.699 ± 0.007	0.876 ± 0.009	0.013 ± 0.000	0.167 ± 0.004	0.697 ± 0.089	0.049 ± 0.001	0.309 ± 0.034
	0.6	0.650 ± 0.012	0.861 ± 0.007	0.009 ± 0.001	0.161 ± 0.004	0.660 ± 0.109	0.047 ± 0.000	0.260 ± 0.030
	0.8	0.606 ± 0.006	0.853 ± 0.008	0.007 ± 0.001	0.157 ± 0.003	0.682 ± 0.097	0.046 ± 0.001	0.214 ± 0.018
	1.0	0.580 ± 0.002	0.846 ± 0.008	0.005 ± 0.000	0.156 ± 0.003	0.703 ± 0.097	0.045 ± 0.001	0.194 ± 0.013
AKRMap	0.0	0.979 ± 0.001	0.994 ± 0.000	0.091 ± 0.001	1.206 ± 0.023	0.232 ± 0.007	0.285 ± 0.003	0.962 ± 0.019
	0.2	0.977 ± 0.001	0.994 ± 0.000	0.089 ± 0.001	0.909 ± 0.014	0.233 ± 0.007	0.210 ± 0.007	0.943 ± 0.013
	0.4	0.974 ± 0.002	0.994 ± 0.000	0.081 ± 0.002	0.726 ± 0.029	0.234 ± 0.006	0.160 ± 0.004	0.927 ± 0.023
	0.6	0.965 ± 0.003	0.992 ± 0.002	0.069 ± 0.003	0.530 ± 0.037	0.241 ± 0.006	0.112 ± 0.004	0.873 ± 0.024
	0.8	0.945 ± 0.004	0.990 ± 0.001	0.054 ± 0.002	0.380 ± 0.019	0.274 ± 0.017	0.083 ± 0.004	0.852 ± 0.010
	1.0	0.641 ± 0.005	0.894 ± 0.003	0.010 ± 0.001	0.149 ± 0.003	0.460 ± 0.007	0.045 ± 0.001	0.299 ± 0.010

Continued on next page.

Table 7 (continued from previous page)

Method	λ	Trust ↑	Cont. ↑	10-NN ↑	$10^2 \times \mathrm{DE}$ ↓	Φ_σ ↓	M_{sv}	M_{ctr}
ISDR	0.0	0.997 ± 0.000	0.995 ± 0.000	0.376 ± 0.000	0.680 ± 0.002	0.218 ± 0.014	0.182 ± 0.000	0.985 ± 0.027
	0.2	0.997 ± 0.000	0.995 ± 0.000	0.376 ± 0.000	0.587 ± 0.003	0.217 ± 0.014	0.163 ± 0.001	0.985 ± 0.027
	0.4	0.997 ± 0.000	0.995 ± 0.000	0.371 ± 0.000	0.490 ± 0.002	0.218 ± 0.014	0.139 ± 0.001	0.985 ± 0.027
	0.6	0.997 ± 0.000	0.995 ± 0.000	0.358 ± 0.000	0.375 ± 0.002	0.218 ± 0.014	0.110 ± 0.000	0.985 ± 0.028
	0.8	0.996 ± 0.000	0.995 ± 0.000	0.319 ± 0.001	0.227 ± 0.000	0.218 ± 0.014	0.067 ± 0.000	0.984 ± 0.027
	1.0	0.992 ± 0.000	0.994 ± 0.000	0.120 ± 0.000	0.064 ± 0.000	0.217 ± 0.014	0.014 ± 0.000	0.985 ± 0.027

C.6 Additional HPDv3 Layouts

Figure 6 complements Figure 1 with the remaining baselines.

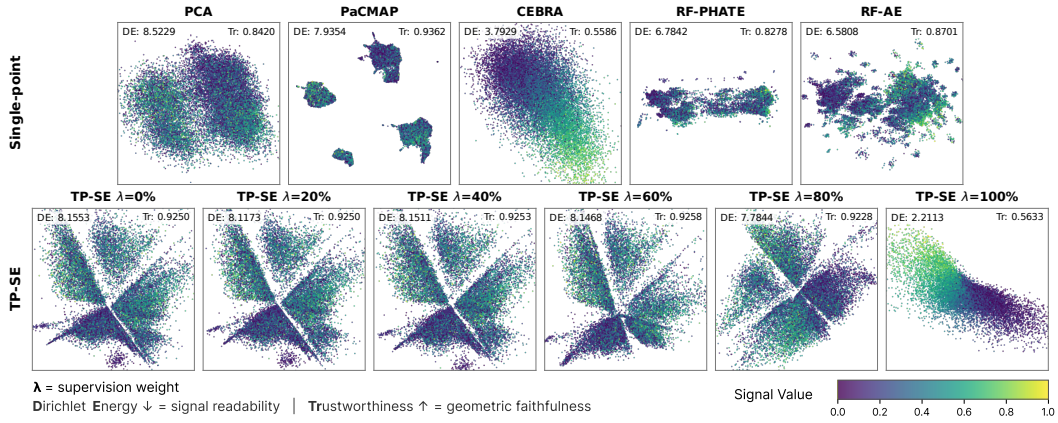

Figure 6: Additional HPDv3 embeddings: fixed baselines (top row) and TP-SE across the standard supervision grid (bottom row). Points are colored by the normalized scalar signal.

C.7 Label Recovery

Label recovery provides a supplementary reference using labels excluded from layout fitting and operating-point selection. On each layout fitted to all inputs, a uniform-weight Euclidean 10-NN classifier predicts held-out labels using the same stratified ten-fold split for every method and seed. We average Accuracy and Macro-F1 over folds within each layout, then report their mean and sample standard deviation across seeds. Operating points retain the Trust–DE selection rule in Appendix C.3. Table 8 reports the fixed methods, and Tables 9 and 10 report every standard-grid weight.

Cell types on Tabula Muris. The cell-type annotations provide an external reference distinct from the dissociation score used for supervision. Of the 44,949 cells in each layout, 44,807 have annotations covering 81 classes; the 142 cells with missing annotations are omitted from label recovery. At $\lambda = 0.8$, ISDR attains 91.70% Accuracy and 80.02% Macro-F1, compared with t-SNE's 92.18% and 80.96%, while reducing DE by 66%. S-UMAP's recovery drops sharply at weights 0.6–1.0 as DE approaches zero (Table 7).

Judge choices on HPDv3. The labels are the judge's three-way decisions in the presented image order: first image, second image, or tie (7,779, 11,838, and 383 samples, respectively). These model-derived labels come from the same judge as the confidence signal and serve as an auxiliary reference. At $\lambda = 0.4$, ISDR attains 60.92% Accuracy and 50.68% Macro-F1, close to t-SNE's 61.44% and 51.13%. Macro-F1 complements Accuracy for this imbalanced label distribution. The full ISDR path in Table 10 shows lower recovery at the strongest supervision weights.

C.8 Measured Runtime on HPDv3

Table 11 reports HPDv3 wall-clock times under the main experimental settings. Times cover method calls, including preprocessing, initialization, optimization, and any setup within the call. Feature

Table 8: Label recovery for fixed methods (%). PCA results are identical across seeds.

Method	Tabula Muris: cell type		HPDv3: judge choice	
	Accuracy ↑	Macro-F1 ↑	Accuracy ↑	Macro-F1 ↑
t-SNE	92.18 ± 0.06	80.96 ± 0.19	61.44 ± 0.06	51.13 ± 0.43
UMAP	90.15 ± 0.10	74.91 ± 0.26	56.36 ± 0.26	44.11 ± 0.20
PaCMAP	89.28 ± 0.18	73.79 ± 0.80	56.15 ± 0.25	43.78 ± 0.26
PCA	19.16	10.51	53.44	34.19
RF-AE	69.23 ± 0.46	44.05 ± 0.22	56.31 ± 0.56	37.55 ± 0.72
RF-PHATE	61.61 ± 0.20	37.48 ± 0.46	56.17 ± 0.17	45.59 ± 0.74
CEBRA	23.35 ± 1.92	9.25 ± 0.46	53.74 ± 0.53	34.15 ± 0.39
Majority class	9.81	0.22	59.19	24.79

Table 9: Tabula Muris label recovery across supervision weights (%).

Method	$\lambda = 0$	$\lambda = 0.2$	$\lambda = 0.4$	$\lambda = 0.6$	$\lambda = 0.8$	$\lambda = 1.0$
			Accuracy ↑			
ISDR	92.01 ± 0.07	92.06 ± 0.04	92.02 ± 0.06	91.95 ± 0.03	91.70 ± 0.06	89.39 ± 0.10
S-UMAP	90.01 ± 0.08	89.92 ± 0.15	89.98 ± 0.06	12.51 ± 0.07	12.47 ± 0.02	12.30 ± 0.10
AKRMap	86.38 ± 0.15	86.08 ± 0.29	85.38 ± 0.44	83.81 ± 0.70	80.66 ± 0.82	35.95 ± 1.72
TP-SE	76.02 ± 0.61	59.91 ± 5.92	43.37 ± 1.31	33.94 ± 2.85	25.48 ± 0.98	19.43 ± 0.61
			Macro-F1 ↑			
ISDR	80.73 ± 0.22	80.90 ± 0.18	80.92 ± 0.35	80.83 ± 0.15	80.02 ± 0.30	73.35 ± 0.25
S-UMAP	74.46 ± 0.15	74.18 ± 0.32	74.53 ± 0.18	3.95 ± 0.04	3.95 ± 0.09	3.78 ± 0.07
AKRMap	65.63 ± 0.33	65.38 ± 1.19	63.95 ± 0.77	60.04 ± 1.26	54.60 ± 0.96	15.89 ± 1.02
TP-SE	45.30 ± 1.29	33.57 ± 4.18	20.73 ± 1.42	15.40 ± 1.75	11.33 ± 1.18	8.48 ± 0.58

Table 10: HPDv3 label recovery across supervision weights (%).

