WHAT CAN WE INFER FROM T-SNE, UMAP, AND OTHER NONLINEAR 2D MAPS OF HIGH-DIMENSIONAL REPRESENTATIONS?

Anonymous authors
Paper under double-blind review

ABSTRACT

A common practice in machine learning and scientific data analysis is to reduce a high-dimensional representation to two dimensions using methods such as t-SNE or UMAP, then color the resulting map by semantic variables—cell type, digit identity, topic label—and read the visible organization as evidence about what the representation contains. A variable that forms clear islands in 2D may be interpreted as strongly represented, while one that appears mixed may be taken as weak or absent. We show that this backward inference can break in ways not captured by geometric fidelity alone: a variable carrying more information in the high-dimensional representation can appear less structured than a weaker one on the same fixed map — the visual ranking can reverse the information ranking. Furthermore, the mixed appearance itself is not diagnostic: nearly identical coarse 2D mixing statistics can arise from opposite underlying information states, one in which all information carried by the original noisy Y-channel is retained and another in which it is entirely absent. These results reveal a decoupling between information present before mapping and semantic expression after it. At the same time, nonlinear 2D maps can retain useful semantic structure: class-conditioned neighborhood relations — which classes tend to occur near one another — can remain similar even when most individual neighbors change. This semantic stability can coexist with substantial reorganization of between-class geometry and class-local persistence profiles. Taken together, these results highlight that semantic patterns in 2D can be informative about selected relationships in a high-dimensional representation without faithfully encoding its information content or geometry. A 2D map can therefore provide a useful view of a semantic relation without establishing the strength of the underlying information, the geometry from which it arose, or the broader scientific interpretation attached to it.

1 INTRODUCTION

High-dimensional representations can encode semantically meaningful information, but their structure is difficult to inspect directly. To make complex structure visually accessible, researchers routinely apply nonlinear dimensionality-reduction methods such as t-SNE (van der Maaten & Hinton, 2008), UMAP (McInnes et al., 2018), PaCMAP (Wang et al., 2021), and TriMap (Amid & Warmuth, 2019) to transform high-dimensional representations into two-dimensional views for visual analysis. Throughout, we refer to this class as *nonlinear dimensionality-reduction methods*. The resulting fixed coordinates are then colored by variables such as class identity, cell type, disease state, or topic (Luecken & Theis, 2019; Clarke et al., 2021; Ahn et al., 2021; Rives et al., 2021; De Donno et al., 2023; Ovek et al., 2024; Fraikin et al., 2024; Fifty et al., 2024; Cashman et al., 2025; Yu et al., 2025; Rohbeck et al., 2025). Clear islands or separation are then used to reason about what the high-dimensional representation encodes, whereas mixed colorings may suggest weak or absent structure. We call this dimensionality reduction followed by semantic interpretation of the fixed map the *post-hoc workflow*. The central question is what such a 2D view actually allows us to infer about the high-dimensional representation that produced it. MNIST (LeCun et al., 1998)—a standard benchmark of 28×28 handwritten-digit images — makes the problem concrete. We compute a single UMAP embedding of the 784-dimensional pixel representation and color the same fixed

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

coordinates by two semantic variables (Figure 1). Digit identity produces comparatively clear regions, whereas ordered bins of total ink mass show a more diffuse spatial organization. The key contrast is absolute: ink mass is directly available from the original pixels through the linear statistic $m(\mathbf{x}) = \sum_{u=1}^{784} x_u$, and its ordered bins are exactly linearly realizable in the original pixel space (Appendix B.7). Thus, exact linear realizability in the source can coexist with weak visual expression in 2D. The clear digit regions raise a further question: even if a declared semantic relation among the digit classes is shown to survive the mapping, does that imply that the geometry realizing it was preserved as well, or could the same relation appear through a substantially different geometry in 2D?

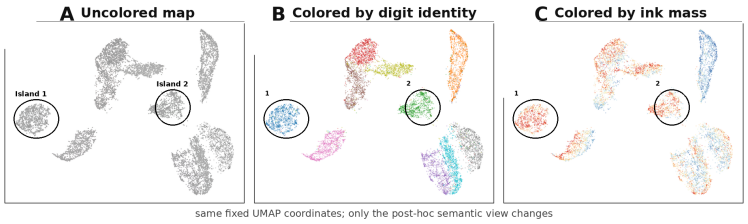


Figure 1: **One fixed MNIST map, two semantic readings. A.** Uncolored UMAP coordinates. **B.** Digit identity forms comparatively clear regions. **C.** Ordered ink-mass deciles show a more diffuse spatial organization on the same fixed coordinates. No point moves across panels.

Prior work has largely focused on the *forward* question—which properties of a high-dimensional representation survive nonlinear dimensionality reduction, and how much geometric quantities such as distances, neighborhoods, and cluster structure change in the process (Venna & Kaski, 2006; Lee & Verleysen, 2009; Rieck & Leitte, 2015; Wattenberg et al., 2016; Böhm et al., 2022; Jeon et al., 2022; Chari & Pachter, 2023; Jeon et al., 2024; Snoeck et al., 2026). The picture is mixed: distortions can be substantial, yet 2D views can retain genuinely useful semantic signal—visual separation correlates with classifier performance (Rauber et al., 2018; Benato et al., 2023; 2024), and label-level quantities such as kNN accuracy can remain informative even when exact neighborhoods are substantially distorted (Lause et al., 2024). Geometric distortion is only one part of the story. These results establish that 2D maps can retain useful semantic signal. We study the complementary *backward* question in post-hoc interpretation: when I color the same 2D map by different variables of interest, what can the resulting visual patterns actually tell me about the high-dimensional representation from which the map was produced? We separate this backward inference into three questions, each building on the last.

Q_1**: Does visual expression track source information?** When one variable forms clear regions and another looks mixed on the same fixed 2D map, can we infer that the first carries more information in the high-dimensional representation? Under controlled conditions, we show that this inference can fail: the information ranking of two variables can strictly reverse on the same fixed map, and nearly equal mixing statistics can accompany opposite high-dimensional information states—one retaining all information carried by the original noisy Y-channel, the other containing none.

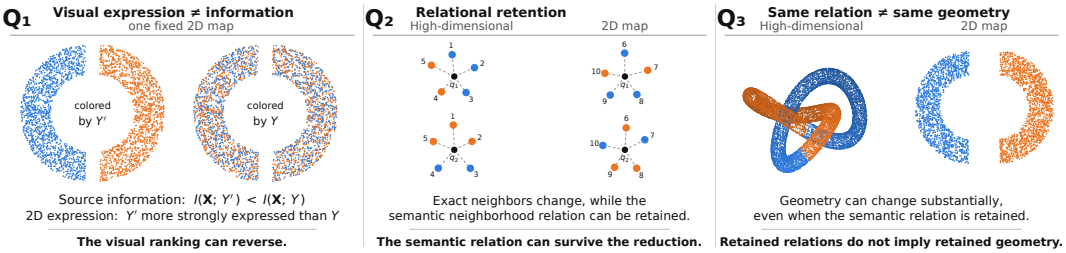


Figure 2: **Conceptual overview of** Q_1**–**Q_3**.** (Q_1) Visual expression need not rank semantic variables by the information they carry in the high-dimensional representation. (Q_2) Exact neighbor identities may change while a declared semantic neighborhood relation is retained. (Q_3) Retention of that relation does not imply preservation of the geometry through which it is realized.