Method	$\lambda = 0$	$\lambda = 0.2$	$\lambda = 0.4$	$\lambda = 0.6$	$\lambda = 0.8$	$\lambda = 1.0$
			Accuracy ↑			
ISDR	61.26 ± 0.25	61.41 ± 0.20	60.92 ± 0.22	60.32 ± 0.27	59.45 ± 0.48	58.23 ± 0.44
S-UMAP	56.58 ± 0.29	56.48 ± 0.25	56.12 ± 0.83	54.47 ± 0.09	54.34 ± 0.46	54.42 ± 0.49
AKRMap	55.90 ± 0.39	55.99 ± 0.58	56.06 ± 0.81	55.20 ± 1.63	55.50 ± 0.70	56.70 ± 2.54
TP-SE	54.69 ± 0.57	54.90 ± 0.24	55.23 ± 0.48	54.78 ± 0.31	55.10 ± 0.20	63.11 ± 0.43
			Macro-F1 ↑			
ISDR	51.31 ± 0.40	51.70 ± 0.34	50.68 ± 0.68	51.22 ± 0.53	49.92 ± 0.74	48.27 ± 0.22
S-UMAP	43.83 ± 0.46	43.95 ± 0.58	43.50 ± 0.69	34.72 ± 0.14	34.58 ± 0.31	34.74 ± 0.37
AKRMap	42.19 ± 1.08	44.96 ± 2.35	52.70 ± 1.44	50.54 ± 1.50	49.46 ± 1.64	48.72 ± 4.06
TP-SE	35.18 ± 0.58	35.42 ± 0.39	35.69 ± 0.60	35.12 ± 0.24	35.50 ± 0.21	47.00 ± 1.71

extraction, metric evaluation, and plotting are excluded. For UMAP and S-UMAP, the fixed random seeds make the UMAP implementation set `n_jobs=1` and disable parallel layout optimization. For ISDR at $\lambda = 0.4$, the 750-step refinement loop takes 107.66 ± 1.84 s of the 164.00 ± 5.95 s total, averaging 0.144 s per refinement step.

C.9 Initialization and Support Ablations

Effect of geometry-first initialization. On HPDv3, we assess the benefit of the geometry-only stage by comparing a t-SNE start with PCA, Gaussian random, and UMAP coordinates passed directly to refinement. All four use the main PCA-50 preprocessing and 750 refinement steps with the settings in Appendix B.3. Table 12(a) shows that the t-SNE start yields the highest trustworthiness and lowest DE. Figure 7 extends the comparison to four supervision weights on seed 42.

Sensitivity to the t-SNE initializer. To assess sensitivity to raw starting coordinates, we vary the two-dimensional initialization before the geometry-only stage. With PCA-50 input features, PCA, Gaussian random, or UMAP coordinates are rescaled to the t-SNE initialization scale, followed by 1000 geometry-only t-SNE iterations and 750 refinement iterations at $\lambda = 0.4$. Refinement uses a fixed weight from iteration zero, no early exaggeration, and momentum 0.8. All three initializers use

Table 11: HPDv3 runtime profile at fixed $N = 20{,}000$. Sweepable methods are shown at $\lambda = 0.4$; single-point baselines are shown at their fixed standard setting. Times are wall-clock seconds, reported as mean $\pm$ standard deviation over five fixed-data seeds.

Method	Parameter	Time (s)
PCA	fixed	0.59 ± 0.04
t-SNE	fixed	37.24 ± 8.41
UMAP	fixed	51.30 ± 21.78
PaCMAP	fixed	10.88 ± 1.91
S-UMAP	$\lambda = 0.4$	38.23 ± 5.66
CEBRA	fixed	14.75 ± 1.14
RF-PHATE	fixed	2938.61 ± 69.91
TP-SE	$\lambda = 0.4$	15.96 ± 2.86
RF-AE	fixed	1826.62 ± 84.20
AKRMap	$\lambda = 0.4$	64.48 ± 8.04
ISDR	$\lambda = 0.4$	164.00 ± 5.95

Table 12: Full initialization ablation results on HPDv3. Values are mean $\pm$ SD over five seeds: population SD in (a), sample SD in (b,c). Trust denotes trustworthiness. The standard ISDR refinement schedule in (a,c) differs from the fixed-weight schedule in (b).

(a) Choice of refinement start. Standard ISDR schedule at $\lambda = 0.4$.

Refinement start	Trust ↑	$100 \times$ DE ↓
t-SNE	$\mathbf{0.9775 \pm 0.0003}$	$\mathbf{2.11 \pm 0.02}$
PCA	0.9620 ± 0.0003	5.42 ± 0.05
Random	0.9625 ± 0.0005	5.34 ± 0.05
UMAP	0.9649 ± 0.0002	4.87 ± 0.11

(b) Sensitivity to the raw initializer. Fixed-weight refinement at $\lambda = 0.4$.

	After geometry-only t-SNE		After ISDR refinement	
Raw initializer	Trust ↑	$100 \times$ DE ↓	Trust ↑	$100 \times$ DE ↓
PCA	0.9869 ± 0.0001	7.243 ± 0.005	0.9816 ± 0.0002	2.598 ± 0.015
Random	0.9869 ± 0.0002	7.253 ± 0.009	0.9816 ± 0.0003	2.591 ± 0.017
UMAP	0.9868 ± 0.0001	7.240 ± 0.015	0.9816 ± 0.0002	2.597 ± 0.028

(c) Refinement from identical t-SNE coordinates. Weights are method-specific.

	ISDR		S-UMAP	
Weight	Trust ↑	$100 \times$ DE ↓	Trust ↑	$100 \times$ DE ↓
0.0	0.9854 ± 0.0002	7.281 ± 0.017	0.9484 ± 0.0018	7.929 ± 0.020
0.2	0.9831 ± 0.0001	4.445 ± 0.019	0.9477 ± 0.0014	7.935 ± 0.009
0.4	0.9776 ± 0.0003	2.111 ± 0.007	0.9290 ± 0.0402	7.532 ± 0.880
0.6	0.9711 ± 0.0006	1.088 ± 0.005	0.6401 ± 0.0010	1.714 ± 0.034
0.8	0.9659 ± 0.0002	0.590 ± 0.007	0.6402 ± 0.0012	1.718 ± 0.028
1.0	0.9539 ± 0.0005	0.328 ± 0.004	0.6392 ± 0.0010	1.699 ± 0.025

the same input row order and optimizer settings. Table 12(b) shows similar trustworthiness and DE after both stages, indicating robustness to the tested raw initializers.

Comparison from a shared t-SNE initialization. To test whether ISDR's advantage over S-UMAP depends on its initialization, both methods start from the same saved t-SNE coordinates for each of five paired seeds. ISDR uses the standard refinement schedule for 750 steps. S-UMAP uses its fuzzy graph and optimizer for 200 epochs, with its input-coordinate rescaling and jitter disabled to preserve the shared start. Table 12(c) and Figure 9 compare the six-point paths. Weights are specific to each method. In every paired seed, each S-UMAP point is dominated by at least one ISDR point in the Trust–DE plane. Thus ISDR's empirical advantage persists with a shared geometric

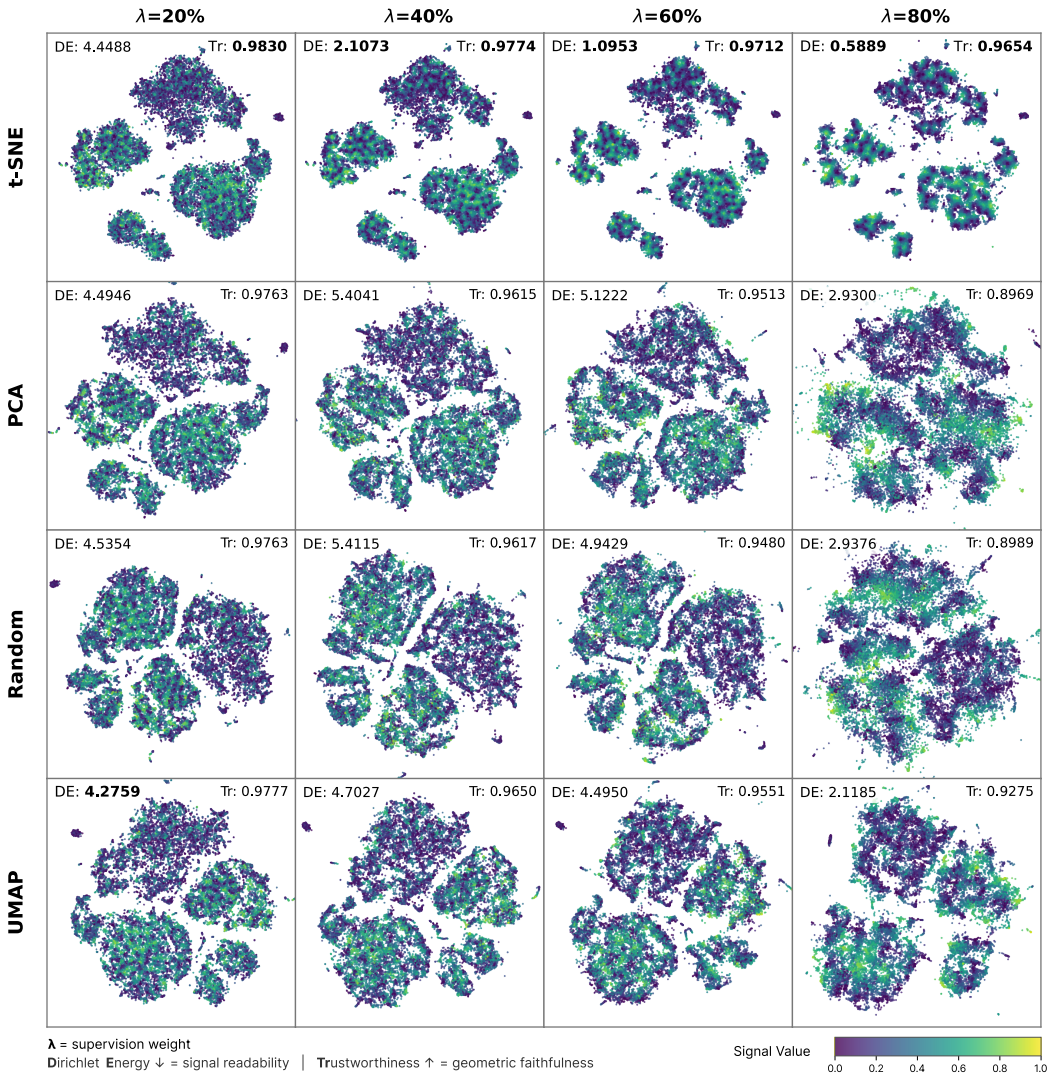

Figure 7: HPDv3 initialization sensitivity on seed 42 with PCA-50 preprocessing. Rows vary the coordinates passed to refinement, and columns vary λ. Colors show the normalized signal.

initialization. Figure 8 qualitatively compares S-UMAP's default and shared t-SNE initializations across five supervision weights, using the same seed-42 samples and normalized signal colors as the other supplementary embedding tables.

NW support size k. The support size k controls the neighborhood used by the NW readout. Figure 10 reruns ISDR on HPDv3 with $k \in \{30, 60, 90, 150\}$ and $\lambda \in \{0, 0.1, 0.3, 0.5, 0.9\}$. Trustworthiness is similar across support sizes. The effect on DE depends on the supervision weight; larger supports yield lower DE at the highest weight. We use $k = \lfloor 3p \rfloor = 90$ in the main experiments.

Support rebuild interval R. The low-dimensional NW support graph is rebuilt every R refinement steps. Figure 11 varies $R \in \{25, 50, 100, 250\}$ on HPDv3 using the same fixed-data seed and default support size $k = 90$. The operating path varies little for $R \leq 100$, while $R = 250$ yields higher DE at high supervision weights. We use $R = 50$ as the default.

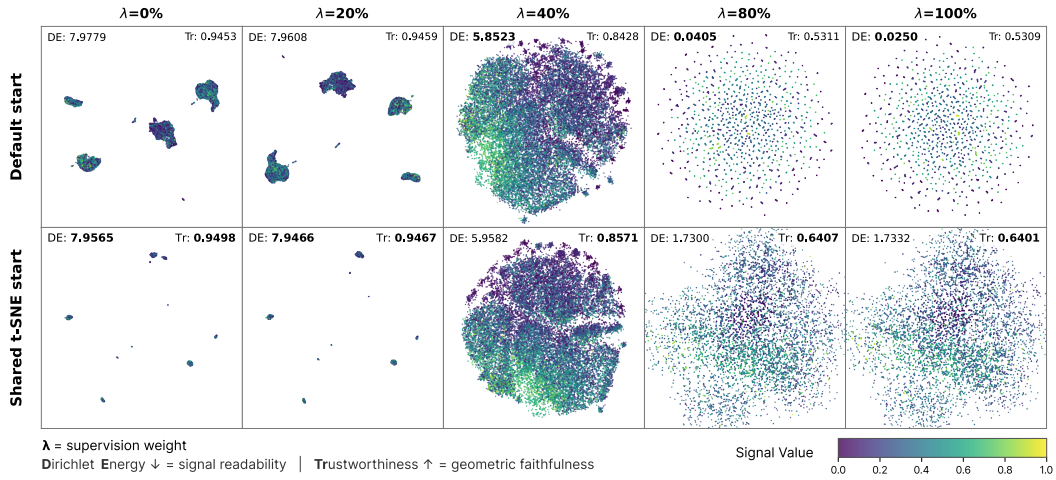

Figure 8: HPDv3 S-UMAP initialization sensitivity on seed 42. Rows vary the initialization (default or shared t-SNE), and columns vary $\lambda \in \{0, 0.2, 0.4, 0.8, 1.0\}$. Colors show the normalized signal. Bold values indicate the lower DE or higher trustworthiness in each column.

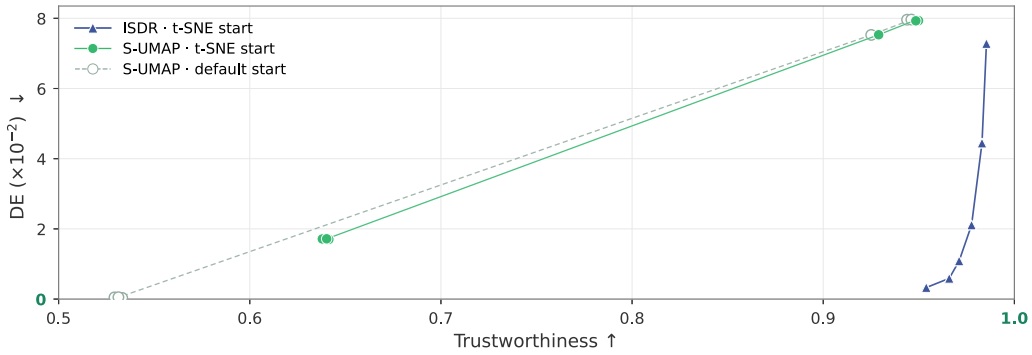

Figure 9: Geometry–signal paths for the shared-start ablation on HPDv3. Points are means over five seeds at six method-specific weights. S-UMAP's default initialization is included as a reference. Overlapping markers are slightly offset horizontally for visibility, while connecting lines retain the exact coordinates. Table 12(c) reports the shared-start means and standard deviations.

Table 13: Linear supervised DR comparison. Values are mean ± population SD; identical results across seeds are shown once. ISDR uses $\lambda = 0.4$ on HPDv3 and $\lambda = 0.8$ on Tabula Muris.

	HPDv3		Tabula Muris	
Method	Trust ↑	100×DE ↓	Trust ↑	100×DE ↓
LOL-Q3	0.6997	7.2357	0.5816	0.3464
HSIC-SPCA	0.7849 ± 0.0021	7.4122 ± 0.0157	0.5822 ± 0.0005	0.3602 ± 0.0100
SIR	0.5666	5.0893	0.5808	0.2291
Bair-SPC-CV	0.7691	7.3960	0.6162	1.5033
ISDR	0.9774 ± 0.0002	2.1077 ± 0.0083	0.9960 ± 0.0001	0.2274 ± 0.0005

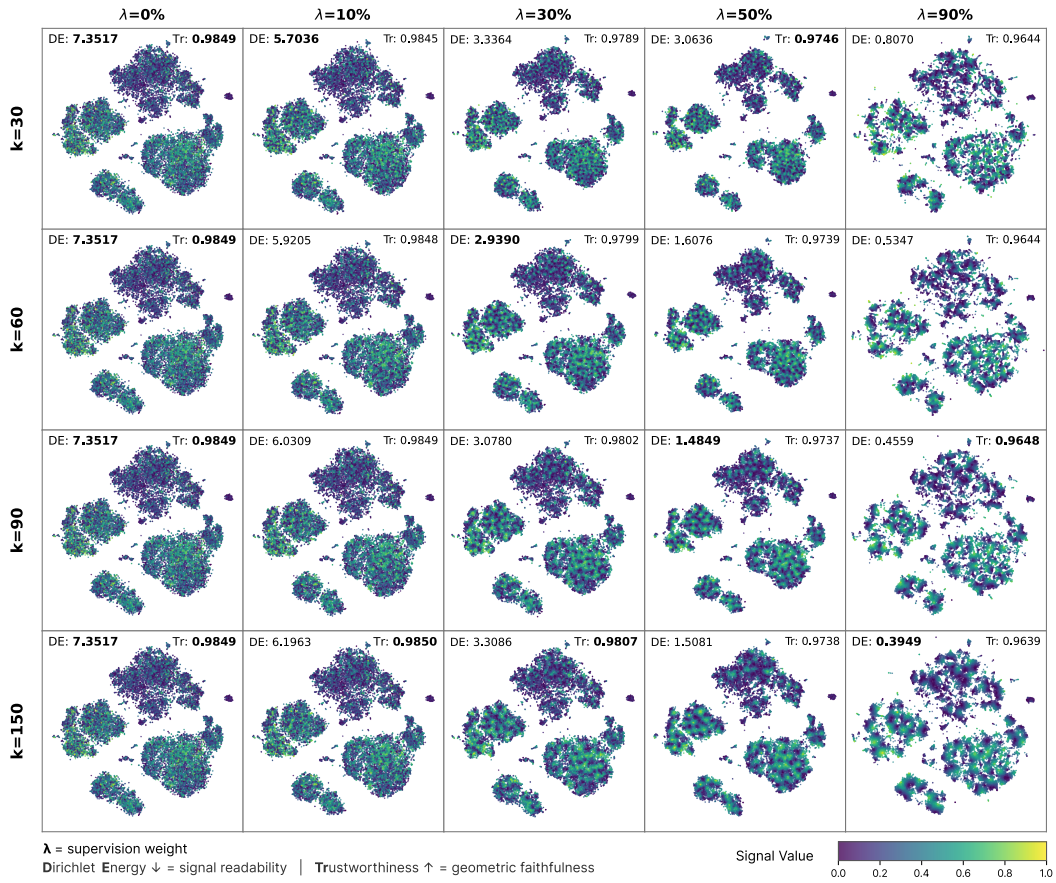

Figure 10: HPDv3 NW support sensitivity on seed 42. Rows vary support size k, columns vary λ, and points are colored by the normalized scalar signal.

C.10 Comparison with Linear Supervised DR

We compare four supervised linear projections to two dimensions on the same HPDv3 and Tabula Muris inputs, using the scalar signals and $k = 10$ metrics from the main experiments. LOL-Q3 [10] partitions the response into three quantile groups and uses their two-dimensional mean-difference subspace; SIR [17] uses ten response-quantile slices. HSIC-SPCA [25] uses an RBF response kernel with median bandwidth, approximated by 1024 random Fourier features. On HPDv3, SIR and HSIC-SPCA operate after PCA-256 preprocessing. Bair-SPC [26] screens features by their absolute correlation with the response, selects the retained fraction by five-fold response-prediction MSE, and refits a two-component PCA. Screening, scaling, PCA, and regression are fitted within each training fold. Table 13 compares these projections with the ISDR representative points.

The linear projections have substantially lower trustworthiness than ISDR on both datasets. On Tabula Muris, SIR achieves similar DE to ISDR.

D Case Study Supplement

The main text uses the ISDR map for a qualitative HPDv3 inspection of LLM-as-a-Judge confidence. This appendix records the data-processing scope behind that inspection and the interface used to examine individual samples.

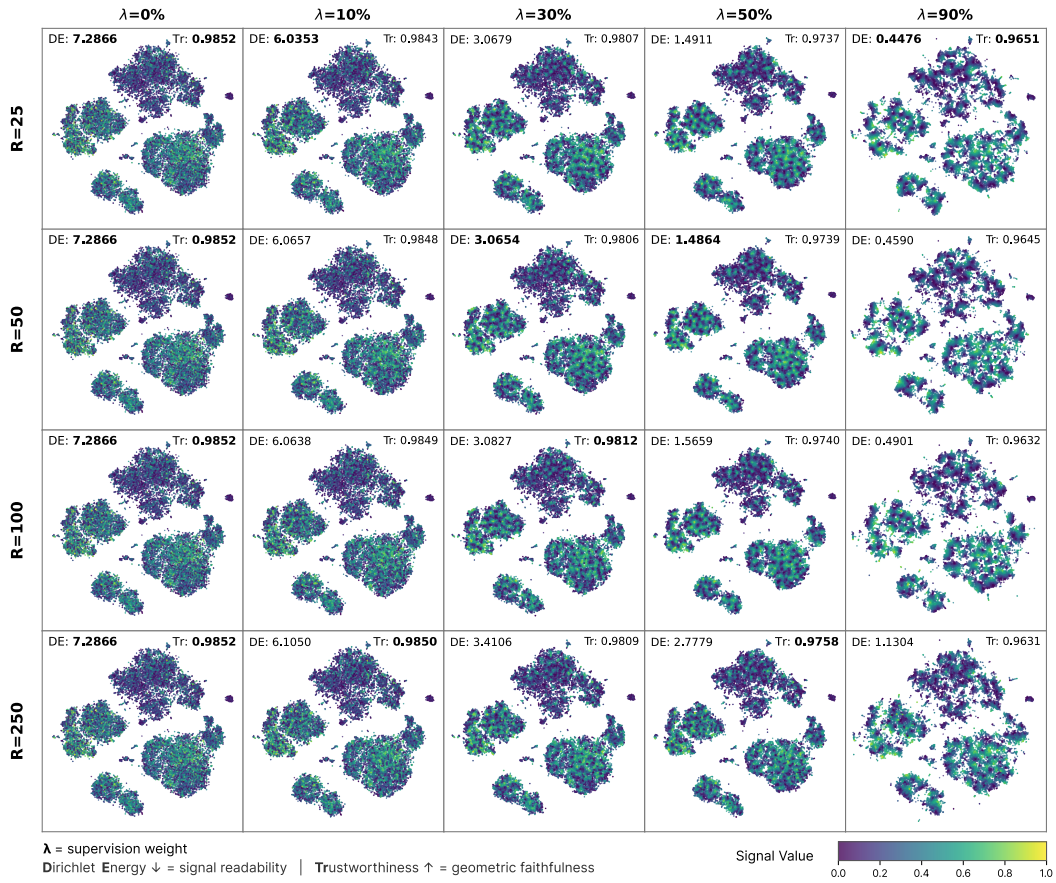

Figure 11: HPDv3 support rebuild sensitivity on seed 42. Rows vary rebuild interval R, columns vary λ, and points are colored by the normalized scalar signal.