Q_2: **What survives at the local level?** Setting aside whole-map appearance, we examine structure at a different scale: individual neighborhoods. Even when most exact neighbors are replaced, can the class composition of those neighborhoods—which classes tend to occur near which—remain stable? Across controlled and natural settings, we find that it can: semantic neighborhood relations remain substantially more stable than the identities of the neighbors carrying them.

Q_3: **Does a surviving relation imply surviving geometry?** When such a semantic relation does transfer intact, is the geometry through which it is realized also preserved—or can the map reorganize separation and internal structure while keeping the relation the same? We show that this decoupling can be substantial: between-class geometry and within-class structure can both be substantially reorganized while the semantic relation remains close.

Figure 2 summarizes this chain. We establish the central contrasts with controlled constructions and examine complementary evidence across thirteen natural conditions spanning vision, text, single-cell genomics, and tabular data. Section 4 marks two boundary cases of the post-hoc workflow. Prior work shows that a 2D map need not preserve source geometry; we add the backward result that visual expression can even reverse the source-information ranking. Thus, a 2D map may usefully display a semantic relation without, by itself, determining information strength, source geometry, or scientific meaning.

2 The Post-hoc Workflow

In the workflow we study, a high-dimensional representation is reduced to two dimensions without reference to any semantic variable; only afterward are the fixed coordinates colored by that variable and interpreted. Schematically:

$$\boxed{\begin{array}{ll} \mathcal{O} \xrightarrow{f} \mathbf{X} \xrightarrow[\mathcal{L}_g,\xi]{g\in\mathcal{M}} \mathbf{Z} & \text{reduction without } Y \\ (\mathbf{Z}, Y) \longrightarrow \text{colored 2D view} \longrightarrow \text{semantic interpretation} & \text{post-hoc interpretation.} \end{array}} \tag{1}$$

The observed data $\mathcal{O}$ are first transformed by f into a high-dimensional representation $\mathbf{X} \in \mathbb{R}^{n\times d_X}$, where f may be the identity (raw observations), a standard preprocessing step such as scaling or PCA, or a learned encoder. A nonlinear dimensionality-reduction method g—here t-SNE (van der Maaten & Hinton, 2008), UMAP (McInnes et al., 2018), PaCMAP (Wang et al., 2021), or TriMap (Amid & Warmuth, 2019)—is then applied to produce two-dimensional coordinates $\mathbf{Z} \in \mathbb{R}^{n\times 2}$ by approximately minimizing a method-specific objective $\mathcal{L}_g(\widetilde{\mathbf{Z}}; \mathbf{X}, \xi)$, where $\widetilde{\mathbf{Z}}$ denotes a candidate 2D realization, ξ collects method settings and randomization, and $\mathcal{M} = \{g_{\mathrm{tSNE}}, g_{\mathrm{UMAP}}, g_{\mathrm{PaCMAP}}, g_{\mathrm{TriMap}}\}$. Only after this map is fixed is a semantic variable Y attached and its visual expression read from the coloring. Crucially, Y is not an argument of $\mathcal{L}_g(\widetilde{\mathbf{Z}}; \mathbf{X}, \xi)$: the reduction can expose structure in $\mathbf{X}$ without being calibrated to the information $\mathbf{X}$ carries about Y. Whether the representation $\mathbf{Z}$ tracks that information is what we test with Q_1. The specific objectives and experimental settings for all four methods are given in Appendix A.4. Throughout, bold uppercase denotes realized data matrices ($\mathbf{X}$, $\mathbf{Z}$), plain uppercase the corresponding random vectors ($X \in \mathbb{R}^{d_X}$, $Z \in \mathbb{R}^2$), and bold lowercase individual observations ($\mathbf{x}_i$, $\mathbf{z}_i$), with $\mathbf{X} = [\mathbf{x}_1^\top; \ldots; \mathbf{x}_n^\top]$ and $\mathbf{Z} = [\mathbf{z}_1^\top; \ldots; \mathbf{z}_n^\top]$. A single mapper is denoted $g \in \mathcal{M}$, with $\mathcal{M}$ the family of audited methods.

Information in the high-dimensional representation. To test whether visual expression tracks information in the high-dimensional representation—and to later separate information content from geometric organization—we must quantify how much information about Y is present in $\mathbf{X}$ before the reduction. We use mutual information $I(X;Y) = \mathrm{H}(Y) - \mathrm{H}(Y \mid X)$, where H denotes entropy, when the data-generating process is known, and the held-out balanced accuracy of a linear logistic readout, $\mathrm{Lin}(\mathbf{X}; Y)$, as a separate accessibility diagnostic. The first captures all decodable information regardless of the complexity of the decoder; the second tests whether a linear rule suffices to recover it. Details are given in Appendices A and B.

Semantic expression in the fixed 2D map. The post-hoc workflow is ultimately evaluated visually: a practitioner judges if Y appears spatially organized in the colored map. For the controlled Q_1 experiments, we operationalize this judgment with a deliberately coarse label-blind readout that measures how strongly Y aligns with a declared spatial partition of the map. Let $\Pi(\mathbf{Z}) = \{\mathcal{P}_1, \ldots, \mathcal{P}_L\}$

be a label-blind partition of the uncolored coordinates and $C_{\mathbf{Z}}(i) \in \{1, \ldots, L\}$ the region label of observation i. For a K-category variable Y, define

$$\mathrm{Mix}(\mathbf{Z}; Y) = \frac{\widehat{\mathrm{H}}(Y \mid C_{\mathbf{Z}})}{\log K}. \tag{2}$$

Here $\widehat{\mathrm{H}}$ denotes the empirical plug-in conditional entropy computed from the observed region-by-class counts. For a given semantic variable Y, Mix measures how strongly its labels align with the spatial organization of the map: it is zero when each region contains a single class and near-maximal when classes are spread uniformly across regions. For these controlled analyses, this provides a quantitative proxy for the coarse spatial organization visible in the colored 2D view. The natural-data examples below are complementary descriptive illustrations and do not use Mix as a calibrated perceptual measure. Implementation details and the balanced-class information identity are given in Appendix B.2.

3 Controlled Tests of Q_1–Q_3

In real-world data, the contributions of individual components to a representation are entangled, making it difficult to trace how high-dimensional semantic information is expressed on a 2D map. We therefore use a controlled generative model in which the influence of information content and metric can be varied independently. The controlled encoding is structurally simpler than natural representations — we show that even this favorable case suffices for backward inference to fail.

3.1 Controlled representation and interventions

The synthetic representation concatenates four blocks, $\mathbf{X}_{\mathrm{syn}} = [\mathbf{X}_{\mathrm{topo}}, \mathbf{X}_Y, \mathbf{X}_{Y'}, \mathbf{X}_{\mathrm{noise}}]$. A noisy torus $\mathbf{X}_{\mathrm{topo}}$ supplies geometric structure but carries no information about the semantic variables. Two independent semantic channels, $\mathbf{X}_Y \in \mathbb{R}^{n\times 5}$ and $\mathbf{X}_{Y'} \in \mathbb{R}^{n\times 3}$, encode Y and Y' as class-specific Gaussian clusters. Isotropic noise $\mathbf{X}_{\mathrm{noise}} \in \mathbb{R}^{n\times 10}$ fills the remaining dimensions. By construction, Y, Y', and torus position are mutually independent, so the information the representation contains about each semantic variable is cleanly separated. The key experimental lever is scaling. Multiplying a block by a scalar α changes its contribution to the metric structure of the input but does not change the information it carries, so this scaling lets us decouple what the mapper sees from what the representation knows. Setting $\alpha = 0$ removes the block and its information entirely. Each experimental condition τ is defined by a choice of scales $(\alpha_Y, \alpha_{Y'}, \alpha_{\mathrm{noise}})$; observations, labels, class centers, and noise draws remain identical across conditions, so with the mapping settings ξ held fixed, differences in $\mathbf{Z}^{(\tau)} = g(\mathbf{X}_{\mathrm{syn}}^{(\tau)}; \xi)$ are induced by the rescaling. This design lets us independently control what semantic information is present in the representation and how strongly it influences the metric structure seen by the reduction. The practitioner in the post-hoc workflow, by contrast, observes only the finished 2D map and must reason backward. This asymmetry is exactly what the controlled setting is designed to expose: we can measure how far that backward inference can depart from the known high-dimensional information state. The specific conditions are introduced where each question is tested below; the complete generator specification and scale choices are recorded in Appendix A.2. For every controlled mutual-information value below, we condition on the realized collection of sampled semantic class centers $\boldsymbol{\mu}$; to keep the main text light, we write $I(\cdot;\cdot)$ for the conditional quantity $I_{\boldsymbol{\mu}}(\cdot;\cdot)$. Appendix A.3 gives the corresponding conditional population law and numerical evaluation convention.

3.2 Q_1: Visual Expression and Information

Practitioners routinely read clear regions in a colored 2D map as evidence that a variable is strongly represented in $\mathbf{X}$, and a mixed coloring as evidence that it is weak or absent. Across three controlled experiments, we show how strongly both readings can diverge from the information the high-dimensional representation actually carries about the variable of interest.

Experiment 1: information present in high dimensions need not be visually expressed (Figure 3A). In the Redirect condition, both semantic channels are retained but their metric influence is set to be very different: $\alpha_Y = 0.1$ and $\alpha_{Y'} = 10$. This makes Y metrically near-invisible and amplifies Y', without removing either variable from the representation. Before mapping, the information

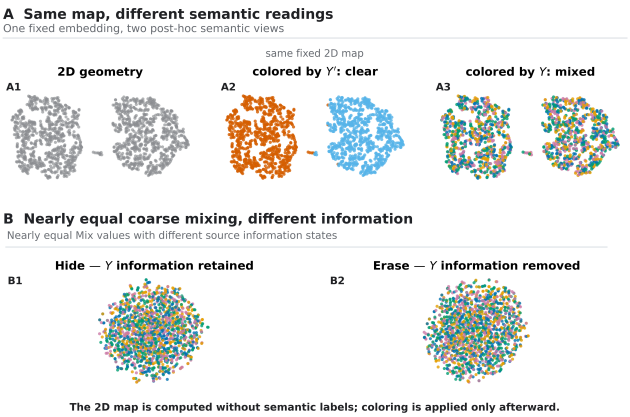


Figure 3: **Controlled Q_1 contrasts. A:** Strongly represented Y appears mixed while Y' is clearly expressed on the same map. **B:** Nearly identical coarse expression accompanies retained versus absent Y-information.

carried by each variable is

$$I(X_{\text{syn}}^{(\text{Redirect})}; Y) = 1.966 \text{ bits}, \qquad I(X_{\text{syn}}^{(\text{Redirect})}; Y') = 0.995 \text{ bit}.$$

Both variables also have high held-out linear-readout accuracy: $\text{Lin}(\mathbf{X}_{\text{syn}}^{(\text{Redirect})}; Y) = .991$ and $\text{Lin}(\mathbf{X}_{\text{syn}}^{(\text{Redirect})}; Y') = .997$, confirming that the information is not only present but linearly accessible. *Single-variable conclusion.* The representation carries nearly two bits of information about Y and makes it almost perfectly linearly accessible, yet its coarse 2D expression is nearly maximally mixed:

$$\text{Mix}(\mathbf{Z}^{(\text{Redirect})}; Y) = 0.997.$$

A practitioner inspecting only the Y coloring would conclude the representation carries little information about the variable, when in fact it carries nearly the theoretical maximum. *Same-map comparison.* Coloring the same frozen coordinates by the competing semantic variable gives

$$\text{Mix}(\mathbf{Z}^{(\text{Redirect})}; Y') = 0.029.$$

On the same map, strongly represented Y remains nearly maximally mixed while Y' is clearly expressed—visual expression can differ sharply even among informative variables. MNIST provides a natural illustration of this absolute phenomenon: ink-mass deciles are exactly linearly realizable from the original pixels, yet their coloring is more diffuse than digit identity on the same fixed UMAP (Figure 1; Appendix B.7). Bibtex provides a complementary same-map multi-view example in which both semantic views remain nontrivially accessible in the source representation (Appendix B.8). Experiment 1 thus establishes the absolute failure mode; Experiment 2 matches cardinality and tests whether the source-information ordering itself can reverse.

Experiment 2: the variable carrying more information can look less expressed in 2D. Experiment 1 establishes an *absolute* failure mode: strongly present information can look weak in isolation. Experiment 2 tests the complementary *relative* failure: can comparing two variables on the same map reverse their source-information ordering? Because the original Y and Y' differ in cardinality, we use the matched-cardinality *Rank* construction: both are independent four-class variables with equal five-dimensional signal blocks, giving a 23-dimensional representation. Y is generated with a larger Gaussian center-radius parameter ($\delta_Y = 6$ versus $\delta_{Y'} = 3$), and the same Redirect scales $\alpha_Y = 0.1$ and $\alpha_{Y'} = 10$ suppress Y while amplifying Y'. Across five independently generated datasets, each mapped with three t-SNE seeds (15 maps total), Y consistently carries more information:

$$I(X_{\text{syn}}^{(\text{Rank})}; Y) = 1.985 \pm .021 \text{ bits} \; > \; I(X_{\text{syn}}^{(\text{Rank})}; Y') = 1.605 \pm .227 \text{ bits}.$$

The linear readout confirms the same ordering: $\text{Lin}(\mathbf{X}_{\text{syn}}^{(\text{Rank})}; Y) = .995 \pm .008$ versus $\text{Lin}(\mathbf{X}_{\text{syn}}^{(\text{Rank})}; Y') = .887 \pm .062$.

On the frozen 2D maps, this ordering reverses in every one of the fifteen 5×3 data/map realizations:

$$\mathrm{Mix}(\mathbf{Z}^{(\mathrm{Rank})}; Y) = .9978 \pm .0006 \;>\; \mathrm{Mix}(\mathbf{Z}^{(\mathrm{Rank})}; Y') = .325 \pm .119 \qquad \text{(lower is clearer)}.$$

The variable carrying more information in high dimensions consistently looks less expressed after dimensionality reduction—a practitioner reading the map would infer the opposite of the actual information ranking.