D.1 HPDv3 Judge Construction

Each HPDv3 sample contains a text prompt and two candidate images, with generator identities and the benchmark's human preference distribution. We use Qwen3-VL-8B-Instruct as the judge model. The high-dimensional representation is its 4096-dimensional layer-19 hidden state [74], and the scalar signal is the top-two logit gap among the A/B/C judge options.

The judge prompts the model to choose between two images or a tie [75]:

> You are a helpful assistant proficient in analyzing text to image generation problems. Please choose which image better matches the text prompt, or indicate if they are equally good: "{prompt}". You are given two images: Image A: the first image; Image B: the second image. In your evaluation, weigh factors such as alignment, aesthetics, creativity, and overall image quality. Avoid any position biases and ensure that the order in which the images were presented does not influence your decision. You can only output A/B/C, where C means both images are equally good (tie). Output:

The model input is a chat-style user message containing the two images, followed by this text instruction. When the processor supports it, the thinking mode is disabled so that the first generated token corresponds directly to the answer option.

D.2 Interactive Inspection Interface

The case study was conducted using the visual analytics interface shown in Figure 12. The map is rendered as a WebGL scatter plot colored by the confidence signal. The interface supports pan, zoom, lasso or box selection, and a sample panel for inspecting the image generation prompt, both images,

judge and human preferences, and confidence values. These functions were used only to inspect regions and examples on top of a fixed ISDR embedding.

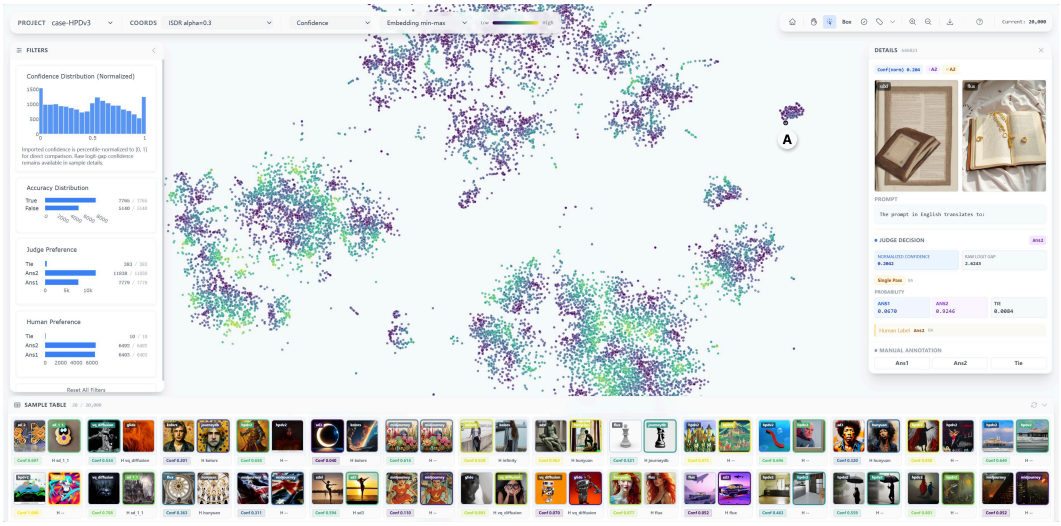

Figure 12: Interface used for the HPDv3 case-study inspection: ISDR embedding (centre), filters and sample table (left/bottom), and the per-sample inspection panel (right). Panel A shows a low-confidence prompt-artifact sample from Region A of Figure 3.

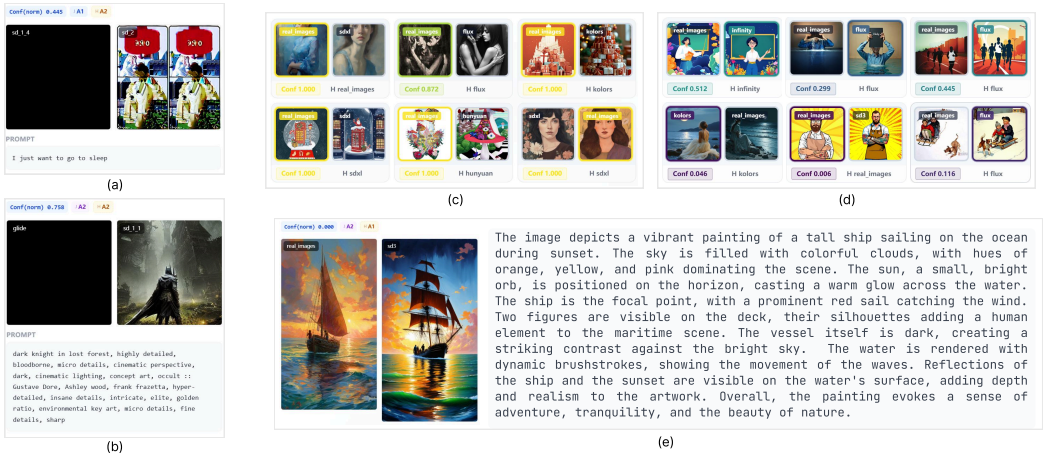

Figure 13: Representative HPDv3 samples inspected in the case study. (a, b) Image-generation failure cases from the high-confidence cluster: (a) the judge prefers a blank image to a poor generation; (b) one candidate is missing/black. (c, d) Sample grids from the high- and low-confidence sides of the local real-image/generated-image transition. (e) Full prompt for the example in Figure 3C.

D.3 Observed Regions

Section 6.4 introduces three regions: the low-confidence prompt-artifact region, the high-confidence blank-image region, and the local transition where real-image and generated-image comparisons change confidence smoothly. Figure 12A gives an example of the low-confidence prompt-artifact region. Figure 13 (a) and (b) show two image-generation failures. In (a), the judge prefers a blank image to a poor generation for the prompt "I just want to go to sleep." Figure 13 (c) and (d) show samples from the high- and low-confidence sides of the local transition. Figure 13 (e) provides the full prompt for the example in Figure 3C, showing how it describes the real image in detail. Together, these regions illustrate why the case study requires both geometric faithfulness and signal readability:

nearby samples remain semantically inspectable, while the scalar field exposes local confidence transitions.

D.4 Checking for Data Artifacts

Prompt artifact. Section 6.4 reported that the low-confidence cluster in Figure 3A consists of samples whose prompt is the literal phrase "The prompt in English translates to:". We checked whether this is introduced by our judge pipeline or already present in the upstream HPDv3 release. A streaming text scan of HPDv3 `train.json` finds 9,448 samples whose `prompt` field contains this phrase: 6,620 are exactly this single sentence with no following text, 2,338 are followed by a newline and a complete English description, and 490 continue inline with an English description on the same line. The low confidence of the judge on the 6,620 "empty-prompt" samples is consistent with the prompt carrying no descriptive content for the comparison.

Blank image. Section 6.4 reported that the high-confidence cluster in Figure 3B contains samples in which one of the two candidate images is blank. We scanned for blank images: an initial filter retains images whose mean grayscale value is below 5; a stricter rule then keeps images whose maximum pixel value is at most 30 and that contain at most 20 pixels with value above 10; pixel-exact all-zero images are tallied separately. At the image level, 1,009 of the 1,093,082 local HPDv3 images (0.0923%) are flagged as blank or near-blank. Among these, 991 (0.0907% of all images, 98.2% of the flagged set) are pixel-exact all-black images, and the remaining 18 (1.8%) are not strictly zero but still nearly black. At the pair level, the full set of 1,164,177 annotated pairs contains 1,954 pairs (0.1678%) with at least one flagged image, of which 33 pairs have both images flagged; outside these 33 both-flagged pairs, every flagged image occupies the `path2` slot. The flagged images are dominated by `sd_1_1` and `sd_1_4`, with a smaller contribution from `glide` and a few cases from `hpdv2` and `midjourney`. The training-split pair listing of 1,139,924 entries contains 4,776 entries (0.4190%) with a flagged image, all in the `path2` slot and referencing 1,007 unique flagged images. The 14,400-entry test-split listing contains no flagged pairs. The high judge confidence on the cluster in Figure 3B is consistent with this distribution: in nearly all affected pairs the blank or near-blank image appears on a single side, so the comparison effectively reduces to assessing the non-blank candidate.

E Tabular Regression Benchmark Results

Table 14: Large-sample tabular regression benchmark design.

Dataset	N	D	Target
California Housing [69]	20,640	8	median house value
Superconductivity [70]	21,263	81	critical temperature

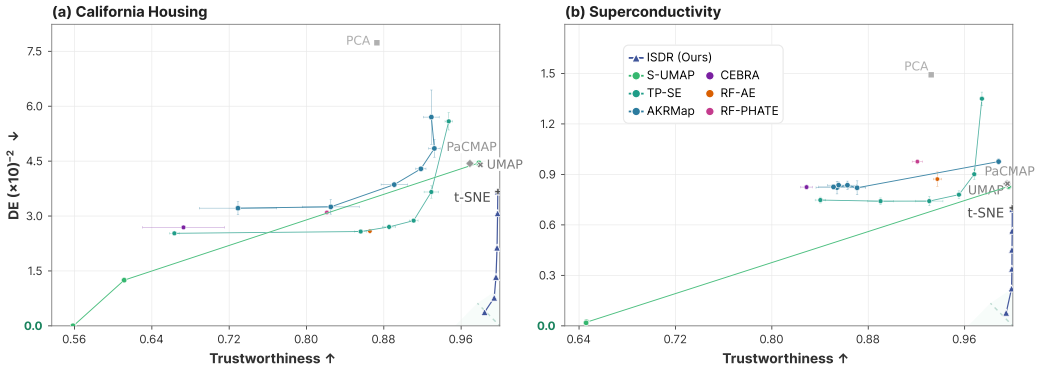

Figure 14: Geometry–signal sweeps on the large-sample tabular regression benchmarks. Points show mean ± one standard deviation over five DR seeds.