Experiment 3: nearly equal coarse expression can accompany retained or absent information in the high-dimensional representation (Figure 3B). Experiments 1 and 2 show that information can be present in the high-dimensional representation despite weak coarse expression. We next test whether near-maximal coarse mixing distinguishes information that remains present from information that has been removed. Hide and Erase use the matched scale tuples $(\alpha_Y, \alpha_{Y'}, \alpha_{\mathrm{noise}}) = (.1, 1, 10)$ and $(0, 1, 10)$, respectively. In Hide, the original noisy Y-channel is retained under an invertible rescaling, so all information carried by that channel is preserved. In Erase, the only Y-dependent channel is zeroed, so the representation is independent of Y under the controlled population model.

$$I(X_{\mathrm{syn}}^{(\mathrm{Hide})}; Y) = 1.966 \text{ bits}, \qquad I(X_{\mathrm{syn}}^{(\mathrm{Erase})}; Y) = 0.$$

Applying the same label-free mapping procedure and the same declared four-region partition rule to the two conditions yields nearly equal coarse mixing statistics:

$$\mathrm{Mix}(\mathbf{Z}^{(\mathrm{Hide})}; Y) = 0.9980, \qquad \mathrm{Mix}(\mathbf{Z}^{(\mathrm{Erase})}; Y) = 0.9976.$$

Thus, in this construction, the declared coarse readout assigns nearly the same expression to two opposite source-information states. This result concerns the declared coarse statistic; the complete embeddings may differ. All three controlled contrasts show the same qualitative pattern under t-SNE, UMAP, PaCMAP, and TriMap with predeclared settings and no target-specific tuning (Appendix Table 4).

Result 1 (Q_1): visual expression is not a calibrated measure of source information.

A variable carrying strong information in the high-dimensional representation can appear weakly expressed in 2D, and the information ranking of two variables can strictly reverse on the same map. Consequently, a plausible visual interpretation in 2D may lead to incorrect conclusions about the underlying high-dimensional representation.

$$\boxed{\text{visual readout of 2D map} \;\not\Rightarrow\; \text{information present in } \mathbf{X}}$$

The supporting calculations, four-mapper breadth audit, and natural-data constructions are given in Appendix B.

3.3 Q_2: Semantic Relation Retention

Q_1 shows that whole-map coloring is an unreliable guide to information in the high-dimensional representation. Semantic structure may still survive at a more local level. To understand which aspects of local structure survive the reduction, we inspect two levels: observation identity—do the same points remain nearby after mapping—and semantic composition—do the classes represented nearby remain similar, even when the specific points change. Let $\mathbf{X}$ denote the high-dimensional representation and $\mathbf{Z}$ its 2D embedding. For observation $i \in [n] := \{1, \ldots, n\}$, let $\mathrm{NN}_A^k(i)$ denote its k nearest neighbors in space $A \in \{\mathbf{X}, \mathbf{Z}\}$. We measure neighbor identity change by the exact-neighbor replacement rate

$$R_{\mathrm{nbr}} = 1 - \frac{1}{n} \sum_{i=1}^{n} \frac{\left| \mathrm{NN}_{\mathbf{X}}^k(i) \cap \mathrm{NN}_{\mathbf{Z}}^k(i) \right|}{k}.$$

$R_{\mathrm{nbr}} = 0$ means every neighbor is preserved; $R_{\mathrm{nbr}} = 1$ means every neighbor has been replaced. Identity, however, is only one level of description: different observations may replace one another

while preserving a similar semantic composition. We capture this second level by the directed class-conditioned relation

$$\mathbf{T}_A(c, r) := \Pr(Y_J = r \mid Y_U = c), \qquad U \sim \mathrm{Unif}([n]),\ J \mid U \sim \mathrm{Unif}\big(\mathrm{NN}_A^k(U)\big). \tag{3}$$

Each row $\mathbf{T}_A(c, \cdot)$ answers: given an observation of class c, which classes tend to occur among its neighbors? The diagonal captures same-class concentration; the off-diagonal entries retain which other classes tend to occur nearby. Equivalently, $\mathbf{T}_A(c, r)$ is the mean over class-c points of the fraction of neighbors belonging to class r; Appendix C.1 gives the pointwise-vector and aggregation forms. Because the same relation is defined in both spaces, we can now ask directly how it changes under dimensionality reduction. We write $\mathbf{T}_\mathbf{X}$ for the relation in $\mathbf{X}$ and $\mathbf{T}_\mathbf{Z}$ for the relation in $\mathbf{Z}$, and compare them by average row-wise total variation,

$$\mathrm{D}_{\mathrm{comp}}(\mathbf{T}_\mathbf{X}, \mathbf{T}_\mathbf{Z}) = \frac{1}{K}\sum_{c=1}^{K} \mathrm{TV}\big(\mathbf{T}_\mathbf{X}(c, \cdot), \mathbf{T}_\mathbf{Z}(c, \cdot)\big), \qquad \mathrm{TV}(\mathbf{p}, \mathbf{q}) := \frac{1}{2}\|\mathbf{p} - \mathbf{q}\|_1.$$

Small $\mathrm{D}_{\mathrm{comp}}$ means that the class-conditioned semantic relation remains close across spaces on average over conditioning classes. The Q_2 audit therefore compares two effects of the same mapping: R_{nbr} measures how strongly exact neighbor identity changes, whereas $\mathrm{D}_{\mathrm{comp}}$ measures how strongly the semantic relation carried by those neighborhoods changes. Because $\mathrm{D}_{\mathrm{comp}}$ is a macro-average, class-wise deviations should be inspected separately when class imbalance is material. **Empirical result.** In the controlled torus, 52–68% of exact neighbors are replaced across the four audited mappers, yet the class-conditioned semantic relation remains nearly unchanged ($\mathrm{D}_{\mathrm{comp}} = 0.0046$; Appendix Table 9). The complete natural cross-mapper audit contains $13 \times 4 = 52$ dataset–method cells (Appendix Table 7): 38/52 have $\mathrm{D}_{\mathrm{comp}} < 0.10$, and across all cells the mean relation discrepancy is $\mathrm{D}_{\mathrm{comp}} = .091 \pm .075$ while mean exact-neighbor overlap is only $.278 \pm .143$. The audit also contains counter-regimes in which the semantic relation is not well retained. We call the recurring pattern—substantial change in neighbor identity alongside much smaller change in the class-conditioned semantic relation—*relational retention*.

Result 2 (Q_2): semantic relations can survive extensive neighborhood reconstruction. Across controlled and natural settings, dimensionality reduction can replace many exact neighbors while leaving the directed class-conditioned semantic relation substantially more stable.

$$\mathrm{NN}_\mathbf{X}^k(i) \neq \mathrm{NN}_\mathbf{Z}^k(i) \ \not\Rightarrow\ \mathbf{T}_\mathbf{X} \not\approx \mathbf{T}_\mathbf{Z}$$

Relational retention is condition-dependent; Section 4 presents a complementary low-retention regime that marks the boundary of this behavior.

3.4 Q_3: Geometry After Relation Retention

Q_2 highlights that a class-conditioned semantic relation can survive dimensionality reduction even when most neighbors are replaced. This raises a sharper question: when the semantic relation is retained, is the geometry realizing it also preserved—or can the map reorganize geometry while keeping the relation close? We compare geometry in $\mathbf{X}$ and $\mathbf{Z}$ from two complementary perspectives: *between classes*, how strongly are semantic groups separated from one another; and *within classes*, what is the internal geometric and persistence structure of each group. For a space $A \in \{\mathbf{X}, \mathbf{Z}\}$, we summarize these by

$$\mathcal{G}_A(Y) = \big(\mathrm{Sil}(A; Y),\ \mathrm{B/W}(A; Y),\ \mathrm{PH}_{\mathrm{intra}}(A; Y)\big).$$

Sil (silhouette score; Rousseeuw 1987) captures pointwise class separation, whereas B/W (between-class to within-class spread ratio) captures class-centroid separation relative to within-class scatter; together they provide complementary views of between-class geometry. $\mathrm{PH}_{\mathrm{intra}}$ summarizes within-class geometry through Vietoris–Rips persistent homology (Edelsbrunner et al., 2002) in homological degrees 0, 1, and 2, each normalized by the class's own internal scale. These quantities summarize the chosen filtration rather than directly counting geometric components, loops, or cavities. Formal definitions and the complete geometry audit are in Appendix D. The

controlled torus separates relation retention from geometric retention. On Baseline, the semantic relation stays close across all four mappers ($\mathrm{D}_{\mathrm{comp}} = 0.0046$), but separation amplifies sharply (Sil: $0.182 \to 0.523$; $\mathrm{B/W}$: $0.513 \to 3.45$ for UMAP; the four-method profile is in Appendix Table 9). An exact-$\mathrm{D}_{\mathrm{comp}} = 0$ witness sharpens this: even with a perfectly preserved relation, one realization can have one connected hard-10NN component per class while another splits each class into two disconnected hard-10NN components (Appendix D.3). Under the separate Q_3 source-to-2D geometry protocol, MNIST shows the same pattern: the digit-conditioned relation is retained ($\mathrm{D}_{\mathrm{comp}} = 0.049$), while the geometry through which it is expressed is strongly reorganized by the mapping: between-class separation amplifies (Sil: $0.042 \to 0.312$; $\mathrm{B/W}$: $0.258 \to 4.20$), while the class-local PH0/PH1/PH2 total-persistence summaries decrease substantially (PH0: $122.7 \to 10.8$; PH1: $4.825 \to .420$; PH2: $.933 \to .0009$; Appendix D.4). The ten-digit islands therefore express one retained aspect—their neighbor-class relations remain close to those in 784 dimensions—while their compactness and clean separation arise from the 2D realization. Under the separate Q_3 UMAP protocol, eleven of the thirteen natural dataset–representation conditions satisfy $\mathrm{D}_{\mathrm{comp}} < 0.10$. Across these retained-relation conditions, the class-local PH0/PH1/PH2 total-persistence summaries decrease in all eleven, while between-class geometry changes more heterogeneously across the complementary silhouette and $\mathrm{B/W}$ views (Appendix D.4). The complementary fixed-native and same-dimensional diagnostics are reported in Appendices D.5 and D.6. Conditions with low semantic-relation retention, such as raw-pixel CIFAR-10 and Paul15, define complementary boundary cases for Q_3.

Result 3 (Q_3): clear islands in 2D need not correspond to islands in the high-dimensional representation.

Even when the semantic relation is preserved, a clean 2D map need not be a window into the high-dimensional geometry. The mapping can amplify separation and substantially reorganize the measured class-local persistence profiles while keeping the relation close, so genuine semantic organization can appear through mapping-induced island geometry.

$$\mathbf{T_X} \approx \mathbf{T_Z} \not\Rightarrow \text{ islands in } \mathbf{Z} \text{ reflect islands in } \mathbf{X}$$

4 Limits and Boundary Cases

We highlight two consecutive boundary cases that clarify the evidential scope of the post-hoc workflow (Figure 4).

Low semantic-relation retention. Raw-pixel CIFAR-10 (Krizhevsky, 2009) provides a clear low-retention regime. Under the separate Q_3 source-to-2D protocol, the class-conditioned relation has $\mathrm{D}_{\mathrm{comp}} = 0.262$; the independent default-UMAP illustration in Figure 4A likewise shows the original classes strongly mixed. This marks the low-retention boundary for the conditional Q_3 analysis.

One way to handle this issue is to apply a transformation, $\mathcal{O} \xrightarrow{f} \mathbf{X} \xrightarrow{g} \mathbf{Z}$, exposing structure that is inaccessible in the raw space (Rives et al., 2021; Böhm et al., 2023). This can yield a much cleaner map—but raises a sharper question: once the visible organization is created upstream by f, what does that organization actually mean?

When learned structure requires external semantic validation. Figure 4B illustrates this second boundary with a supervised, in-sample existence construction: the encoder is trained and audited on the same stratified 5,000-image cohort (Appendix E.1). After learning $\mathbf{X} = f(\mathcal{O})$ and reducing with g_{UMAP}, the unlabeled map contains five compact, well-separated islands. Panel C recolors the same $\mathbf{Z}$ by the original CIFAR-10 classes, revealing that each island combines two classes from a deliberately arbitrary pairing while the original categories are interleaved within every island. A matched coherent control attains comparably favorable internal diagnostics (Appendix E.1), showing that scientific relevance requires evidence beyond clean islands and strong linear readout alone.

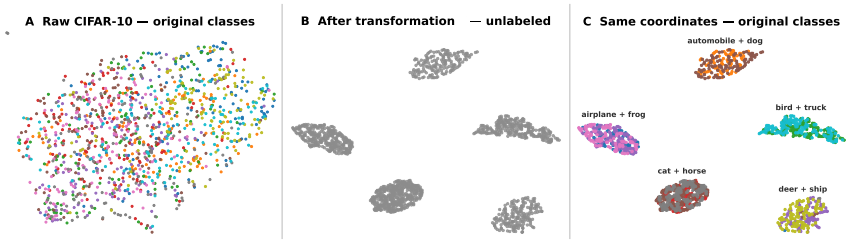


Figure 4: **Two limits of the embedding workflow. A:** In raw CIFAR-10 pixels, the original classes remain mixed. **B:** After supervised representation learning, the same data forms five clean islands. **C:** Recoloring by the original CIFAR-10 classes reveals that the islands reflect a deliberately imposed pairing, not the original categories—clean geometry alone does not establish semantic relevance.