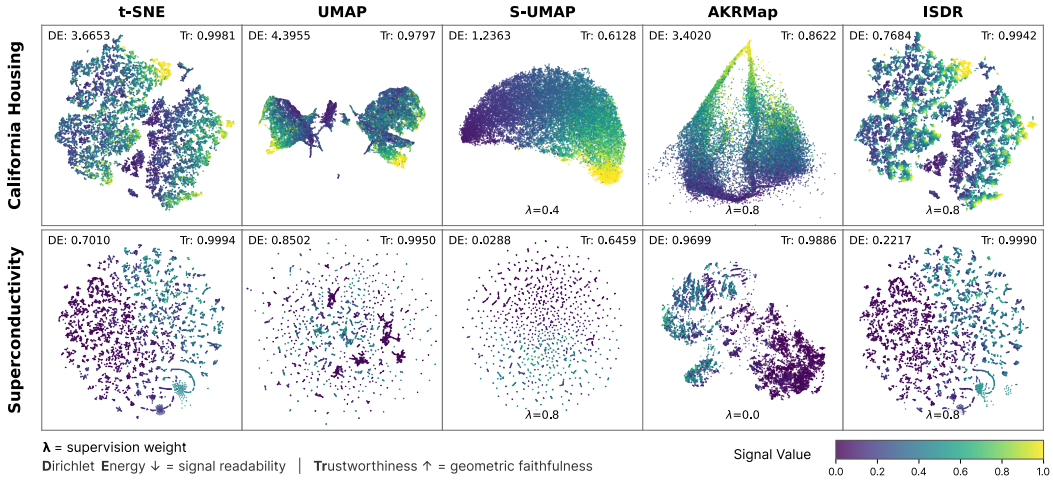

Figure 15: Representative train-set embeddings for the regression benchmarks. Supervised columns use the Pareto-knee rule in Appendix C.3. Colors show the normalized target.

We supplement the main HPDv3 and Tabula Muris experiments with two large-sample public tabular regression benchmarks (Table 14) whose scalar targets are observed labels rather than constructed signals. The benchmarks span geographic/economic and materials science domains and both have at least 20,000 samples. All features are standardized per column, and the original continuous target is used as the scalar field. Because these datasets are already low-dimensional ($D \leq 81$), no PCA pre-reduction is applied before the dimensionality-reduction methods. For ISDR, the scalar target is z-scored inside the NW regularizer so that the supervision weight is not tied to dataset-specific physical units, following Appendix B.6. We use the same five DR seeds and method grid as the main experiments. Both regression datasets use perplexity 30.

Figures 14 and 15 show the sweeps and representative embeddings. Table 15 reports all operating points and runtimes.

On California Housing, the ISDR Pareto knee ($\lambda = 0.8$) reduces $10^2 \times \mathrm{DE}$ from the geometry-only value 3.67 to 0.77, with trustworthiness changing from 0.998 to 0.994. S-UMAP at $\lambda = 0.6$ reaches DE rounded to zero but trustworthiness 0.559. On Superconductivity, the ISDR Pareto knee ($\lambda = 0.8$) reduces $10^2 \times \mathrm{DE}$ from the t-SNE anchor's 0.70 to 0.22 while retaining trustworthiness 0.999 and 10-NN overlap 0.711.

E.1 Full Regression Grids

Table 15 reports every operating point for both California Housing and Superconductivity.

Table 15: Full regression grids on the large-sample tabular regression benchmarks (California Housing and Superconductivity). Values are mean±SD over five DR seeds. Methods are grouped by row; swept methods report all six standard-grid points $\lambda \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$.

Method	Parameter	Time (s)	Trust ↑	Cont. ↑	10-NN ↑	$10^2 \times \mathrm{DE}$ ↓	Φ_σ ↓	M_{ctr}
Dataset: California Housing								
PCA	–	0.0±0.0	0.873±0.000	0.982±0.000	0.049±0.000	7.74±0.00	0.153±0.000	0.999±0.000
t-SNE	–	32.0±1.3	0.998±0.000	0.986±0.000	0.523±0.000	3.67±0.00	0.147±0.000	0.982±0.000
UMAP	–	33.6±0.7	0.980±0.000	0.991±0.000	0.269±0.001	4.40±0.01	0.282±0.004	0.976±0.003
PaCMAP	–	20.3±0.4	0.969±0.001	0.991±0.000	0.212±0.001	4.44±0.01	0.256±0.001	0.951±0.001
S-UMAP	$\lambda = 0.0$	39.4±0.4	0.978±0.001	0.991±0.000	0.252±0.001	4.42±0.03	0.303±0.009	0.972±0.004
	$\lambda = 0.2$	39.2±0.5	0.978±0.001	0.991±0.000	0.251±0.001	4.45±0.05	0.306±0.005	0.973±0.005
	$\lambda = 0.4$	49.9±0.5	0.611±0.002	0.830±0.001	0.016±0.000	1.25±0.02	0.483±0.002	0.364±0.007
	$\lambda = 0.6$	48.5±0.6	0.559±0.001	0.645±0.002	0.002±0.000	0.00±0.00	0.570±0.003	0.431±0.029
	$\lambda = 0.8$	49.3±2.1	0.559±0.001	0.646±0.001	0.002±0.000	0.00±0.00	0.571±0.005	0.430±0.024

Continued on next page.

Table 15 (continued from previous page)

Method	Parameter	Time (s)	Trust ↑	Cont. ↑	10-NN ↑	$10^2 \times \mathrm{DE} \downarrow$	$\Phi_\sigma \downarrow$	M_{ctr}
	$\lambda = 1.0$	49.9±2.6	0.558±0.001	0.646±0.002	0.002±0.000	0.00±0.00	0.569±0.006	0.413±0.026
CEBRA	–	14.5±0.6	0.673±0.042	0.872±0.004	0.018±0.003	2.69±0.03	0.534±0.016	0.570±0.013
TP-SE	$\lambda = 0.0$	12.9±0.5	0.947±0.004	0.986±0.001	0.098±0.006	5.59±0.24	0.272±0.061	1.039±0.066
	$\lambda = 0.2$	12.9±0.4	0.929±0.007	0.974±0.007	0.078±0.007	3.66±0.17	0.326±0.080	0.965±0.063
	$\lambda = 0.4$	12.5±0.6	0.911±0.005	0.956±0.008	0.059±0.002	2.87±0.04	0.320±0.082	0.919±0.014
	$\lambda = 0.6$	12.3±0.5	0.885±0.007	0.941±0.009	0.048±0.004	2.70±0.04	0.323±0.071	0.861±0.028
	$\lambda = 0.8$	12.1±0.3	0.856±0.006	0.913±0.013	0.038±0.002	2.58±0.05	0.428±0.058	0.773±0.046
	$\lambda = 1.0$	12.0±0.9	0.663±0.005	0.881±0.011	0.019±0.001	2.53±0.03	0.517±0.053	0.717±0.043
RF-PHATE	–	19.7±1.0	0.821±0.001	0.937±0.001	0.043±0.001	3.10±0.02	0.256±0.003	0.821±0.009
RF-AE	–	49.7±2.4	0.866±0.001	0.909±0.004	0.051±0.001	2.59±0.02	0.315±0.014	0.812±0.017
AKRMap	$\lambda = 0.0$	38.6±1.8	0.929±0.008	0.990±0.001	0.091±0.004	5.70±0.74	0.182±0.034	0.957±0.022
	$\lambda = 0.2$	40.8±2.4	0.932±0.006	0.990±0.001	0.087±0.005	4.85±0.24	0.190±0.024	0.920±0.038
	$\lambda = 0.4$	40.3±1.6	0.918±0.005	0.987±0.001	0.075±0.005	4.29±0.05	0.181±0.021	0.859±0.003
	$\lambda = 0.6$	39.4±1.3	0.891±0.014	0.981±0.002	0.057±0.006	3.86±0.07	0.174±0.016	0.814±0.011
	$\lambda = 0.8$	42.9±7.0	0.825±0.029	0.953±0.009	0.034±0.003	3.25±0.20	0.268±0.060	0.716±0.043
	$\lambda = 1.0$	39.7±1.1	0.729±0.040	0.860±0.011	0.016±0.002	3.22±0.18	0.594±0.027	0.568±0.029
ISDR	$\lambda = 0.0$	53.3±1.1	0.998±0.000	0.986±0.000	0.520±0.000	3.67±0.00	0.148±0.000	0.986±0.000
	$\lambda = 0.2$	84.7±1.8	0.998±0.000	0.986±0.000	0.517±0.000	3.07±0.00	0.147±0.000	0.983±0.000
	$\lambda = 0.4$	84.9±0.7	0.997±0.000	0.986±0.000	0.499±0.000	2.14±0.00	0.148±0.000	0.984±0.000
	$\lambda = 0.6$	83.6±0.2	0.996±0.000	0.986±0.000	0.465±0.000	1.34±0.00	0.148±0.000	0.981±0.000
	$\lambda = 0.8$	83.6±1.5	0.994±0.000	0.985±0.000	0.412±0.000	0.77±0.00	0.149±0.000	0.982±0.000
	$\lambda = 1.0$	64.7±0.8	0.984±0.000	0.983±0.000	0.158±0.000	0.37±0.00	0.149±0.000	0.983±0.000
Dataset: Superconductivity								
PCA	–	0.1±0.0	0.932±0.000	0.994±0.000	0.440±0.000	1.49±0.00	0.156±0.000	0.841±0.000
t-SNE	–	34.9±1.7	0.999±0.000	0.998±0.000	0.791±0.000	0.70±0.00	1.193±0.000	0.832±0.000
UMAP	–	46.0±3.6	0.996±0.000	0.976±0.001	0.629±0.002	0.84±0.01	65.549±23.216	1.732±0.383
PaCMAP	–	21.5±0.9	0.995±0.000	0.990±0.000	0.595±0.003	0.84±0.02	18.513±7.455	1.315±0.282
S-UMAP	$\lambda = 0.0$	49.4±1.5	0.997±0.000	0.980±0.002	0.624±0.001	0.83±0.01	81.494±53.912	1.712±0.297
	$\lambda = 0.2$	49.3±1.7	0.997±0.000	0.980±0.001	0.621±0.002	0.83±0.01	93.587±52.101	1.868±0.157
	$\lambda = 0.4$	50.0±0.5	0.996±0.000	0.980±0.002	0.623±0.003	0.83±0.01	82.207±41.492	2.025±0.107
	$\lambda = 0.6$	52.0±1.1	0.646±0.001	0.667±0.004	0.010±0.000	0.02±0.00	90.806±45.722	0.955±0.038
	$\lambda = 0.8$	52.2±0.8	0.646±0.001	0.667±0.004	0.010±0.000	0.02±0.01	71.943±6.094	0.909±0.088
	$\lambda = 1.0$	53.2±0.9	0.646±0.001	0.663±0.006	0.009±0.000	0.02±0.00	94.719±34.948	0.915±0.110
CEBRA	–	14.2±0.6	0.829±0.005	0.970±0.002	0.317±0.009	0.82±0.01	0.525±0.014	0.431±0.020
TP-SE	$\lambda = 0.0$	12.9±0.6	0.974±0.001	0.993±0.001	0.435±0.022	1.35±0.04	0.220±0.011	0.837±0.095
	$\lambda = 0.2$	12.4±0.9	0.968±0.001	0.990±0.002	0.412±0.007	0.90±0.03	0.238±0.032	0.845±0.017
	$\lambda = 0.4$	12.3±1.2	0.955±0.002	0.987±0.002	0.385±0.004	0.78±0.02	0.275±0.043	0.817±0.034
	$\lambda = 0.6$	12.6±0.5	0.931±0.011	0.984±0.003	0.367±0.012	0.74±0.03	0.299±0.046	0.780±0.044
	$\lambda = 0.8$	12.8±0.4	0.890±0.011	0.980±0.001	0.364±0.008	0.74±0.02	0.312±0.025	0.766±0.046
	$\lambda = 1.0$	12.8±0.4	0.840±0.004	0.974±0.001	0.341±0.004	0.75±0.02	0.337±0.024	0.671±0.052
RF-PHATE	–	28.7±1.2	0.921±0.004	0.959±0.008	0.360±0.008	0.98±0.01	82.234±49.493	1.482±0.261
RF-AE	–	57.8±2.2	0.937±0.003	0.963±0.004	0.369±0.006	0.87±0.04	45.495±36.269	1.246±0.185
AKRMap	$\lambda = 0.0$	38.6±0.4	0.988±0.001	0.998±0.000	0.556±0.005	0.98±0.02	0.217±0.011	0.690±0.020
	$\lambda = 0.2$	41.5±2.5	0.871±0.007	0.972±0.003	0.331±0.013	0.82±0.04	0.732±0.029	0.355±0.024
	$\lambda = 0.4$	41.4±1.1	0.863±0.008	0.968±0.002	0.318±0.010	0.84±0.02	0.759±0.044	0.354±0.017
	$\lambda = 0.6$	42.9±3.5	0.854±0.017	0.966±0.004	0.315±0.019	0.82±0.04	0.745±0.035	0.351±0.020
	$\lambda = 0.8$	43.3±5.1	0.855±0.009	0.965±0.003	0.311±0.013	0.84±0.02	0.757±0.051	0.349±0.023
	$\lambda = 1.0$	41.6±3.3	0.851±0.012	0.965±0.003	0.305±0.005	0.83±0.02	0.763±0.054	0.348±0.013
ISDR	$\lambda = 0.0$	57.3±1.9	1.000±0.000	0.998±0.000	0.790±0.000	0.69±0.00	1.022±0.000	0.837±0.000
	$\lambda = 0.2$	86.5±2.1	0.999±0.000	0.998±0.000	0.781±0.000	0.56±0.00	1.020±0.000	0.842±0.000
	$\lambda = 0.4$	88.3±1.9	0.999±0.000	0.998±0.000	0.770±0.000	0.45±0.00	0.946±0.000	0.837±0.000
	$\lambda = 0.6$	87.2±1.5	0.999±0.000	0.998±0.000	0.749±0.000	0.34±0.00	0.991±0.000	0.829±0.000
	$\lambda = 0.8$	87.1±1.2	0.999±0.000	0.998±0.000	0.711±0.000	0.22±0.00	0.996±0.000	0.835±0.000
	$\lambda = 1.0$	67.3±2.5	0.994±0.000	0.997±0.000	0.357±0.000	0.08±0.00	1.105±0.000	0.827±0.000

F Traditional Classification Benchmark Results

We include traditional classification benchmarks as supplementary diagnostics, encoding labels as integer-valued scalar signals and evaluating the same trustworthiness–DE plane. DE therefore depends on the chosen numerical label ordering.

Table 16: Traditional classification benchmark design.

Dataset	Source	N	Dim.	Encoded signal
MNIST	torchvision MNIST	10,000	784	digit label 0–9
Fashion-MNIST	torchvision Fashion-MNIST	10,000	784	garment label 0–9
OrganCMNIST	MedMNIST OrganCMNIST	10,000	784	organ label 0–10
PathMNIST	MedMNIST PathMNIST	10,000	2352	tissue label 0–8

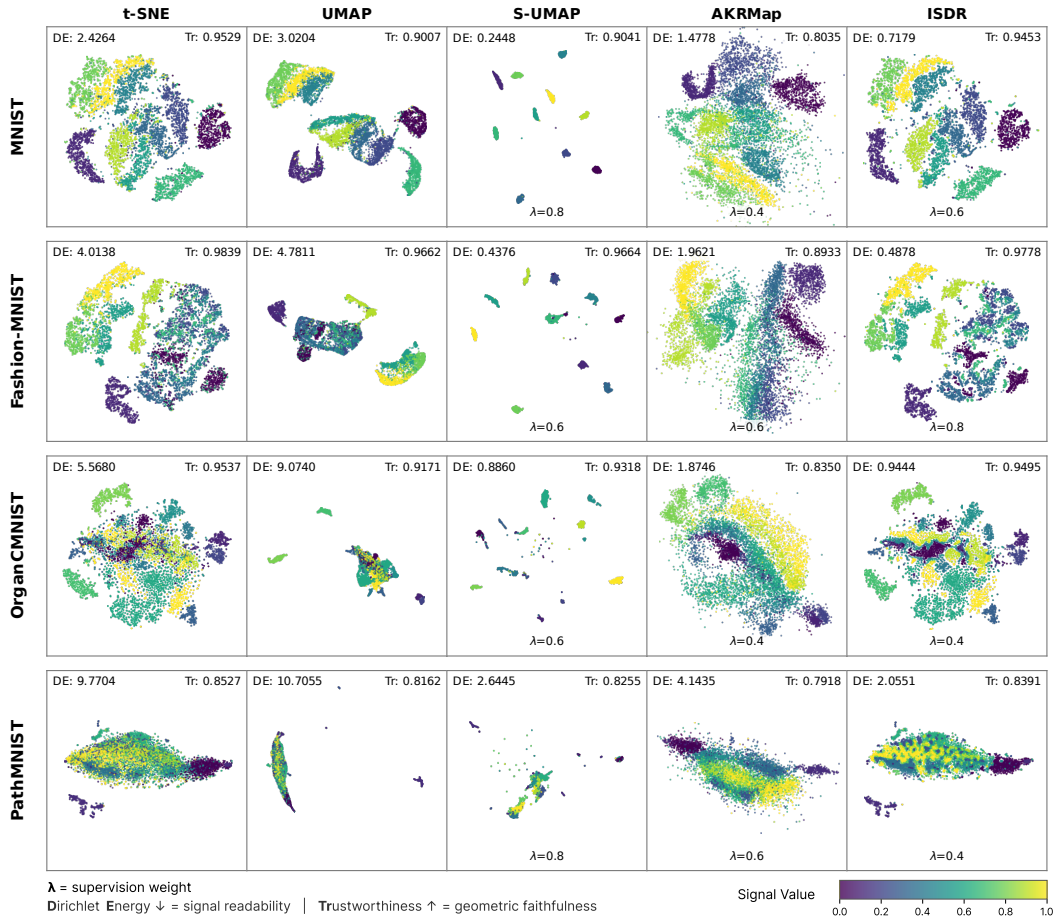

Figure 16: Traditional classification embeddings on four supplementary datasets. Columns show t-SNE, UMAP, S-UMAP, AKRMap, and ISDR. Supervised columns use the Pareto-knee rule in Appendix C.3 on the unified grid. Points are colored by the normalized encoded label signal.

F.1 Protocol

Each benchmark uses $N = 10{,}000$ examples with an $80/20$ split fixed by data seed 42 and split seed 42. The five reported DR seeds change only method stochasticity, not the sampled examples or the train/test split. We use the same methods as in the main experiments and the standard supervision grid $\lambda \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ for S-UMAP, TP-SE, AKRMap, and ISDR. CEBRA, RF-PHATE, and RF-AE are treated as single-point supervised baselines, as in the main protocol. Because the nonparametric maps are fitted to the training set, we report visualization metrics on that set. The categorical proxy columns report the scale-invariant geometry distortion Φ_σ and a same-label contraction proxy M_{ctr}, computed from sampled pairs with identical class labels. Table 16 summarizes the four datasets used.

F.2 Full Classification Grids

Table 17 reports every standard-grid point for the four supplementary classification datasets. Figure 16 shows representative single-seed embeddings alongside the t-SNE and UMAP anchors.

Table 17: Full traditional classification grids on the four supplementary datasets (MNIST, Fashion-MNIST, OrganCMNIST, PathMNIST). Values are mean±SD over five DR seeds. Categorical proxy columns use identical-label sampled pairs for M_{ctr}. A slash (/) marks degenerate S-UMAP runs for which metric values are not reported.