5 Related Work

Geometric fidelity and inference from embeddings. A large body of work characterizes which properties of a high-dimensional representation survive dimensionality reduction: local neighborhood preservation and rank distortions (Venna & Kaski, 2006; Lee & Verleysen, 2009), correspondence between cluster structure in projections and in high dimensions (Jeon et al., 2022), distortions of neighborhoods, mixing, and inter-class relationships (Chari & Pachter, 2023), and the distinction between class and cluster structure across spaces (Jeon et al., 2024). Fundamental dimensional limits also constrain neighborhood preservation (Snoeck et al., 2026). Complementary inverse and recoverability work asks what properties of the input can be inferred from a low-dimensional output, including cluster strength from t-SNE (Bergam et al., 2026) and properties implied by UMAP outputs (Draganov & Dohn, 2023). These works ask whether geometry is preserved or recoverable; we ask whether visual expression in the map reliably tracks the information a representation carries about a post-hoc semantic variable. **Label-level structure despite geometric distortion.** Lause et al. (2024) show that label-level measures such as kNN accuracy can remain informative even with substantially distorted exact neighborhoods. Our Q_2 extends this direction by comparing the full directed class-destination distribution, including off-diagonal preferences, so that replacing a neighbor by one class is distinguished from replacing it by another. Q_3 then exploits verified relation retention to ask how much of the geometry realizing that relation is preserved — a question that label-level accuracy alone cannot resolve. **Constructing and interpreting semantic views.** Existing literature modifies or explains the visualization itself: conditional t-SNE discounts known structure to expose complementary patterns (Kang et al., 2021); contrastive approaches build representations designed to expose particular structure (Böhm et al., 2023); VERA explains regions of an existing map using high-dimensional attributes (Poličar & Zupan, 2024); and surveys document recurring interpretive failure modes (Cashman et al., 2025; Jeon et al., 2025; 2026). These works change or annotate the view, whereas we hold the realized map fixed and test what the subsequent semantic reading can support. **Positioning.** The forward literature establishes that a 2D map need not be a faithful geometric copy of its source; we identify a distinct backward decoupling in which the ranking suggested by visual expression can reverse the source-information ranking. Geometric fidelity and backward semantic inference therefore address complementary questions: one asks whether structure is preserved, while the other asks what visible patterns support about source information. Appendix E.2 provides a point-by-point comparison.

6 Conclusion

In the post-hoc workflow, practitioners reduce high-dimensional data to 2D to reason about what structure a representation contains. We have shown that this reasoning can fail sharply: visual expression can reverse the information ranking, semantic relations can survive substantial geometric reorganization, and clean 2D islands need not reflect high-dimensional geometry. A colored 2D map can describe which semantic relations are expressed after reduction, but cannot establish the information behind them, the geometry realizing them, or their scientific meaning. Each requires its own evidence.

Reproducibility Statement

The appendix specifies the controlled constructions, mapper settings, evaluation metrics, seed conventions, and analysis procedures used for the reported results. It also documents the natural-data inventory, the Q_1 information calculations and multi-seed/cross-mapper checks, the Q_2 relation audit, and the Q_3 geometry and persistent-homology analyses.

AI Use Statement

Generative AI tools were used for language editing (including paraphrasing and polishing), code development and refactoring, and to support literature review. The authors take responsibility for verification of the reported outputs, mathematical claims, and final manuscript.

References

10x Genomics. 3k PBMCs from a healthy donor. Single Cell Gene Expression public dataset, Cell Ranger 1.1.0, 2016. URL https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc3k.

Seoyoung Ahn, Gregory J Zelinsky, and Gary Lupyan. Use of superordinate labels yields more robust and human-like visual representations in convolutional neural networks. *Journal of Vision*, 21(13):13, 2021.

Ehsan Amid and Manfred K. Warmuth. TriMap: Large-scale dimensionality reduction using triplets. *CoRR*, abs/1910.00204, 2019.

Maayan Baron, Adrian Veres, Samuel L. Wolock, Aubrey L. Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K. Wagner, Shai S. Shen-Orr, Allon M. Klein, Douglas A. Melton, and Itai Yanai. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Systems*, 3(4):346–360.e4, 2016.

Bárbara C. Benato, Alexandre X. Falcão, and Alexandru C. Telea. Linking data separation, visual separation, classifier performance using multidimensional projections. In *Computer Vision, Imaging and Computer Graphics Theory and Applications*, volume 2103 of *Communications in Computer and Information Science*, pp. 229–255. Springer, 2024.

Bárbara Caroline Benato, Alexandre Xavier Falcão, and Alexandru-Cristian Telea. Linking data separation, visual separation, and classifier performance using pseudo-labeling by contrastive learning. In *VISIGRAPP (5: VISAPP)*, pp. 315–324. SCITEPRESS, 2023.

Noah Bergam, Szymon Snoeck, and Nakul Verma. t-SNE exaggerates clusters, provably. In *ICLR*. OpenReview.net, 2026. URL https://openreview.net/forum?id=rrbKiGMA5s.

Jan Niklas Böhm, Philipp Berens, and Dmitry Kobak. Attraction-repulsion spectrum in neighbor embeddings. *J. Mach. Learn. Res.*, 23:95:1–95:32, 2022.

Jan Niklas Böhm, Philipp Berens, and Dmitry Kobak. Unsupervised visualization of image datasets using contrastive learning. In *ICLR*. OpenReview.net, 2023.

Dylan Cashman, Mark S. Keller, Hyeon Jeon, Bum Chul Kwon, and Qianwen Wang. A critical analysis of the usage of dimensionality reduction in four domains. *IEEE Trans. Vis. Comput. Graph.*, 31(10):9405–9423, 2025.

Tara Chari and Lior Pachter. The specious art of single-cell genomics. *PLoS Comput. Biol.*, 19(8): e1011288, 2023.

Tarin Clanuwat, Mikel Bober-Irizar, Asanobu Kitamoto, Alex Lamb, Kazuaki Yamamoto, and David Ha. Deep learning for classical Japanese literature. *CoRR*, abs/1812.01718, 2018.

Zoe A. Clarke, Tallulah S. Andrews, Jawairia Atif, Delaram Pouyabahar, Brendan T. Innes, Sonya A. MacParland, and Gary D. Bader. Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. *Nature Protocols*, 16(6):2749–2764, 2021.

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

Carlo De Donno, Soroor Hediyeh-Zadeh, Amir Ali Moinfar, Marco Wagenstetter, Luke Zappia, Mohammad Lotfollahi, and Fabian J Theis. Population-level integration of single-cell datasets enables multi-scale analysis across samples. *Nature Methods*, 20(11):1683–1692, 2023.

Andrew Draganov and Simon Dohn. Relating tSNE and UMAP to classical dimensionality reduction. *CoRR*, abs/2306.11898, 2023. Version 2, revised 14 June 2024; first submitted 20 June 2023.

Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. *Discret. Comput. Geom.*, 28(4):511–533, 2002.

Christopher Fifty, Dennis Duan, Ronald G. Junkins, Ehsan Amid, Jure Leskovec, Christopher Ré, and Sebastian Thrun. Context-aware meta-learning. In *ICLR*. OpenReview.net, 2024.

Archibald Fraikin, Adrien Bennetot, and Stéphanie Allassonnière. T-Rep: Representation learning for time series using time-embeddings. In *ICLR*. OpenReview.net, 2024.

Nilsel Ilter and H. Altay Güvenir. Dermatology. `https://doi.org/10.24432/C5FK5P`, 1998. UCI-MLR.