Method	Parameter	Trust ↑	Cont. ↑	10-NN ↑	$10^2 \times \mathrm{DE} \downarrow$	$\Phi_\sigma \downarrow$	M_{ctr}
Dataset: MNIST							
PCA	–	0.665±0.000	0.925±0.000	0.031±0.000	13.933±0.000	0.274±0.000	0.821±0.000
t-SNE	–	0.953±0.000	0.962±0.000	0.393±0.000	2.423±0.005	0.265±0.000	0.420±0.000
UMAP	–	0.901±0.000	0.962±0.000	0.224±0.001	3.018±0.038	0.357±0.006	0.330±0.002
PaCMAP	–	0.888±0.001	0.955±0.000	0.208±0.001	2.950±0.121	0.406±0.001	0.275±0.002
S-UMAP	$\lambda = 0.0$	0.905±0.001	0.945±0.001	0.229±0.001	0.481±0.026	0.388±0.012	0.116±0.006
	$\lambda = 0.2$	0.905±0.001	0.936±0.003	0.229±0.001	0.358±0.025	0.390±0.014	0.091±0.003
	$\lambda = 0.4$	0.904±0.001	0.929±0.002	0.227±0.001	0.283±0.023	0.379±0.010	0.083±0.007
	$\lambda = 0.6$	0.904±0.000	0.925±0.002	0.227±0.001	0.263±0.012	0.367±0.020	0.075±0.003
	$\lambda = 0.8$	0.904±0.001	0.918±0.007	0.228±0.001	0.251±0.015	0.360±0.007	0.075±0.004
	$\lambda = 1.0$				/		
CEBRA	–	0.672±0.002	0.919±0.002	0.026±0.001	0.654±0.175	0.316±0.010	0.255±0.005
TP-SE	$\lambda = 0.0$	0.806±0.008	0.943±0.007	0.087±0.006	8.441±0.251	0.305±0.012	0.724±0.025
	$\lambda = 0.2$	0.804±0.006	0.941±0.004	0.082±0.004	5.085±0.356	0.296±0.008	0.550±0.023
	$\lambda = 0.4$	0.783±0.011	0.929±0.007	0.067±0.005	1.874±0.382	0.274±0.022	0.453±0.024
	$\lambda = 0.6$	0.744±0.007	0.921±0.004	0.048±0.004	1.325±0.241	0.258±0.018	0.421±0.007
	$\lambda = 0.8$	0.709±0.004	0.918±0.004	0.035±0.001	1.237±0.304	0.259±0.015	0.394±0.005
	$\lambda = 1.0$	0.694±0.001	0.914±0.004	0.031±0.001	1.055±0.137	0.265±0.013	0.374±0.008
RF-PHATE	–	0.702±0.002	0.902±0.003	0.037±0.001	5.281±0.225	1.179±0.176	0.260±0.003
RF-AE	–	0.776±0.004	0.923±0.003	0.066±0.002	1.865±0.104	0.761±0.098	0.320±0.002
AKRMap	$\lambda = 0.0$	0.849±0.006	0.963±0.004	0.180±0.006	5.579±0.639	0.226±0.007	0.545±0.036
	$\lambda = 0.2$	0.838±0.003	0.961±0.001	0.164±0.004	1.937±0.119	0.237±0.006	0.460±0.010
	$\lambda = 0.4$	0.814±0.010	0.952±0.002	0.133±0.007	1.196±0.159	0.249±0.022	0.450±0.035
	$\lambda = 0.6$	0.785±0.010	0.941±0.003	0.100±0.003	0.662±0.060	0.245±0.012	0.421±0.021
	$\lambda = 0.8$	0.733±0.007	0.923±0.001	0.057±0.003	0.433±0.028	0.250±0.003	0.385±0.032
	$\lambda = 1.0$	0.672±0.005	0.888±0.003	0.026±0.001	0.427±0.010	0.326±0.031	0.359±0.015
ISDR	$\lambda = 0.0$	0.951±0.001	0.961±0.000	0.391±0.001	2.423±0.041	0.272±0.001	0.407±0.001
	$\lambda = 0.2$	0.950±0.001	0.961±0.000	0.390±0.001	1.641±0.034	0.273±0.002	0.403±0.002
	$\lambda = 0.4$	0.948±0.000	0.961±0.000	0.387±0.001	1.049±0.034	0.274±0.001	0.398±0.001
	$\lambda = 0.6$	0.945±0.001	0.961±0.000	0.384±0.001	0.685±0.032	0.277±0.002	0.390±0.001
	$\lambda = 0.8$	0.942±0.001	0.960±0.000	0.379±0.002	0.499±0.030	0.279±0.002	0.377±0.001
	$\lambda = 1.0$	0.914±0.001	0.958±0.000	0.252±0.004	0.419±0.031	0.272±0.002	0.393±0.001
Dataset: Fashion-MNIST							
PCA	–	0.877±0.000	0.969±0.000	0.051±0.000	7.063±0.000	0.132±0.000	0.695±0.000
t-SNE	–	0.984±0.000	0.981±0.000	0.370±0.000	4.022±0.006	0.179±0.000	0.549±0.000
UMAP	–	0.966±0.000	0.981±0.001	0.218±0.001	4.779±0.033	0.374±0.008	0.374±0.005
PaCMAP	–	0.959±0.001	0.979±0.000	0.193±0.002	4.885±0.096	0.270±0.003	0.368±0.007
S-UMAP	$\lambda = 0.0$	0.965±0.001	0.968±0.001	0.225±0.002	2.487±0.018	0.368±0.006	0.237±0.003
	$\lambda = 0.2$	0.965±0.000	0.953±0.002	0.226±0.001	1.316±0.090	0.330±0.011	0.210±0.008
	$\lambda = 0.4$	0.966±0.001	0.925±0.003	0.228±0.001	0.582±0.035	0.347±0.010	0.135±0.002
	$\lambda = 0.6$	0.965±0.001	0.903±0.007	0.228±0.001	0.430±0.021	0.351±0.039	0.104±0.008
	$\lambda = 0.8$	0.965±0.001	0.890±0.016	0.227±0.001	0.435±0.045	0.344±0.029	0.111±0.004
	$\lambda = 1.0$				/		
CEBRA	–	0.796±0.003	0.937±0.001	0.031±0.002	2.152±0.131	0.330±0.034	0.280±0.017
TP-SE	$\lambda = 0.0$	0.931±0.001	0.975±0.001	0.086±0.001	6.406±0.095	0.159±0.008	0.636±0.012
	$\lambda = 0.2$	0.934±0.000	0.976±0.001	0.086±0.001	5.663±0.121	0.173±0.014	0.577±0.027
	$\lambda = 0.4$	0.932±0.001	0.970±0.003	0.085±0.002	4.686±0.058	0.207±0.026	0.539±0.036
	$\lambda = 0.6$	0.917±0.004	0.963±0.004	0.071±0.002	3.757±0.140	0.211±0.032	0.498±0.014
	$\lambda = 0.8$	0.884±0.003	0.955±0.003	0.056±0.001	3.079±0.378	0.227±0.020	0.483±0.018
	$\lambda = 1.0$	0.813±0.002	0.938±0.004	0.035±0.001	2.449±0.336	0.303±0.043	0.455±0.009
RF-PHATE	–	0.866±0.004	0.949±0.002	0.057±0.002	3.930±0.121	0.534±0.006	0.166±0.001

Continued on next page.

Table 17 (continued from previous page)

Method	Parameter	Trust ↑	Cont. ↑	10-NN ↑	$10^2 \times \mathrm{DE} \downarrow$	$\Phi_\sigma \downarrow$	M_{ctr}
RF-AE	–	0.900±0.010	0.949±0.004	0.084±0.008	3.294±0.097	0.332±0.018	0.381±0.021
AKRMap	$\lambda = 0.0$	0.951±0.001	0.985±0.001	0.177±0.002	5.171±0.174	0.183±0.019	0.514±0.013
	$\lambda = 0.2$	0.945±0.002	0.983±0.001	0.155±0.002	3.198±0.152	0.178±0.006	0.494±0.004
	$\lambda = 0.4$	0.931±0.001	0.976±0.000	0.121±0.002	2.464±0.168	0.179±0.002	0.507±0.015
	$\lambda = 0.6$	0.909±0.008	0.969±0.001	0.094±0.003	1.884±0.101	0.181±0.007	0.497±0.008
	$\lambda = 0.8$	0.867±0.007	0.953±0.002	0.064±0.002	1.133±0.150	0.194±0.008	0.454±0.021
	$\lambda = 1.0$	0.785±0.016	0.886±0.005	0.025±0.002	1.023±0.200	0.285±0.013	0.426±0.024
ISDR	$\lambda = 0.0$	0.983±0.000	0.981±0.000	0.367±0.000	4.055±0.015	0.187±0.001	0.529±0.001
	$\lambda = 0.2$	0.983±0.000	0.981±0.000	0.366±0.001	3.171±0.010	0.187±0.000	0.527±0.001
	$\lambda = 0.4$	0.981±0.000	0.980±0.000	0.358±0.001	1.913±0.009	0.187±0.000	0.526±0.001
	$\lambda = 0.6$	0.979±0.000	0.980±0.000	0.350±0.001	1.057±0.014	0.188±0.001	0.522±0.001
	$\lambda = 0.8$	0.977±0.000	0.978±0.000	0.335±0.000	0.499±0.010	0.190±0.001	0.516±0.001
	$\lambda = 1.0$	0.963±0.001	0.975±0.000	0.196±0.002	0.243±0.027	0.184±0.000	0.532±0.001
Dataset: OrganCMNIST							
PCA	–	0.801±0.000	0.939±0.000	0.047±0.000	13.146±0.000	0.212±0.000	0.714±0.000
t-SNE	–	0.954±0.000	0.934±0.000	0.317±0.000	5.601±0.035	0.157±0.000	0.668±0.000
UMAP	–	0.919±0.001	0.943±0.003	0.174±0.001	8.962±0.111	0.465±0.040	0.451±0.050
PaCMAP	–	0.919±0.002	0.898±0.001	0.172±0.003	6.193±0.143	0.240±0.007	0.449±0.012
S-UMAP	$\lambda = 0.0$	0.932±0.000	0.903±0.003	0.197±0.001	3.304±0.104	0.450±0.037	0.248±0.010
	$\lambda = 0.2$	0.933±0.001	0.883±0.002	0.198±0.001	1.747±0.092	0.318±0.022	0.224±0.009
	$\lambda = 0.4$	0.931±0.000	0.833±0.005	0.198±0.001	1.036±0.064	0.351±0.009	0.178±0.013
	$\lambda = 0.6$	0.931±0.001	0.777±0.011	0.197±0.001	0.882±0.029	0.370±0.011	0.153±0.013
	$\lambda = 0.8$	0.930±0.001	0.785±0.018	0.196±0.001	0.904±0.023	0.403±0.025	0.130±0.009
	$\lambda = 1.0$	/					
CEBRA	–	0.704±0.006	0.829±0.018	0.020±0.001	4.347±1.342	0.252±0.017	0.331±0.009
TP-SE	$\lambda = 0.0$	0.850±0.003	0.934±0.003	0.065±0.001	12.419±0.255	0.209±0.013	0.715±0.020
	$\lambda = 0.2$	0.855±0.002	0.933±0.003	0.068±0.001	11.823±0.214	0.203±0.013	0.693±0.044
	$\lambda = 0.4$	0.851±0.003	0.914±0.004	0.062±0.003	10.295±0.708	0.266±0.028	0.581±0.009
	$\lambda = 0.6$	0.840±0.004	0.900±0.006	0.055±0.002	7.320±1.409	0.327±0.033	0.531±0.009
	$\lambda = 0.8$	0.817±0.010	0.880±0.010	0.045±0.002	4.903±0.934	0.373±0.045	0.497±0.015
	$\lambda = 1.0$	0.742±0.006	0.853±0.007	0.029±0.001	3.993±0.708	0.369±0.028	0.497±0.026
RF-PHATE	–	0.803±0.002	0.862±0.002	0.043±0.001	9.067±0.229	0.665±0.064	0.271±0.006
RF-AE	–	0.852±0.002	0.849±0.005	0.067±0.001	4.625±0.072	0.340±0.020	0.368±0.005
AKRMap	$\lambda = 0.0$	0.864±0.003	0.956±0.001	0.103±0.002	11.165±0.222	0.166±0.004	0.776±0.012
	$\lambda = 0.2$	0.847±0.003	0.941±0.003	0.075±0.001	3.837±0.329	0.166±0.004	0.765±0.017
	$\lambda = 0.4$	0.832±0.002	0.929±0.002	0.059±0.001	2.212±0.261	0.170±0.004	0.732±0.016
	$\lambda = 0.6$	0.814±0.004	0.898±0.002	0.046±0.001	1.306±0.222	0.180±0.004	0.670±0.045
	$\lambda = 0.8$	0.779±0.006	0.856±0.007	0.030±0.000	1.071±0.082	0.203±0.008	0.545±0.028
	$\lambda = 1.0$	0.706±0.008	0.768±0.005	0.018±0.001	1.041±0.178	0.283±0.021	0.459±0.019
ISDR	$\lambda = 0.0$	0.952±0.000	0.934±0.000	0.313±0.001	5.715±0.058	0.156±0.000	0.660±0.003
	$\lambda = 0.2$	0.951±0.000	0.933±0.000	0.310±0.001	2.019±0.038	0.157±0.000	0.658±0.002
	$\lambda = 0.4$	0.949±0.000	0.932±0.000	0.307±0.000	0.968±0.042	0.157±0.000	0.655±0.002
	$\lambda = 0.6$	0.947±0.001	0.930±0.001	0.303±0.000	0.543±0.035	0.159±0.001	0.648±0.002
	$\lambda = 0.8$	0.946±0.000	0.931±0.001	0.300±0.001	0.387±0.019	0.159±0.000	0.647±0.002
	$\lambda = 1.0$	0.938±0.001	0.928±0.001	0.234±0.003	0.378±0.035	0.159±0.000	0.654±0.002
Dataset: PathMNIST							
PCA	–	0.799±0.000	0.937±0.000	0.047±0.000	9.288±0.001	0.224±0.000	0.746±0.000
t-SNE	–	0.853±0.000	0.941±0.000	0.099±0.000	9.815±0.038	0.169±0.001	0.794±0.001
UMAP	–	0.815±0.001	0.940±0.001	0.060±0.001	10.610±0.104	0.667±0.250	0.856±0.022
PaCMAP	–	0.810±0.001	0.790±0.006	0.050±0.001	10.213±0.061	0.308±0.002	0.750±0.003
S-UMAP	$\lambda = 0.0$	0.817±0.001	0.913±0.001	0.061±0.001	6.636±0.060	0.448±0.087	0.648±0.006
	$\lambda = 0.2$	0.820±0.001	0.898±0.001	0.063±0.001	5.046±0.071	0.426±0.083	0.605±0.016
	$\lambda = 0.4$	0.824±0.001	0.862±0.012	0.063±0.001	3.614±0.124	0.528±0.042	0.505±0.033
	$\lambda = 0.6$	0.826±0.001	0.825±0.004	0.064±0.001	2.677±0.049	0.495±0.058	0.426±0.031
	$\lambda = 0.8$	0.825±0.000	0.793±0.027	0.063±0.001	2.634±0.061	0.640±0.067	0.463±0.046
	$\lambda = 1.0$	/					
CEBRA	–	0.709±0.005	0.735±0.009	0.022±0.001	5.288±0.196	0.318±0.003	0.446±0.011