Hyeon Jeon, Hyung-Kwon Ko, Jaemin Jo, Youngtaek Kim, and Jinwook Seo. Measuring and explaining the inter-cluster reliability of multidimensional projections. *IEEE Trans. Vis. Comput. Graph.*, 28(1):551–561, 2022.

Hyeon Jeon, Yun-Hsin Kuo, Michaël Aupetit, Kwan-Liu Ma, and Jinwook Seo. Classes are not clusters: Improving label-based evaluation of dimensionality reduction. *IEEE Trans. Vis. Comput. Graph.*, 30(1):781–791, 2024.

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 *CHI*, pp. 394:1–394:24. ACM, 2025.

Hyeon Jeon, Jeongin Park, Sungbok Shin, and Jinwook Seo. Stop misusing t-SNE and UMAP for visual analytics. *IEEE Trans. Vis. Comput. Graph.*, 2026. URL `https://arxiv.org/abs/2506.08725`. To be presented at IEEE VIS 2026.

Bo Kang, Darío García García, Jefrey Lijffijt, Raúl Santos-Rodríguez, and Tijl De Bie. Conditional t-SNE: more informative t-SNE embeddings. *Mach. Learn.*, 110(10):2905–2940, 2021.

Ioannis Katakis, Grigorios Tsoumakas, and Ioannis Vlahavas. Multilabel text classification for automated tag suggestion. In *ECML/PKDD 2008 Discovery Challenge*, pp. 75–83, 2008.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, volume 33, pp. 18661–18673, 2020.

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

Ken Lang. NewsWeeder: Learning to filter netnews. In *ICML*, pp. 331–339. Morgan Kaufmann, 1995.

Jan Lause, Philipp Berens, and Dmitry Kobak. The art of seeing the elephant in the room: 2D embeddings of single-cell data do make sense. *PLoS Comput. Biol.*, 20(10):e1012403, 2024.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, 1998.

John Aldo Lee and Michel Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*, 72(7-9):1431–1443, 2009.

Malte D. Luecken and Fabian J. Theis. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, 2019.

Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *CoRR*, abs/1802.03426, 2018.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Damla Ovek, Ozlem Keskin, and Attila Gursoy. ProInterVal: validation of protein–protein interfaces through learned interface representations. *Journal of Chemical Information and Modeling*, 64(8): 2979–2987, 2024.

Franziska Paul, Ya'ara Arkin, Amir Giladi, Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Deborah Winter, David Lara-Astiaso, Meital Gury, Assaf Weiner, Eyal David, Nadav Cohen, Felicia Kathrine Bratt Lauridsen, Simon Haas, Andreas Schlitzer, Alexander Mildner, Florent Ginhoux, Steffen Jung, Andreas Trumpp, Bo Torben Porse, Amos Tanay, and Ido Amit. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, 163 (7):1663–1677, 2015.

Pavlin G. Poličar and Blaž Zupan. VERA: generating visual explanations of two-dimensional embeddings via region annotation. *CoRR*, abs/2406.04808, 2024.

Paulo E. Rauber, Alexandre X. Falcão, and Alexandru C. Telea. Projections as visual aids for classification system design. *Inf. Vis.*, 17(4):282–305, 2018.

Bastian Rieck and Heike Leitte. Persistent homology for the evaluation of dimensionality reduction schemes. *Comput. Graph. Forum*, 34(3):431–440, 2015.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. USA*, 118(15):e2016239118, 2021.

Martin Rohbeck, Edward De Brouwer, Charlotte Bunne, Jan-Christian Huetter, Anne Biton, Kelvin Y. Chen, Aviv Regev, and Romain Lopez. Modeling complex system dynamics with flow matching across time and conditions. In *ICLR*. OpenReview.net, 2025.

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

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. CARER: contextualized affect representations for emotion recognition. In *EMNLP*, pp. 3687–3697. Association for Computational Linguistics, 2018.

Szymon Snoeck, Noah Bergam, and Nakul Verma. Compressibility barriers to neighborhood-preserving data visualization. In *ALT*, volume 313 of *Proceedings of Machine Learning Research*, pp. 1–30, 2026.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *J. Mach. Learn. Res.*, 9(86):2579–2605, 2008.

Jarkko Venna and Samuel Kaski. Local multidimensional scaling. *Neural Networks*, 19(6-7):889–899, 2006.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *CoRR*, abs/2212.03533, 2022.

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. *J. Mach. Learn. Res.*, 22:201:1–201:73, 2021.

Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-SNE effectively. *Distill*, 1(10): e2, 2016.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *CoRR*, abs/1708.07747, 2017.

Han Yu, Hanrui Lyu, YiXun Xu, Charlie Windolf, Eric Kenji Lee, Fan Yang, Andrew M. Shelton, Olivier Winter, International Brain Laboratory, Eva L. Dyer, Chandramouli Chandrasekaran, Nicholas A. Steinmetz, Liam Paninski, and Cole Lincoln Hurwitz. In vivo cell-type and brain region classification via multimodal contrastive learning. In *ICLR*. OpenReview.net, 2025.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NIPS*, pp. 649–657, 2015.

APPENDIX ROADMAP

Appendix A collects the shared experimental and analytical foundations used by more than one of Q_1–Q_3: notation, the controlled torus construction, the fixed-center Gaussian-channel convention, dimensionality-reduction settings, the natural-data inventory, and evaluation/reproducibility conventions. Appendix B supports Q_1 with the information-preservation/removal properties, the declared coarse Mix readout, the three controlled contrasts, a pinned four-mapper audit, the MNIST ink-mass illustration, and the Bibtex natural multi-view construction. Appendix C supports Q_2 with pointwise semantic composition, natural-data $\mathrm{D}_{\mathrm{comp}}$ breadth, the replacement decomposition, and locality controls. Appendix D supports Q_3 with geometry diagnostics, the controlled four-method profile, the exact-relation witness, the class-local persistent-homology audit, the fixed-native control, and the same-dimensional diagnostic. Appendix E collects the outer boundary case and positioning material, including matched CIFAR and the novelty map against neighboring work.

APPENDIX CONTENTS

Appendix A: Shared Experimental and Analytical Foundations 14
A.1 Notation convention 14
A.2 Controlled torus construction 14
A.3 Probability model and mutual-information convention 15
A.4 Nonlinear dimensionality-reduction methods and settings 15
A.5 Natural-data panel 18
A.6 Evaluation and reproducibility conventions 19

Appendix B: Q_1 Support—High-Dimensional Information and Visual Expression 20
B.1 Information under scaling and removal 20
B.2 Coarse map-expression statistic 21
B.3 Experiment 1: information present but not visually expressed 21
B.4 Experiment 2: strict rank reversal at matched cardinality 22
B.5 Experiment 3: retained versus absent information with nearly equal coarse expression . . 22
B.6 Pinned four-mapper audit of the Q1 contrasts 22
B.7 MNIST ink-mass illustration 23
B.8 Bibtex natural multi-view construction 23

Appendix C: Q_2 Support—Semantic Relation Retention 25
C.1 Pointwise semantic composition and aggregation 25
C.2 Natural-data $\mathrm{D}_{\mathrm{comp}}$ breadth 25
C.3 Replacement decomposition and semantic substitutability 26

Appendix D: Q_3 Support—Geometry After Relation Retention 26

D.1 Geometry diagnostics 27
D.2 Controlled four-method geometry profile 28
D.3 Exact semantic relation, different connectivity and separation 28
D.4 Class-local PH audit 29
D.5 Complementary fixed-native realization control 29
D.6 Same-dimensional UMAP diagnostic 30

Appendix E: Limits and Positioning 31
E.1 Outer boundary: matched CIFAR 31
E.2 Novelty map across neighboring questions 32

A Shared Experimental and Analytical Foundations

This appendix collects notation, constructions, datasets, mapper settings, and evaluation conventions used by more than one of Q_1–Q_3. Question-specific definitions, diagnostics, and evidential contrasts remain in their respective appendices.