Continued on next page.

Table 17 (continued from previous page)

Method	Parameter	Trust ↑	Cont. ↑	10-NN ↑	$10^2 \times \mathrm{DE} \downarrow$	$\Phi_\sigma \downarrow$	M_{ctr}
TP-SE	$\lambda = 0.0$	0.814±0.001	0.938±0.008	0.046±0.002	11.810±0.240	0.208±0.013	0.795±0.010
	$\lambda = 0.2$	0.814±0.001	0.927±0.013	0.049±0.003	10.890±0.730	0.202±0.006	0.778±0.019
	$\lambda = 0.4$	0.814±0.001	0.893±0.020	0.050±0.001	9.525±0.558	0.319±0.088	0.744±0.011
	$\lambda = 0.6$	0.805±0.002	0.840±0.014	0.048±0.001	7.196±0.407	0.398±0.053	0.698±0.018
	$\lambda = 0.8$	0.800±0.001	0.783±0.022	0.046±0.000	6.578±0.143	0.547±0.101	0.671±0.010
	$\lambda = 1.0$	0.741±0.008	0.717±0.011	0.030±0.002	5.239±0.150	0.502±0.119	0.585±0.023
RF-PHATE	–	0.794±0.000	0.827±0.006	0.021±0.000	9.500±0.086	0.762±0.010	0.529±0.001
RF-AE	–	0.796±0.001	0.781±0.040	0.027±0.000	8.873±0.123	0.455±0.019	0.605±0.009
AKRMap	$\lambda = 0.0$	0.812±0.000	0.957±0.002	0.048±0.001	11.129±0.151	0.166±0.009	0.778±0.004
	$\lambda = 0.2$	0.804±0.002	0.949±0.001	0.045±0.001	5.834±0.402	0.162±0.004	0.770±0.006
	$\lambda = 0.4$	0.798±0.001	0.938±0.002	0.041±0.001	4.484±0.516	0.164±0.004	0.764±0.009
	$\lambda = 0.6$	0.792±0.001	0.920±0.001	0.038±0.001	3.848±0.230	0.160±0.002	0.749±0.008
	$\lambda = 0.8$	0.773±0.008	0.873±0.009	0.032±0.003	3.554±0.406	0.172±0.012	0.726±0.012
	$\lambda = 1.0$	0.711±0.031	0.698±0.013	0.025±0.004	3.495±0.710	0.309±0.059	0.676±0.015
ISDR	$\lambda = 0.0$	0.850±0.000	0.942±0.001	0.098±0.000	9.814±0.055	0.173±0.002	0.794±0.002
	$\lambda = 0.2$	0.845±0.001	0.942±0.001	0.095±0.001	4.949±0.047	0.174±0.003	0.785±0.002
	$\lambda = 0.4$	0.838±0.001	0.942±0.001	0.089±0.000	2.101±0.037	0.179±0.001	0.774±0.002
	$\lambda = 0.6$	0.832±0.001	0.939±0.001	0.082±0.000	1.138±0.036	0.193±0.002	0.768±0.002
	$\lambda = 0.8$	0.827±0.001	0.934±0.001	0.076±0.001	0.848±0.069	0.209±0.003	0.762±0.004
	$\lambda = 1.0$	0.819±0.001	0.917±0.002	0.051±0.001	0.375±0.021	0.179±0.001	0.775±0.001

G Additional Discussion

A common concern in DR-based visualization is misuse or overinterpretation: approximate visual patterns can lead to unsupported conclusions [1, 2]. Our work addresses this concern by controlling geometric distortion while exposing task-relevant scalar structure. In the case study, however, we found that interpreting faithful patterns can itself become challenging. Many studies of supervised DR and its visualization applications focus on tasks such as image classification, where each data point corresponds to a single image with a label. When analysts find a pattern in the DR visualization, they can inspect it by sampling images from the corresponding region. However, in the LLM-as-a-Judge task, each data point corresponds to a judgment comprising a criterion, an image generation prompt, a pair of images, and a judgment result. Understanding a single sample requires examining these components together, making regions harder to interpret. Recent work [76, 77] helps relate patterns in projections to the original high-dimensional features in classification tasks. As representations grow more abstract and complex, helping analysts interpret patterns in DR visualizations becomes increasingly important and challenging.

Parametric and out-of-sample extensions. The NW regularizer could be applied to an encoder's two-dimensional output, such as an autoencoder bottleneck. This requires maintaining layout-local supports during training, which ordinary random minibatches do not reproduce. One possible approach is an iterative, diffusion-inspired mapper that gradually refines the layout while maintaining local supports, with its timestep tracking progress from the initial layout ($t = 0$) to the refined layout ($t = 1$). Training convergence and out-of-sample generalization for such a model remain open.

Scalar signals. We focus on scalar signals displayed through colormaps [78, 54]. For multidimensional signals, visual encodings such as glyphs, small multiples, or interaction expose different relationships and require corresponding readability objectives. Preserving relationships in both the input and signal spaces also connects this extension to multi-view DR [79].

Broader impact. This work may facilitate the exploration of continuous signals in high-dimensional data while preserving local geometric context, supporting applications such as model evaluation and scientific analysis. However, DR visualizations may still encourage overinterpretation of uncertain or misleading patterns, so findings should be validated against original data and task-specific evidence.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the method, theory, and benchmark claims, and the scope is supported by Section 4, Section 6, and Appendix C.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See section 7.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: Section 4 states the theoretical results and Appendix A gives the assumptions and formal proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 6 and Appendices B–C disclose datasets, seeds, metrics, hyperparameters, sweep grids, and full canonical result tables for the main claims.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code and public-data reproduction instructions are available at `https://github.com/Circular-Aflame/ISDR`. Appendix B and Appendix B.2 give the method settings and data sources.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The experimental setup, data splits, metrics, method hyperparameters, and sweep grids are specified in Section 6 and Appendices B–C.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Main and appendix tables report mean $\pm$ standard deviation over five fixed-data seeds for the canonical experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix B.1 reports the server hardware, memory, and feature-caching setup; Appendix C.8 gives per-method wall-clock times on HPDv3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is a visualization-method study using public or research benchmark data and does not introduce deceptive, privacy-invasive, or harmful deployment mechanisms.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix G discusses the potential positive use of the method as an exploratory aid for inspecting scalar signals on high-dimensional representations, as well as the main negative risk of over-interpreting approximate visual patterns or developing unwarranted confidence in proxy-dependent structure.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a new high-risk model, image generator, language model, or scraped dataset requiring special release safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Appendix B.2 lists the public source, citation, and license or terms for HPDv3/HPSv3, Qwen3-VL-8B-Instruct, Tabula Muris FACS, MNIST, Fashion-MNIST, and the used MedMNIST subsets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.

- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The ISDR implementation is released under the MIT license at `https://github.com/Circular-Aflame/ISDR`, with installation, dataset preparation, benchmark, and ablation instructions.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The case study is an internal data analysis by the research team, with no recruited participants or crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The case study analyzes existing public data and does not involve a human-subject study.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Qwen3-VL-8B-Instruct is used as an existing benchmark model to construct HPDv3 representations and judge-confidence signals; it is not part of the proposed optimization method. This use is documented in Section 6, Appendix B.2 and Appendix D.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.