A.1 Notation convention

We use bold uppercase symbols for realized data matrices, plain uppercase symbols for the corresponding random variables or random vectors, and bold lowercase symbols for realized observation vectors. The symbols below are reserved throughout the manuscript.

Symbol	Meaning
$\mathbf{X} \in \mathbb{R}^{n \times d_X}$, $X \in \mathbb{R}^{d_X}$, $\mathbf{x}_i$	realized high-dimensional matrix, random observation, row i
$\mathbf{Z} \in \mathbb{R}^{n \times 2}$, $Z \in \mathbb{R}^2$, $\mathbf{z}_i$	realized 2D embedding, random 2D coordinate, row i
$V \in \{Y, Y'\}$, X_V, $\mathbf{X}_V$	semantic target, associated channel random vector, realized channel matrix
n, K, L	observations, semantic classes, and coarse map regions
K_V, d_V	class count and channel dimension for semantic target V
i, j	deterministic observation indices
U, J	random anchor and neighbor indices
c, r	conditioning and destination class indices
k	kNN neighborhood size
τ	experimental condition
$H(\cdot)$, $I(\cdot;\cdot)$	entropy and mutual information
$\mathbf{I}_d$	$d \times d$ identity matrix
$\mathrm{NN}_A^k(i)$	directed kNN neighborhood of observation i in space A
$\boldsymbol{\pi}_i^A$, $\mathbf{T}_A$	pointwise semantic composition and class-conditioned relation
$\mathcal{M}$, g	mapper family and one mapper
$\boldsymbol{\mu}_{V,c}$, $\boldsymbol{\mu}$	one Gaussian class center and the complete realized center configuration
$\bar{\mathbf{a}}_c$, $\mathcal{G}_A(Y)$	geometric class centroid and geometry profile

A.2 Controlled torus construction

The fixed Q_1 torus witness uses $n = 1500$ and seed 42. The separate Q_2/Q_3 relation/geometry audit uses a separately generated seed-42 torus realization with $n = 2000$; the two records are not pooled. In either case, independently sample $\theta_i, \phi_i \sim \mathrm{Unif}(0, 2\pi)$ and $\boldsymbol{\epsilon}_i \sim \mathcal{N}(\mathbf{0}, 0.2^2\mathbf{I}_3)$, then

$$\mathbf{x}_{\mathrm{topo},i} = ((5 + 2\cos\phi_i)\cos\theta_i, (5 + 2\cos\phi_i)\sin\theta_i, 2\sin\phi_i) + \boldsymbol{\epsilon}_i.$$

Semantic labels are independent of torus position: $Y_i \sim \mathrm{Unif}\{1,\dots,4\}$ and $Y_i' \sim \mathrm{Unif}\{1,2\}$. Channel-specific observations are

$$\mathbf{x}_{Y,i} = \boldsymbol{\mu}_{Y,Y_i} + \boldsymbol{\eta}_{Y,i}, \qquad \mathbf{x}_{Y',i} = \boldsymbol{\mu}_{Y',Y_i'} + \boldsymbol{\eta}_{Y',i},$$

with independent noise $\boldsymbol{\eta}_{V,i} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_V})$. For each semantic target $V \in \{Y, Y'\}$, class centers are generated by drawing $\widetilde{\boldsymbol{\mu}}_{V,c} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_V})$ independently for each class c and setting

$$\boldsymbol{\mu}_{V,c} = \delta_V \frac{\widetilde{\boldsymbol{\mu}}_{V,c}}{\|\widetilde{\boldsymbol{\mu}}_{V,c}\|_2}.$$

Sampling random directions at fixed radius avoids building a special symmetric center geometry into the construction and provides independent center configurations for the multi-seed robustness check in Experiment 2. We use $(d_Y, \delta_Y) = (5, 6)$ and $(d_{Y'}, \delta_{Y'}) = (3, 4)$; nuisance is $\mathbf{x}_{\mathrm{noise},i} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{10})$. The resulting Baseline is 21-dimensional. The separate matched-cardinality rank witness uses two five-dimensional four-class channels and is therefore 23-dimensional.

For condition τ, the controlled representation is

$$\mathbf{X}_{\mathrm{syn}}^{(\tau)} = \big[\mathbf{X}_{\mathrm{topo}}, \alpha_Y^{(\tau)}\mathbf{X}_Y, \alpha_{Y'}^{(\tau)}\mathbf{X}_{Y'}, \alpha_{\mathrm{noise}}^{(\tau)}\mathbf{X}_{\mathrm{noise}}\big].$$

Within each controlled audit, the four conditions reuse the same sampled rows and differ only by deterministic scales $(\alpha_Y, \alpha_{Y'}, \alpha_{\mathrm{noise}})$: Baseline $(1,1,1)$, Hide $(.1,1,10)$, Erase $(0,1,10)$, and Redirect $(.1,10,1)$. Neither semantic view is supplied to the dimensionality-reduction method. The separate matched-cardinality rank-reversal witness uses $n = 2000$ with two independent five-dimensional four-class semantic channels and the Redirect scales $\alpha_Y = .1$ and $\alpha_{Y'} = 10$.

A.3 Probability model and mutual-information convention

For each semantic target $V \in \{Y, Y'\}$, let X_V denote its associated Gaussian-channel random vector, with class count K_V and channel dimension d_V. Let

$$\boldsymbol{\mu} := \{\boldsymbol{\mu}_{V,c} : V \in \{Y, Y'\},\ c = 1, \dots, K_V\}$$

denote the complete realized collection of sampled class-center vectors. All controlled mutual-information statements condition on this realized center configuration; we write $I_{\boldsymbol{\mu}}(\cdot;\cdot)$ for mutual information under the resulting conditional population law. We write $\mathbb{E}_{\boldsymbol{\mu}}$ for expectation over the joint law of (X_V, V) conditional on $\boldsymbol{\mu}$. This conditioning is essential: integrating the randomly sampled centers out before drawing a single observation defines a different distribution, and treating the finite realized rows as an empirical distribution defines another one.

Each semantic channel therefore has the form

$$X_V \mid V = c \sim \mathcal{N}(\boldsymbol{\mu}_{V,c}, \mathbf{I}_{d_V}), \qquad \Pr(V = c) = \frac{1}{K_V}.$$

Under fixed centers the posterior is the standard softmax over squared distances,

$$p(V = c \mid X_V = \mathbf{x}_V) = \frac{\exp\big[-\frac{1}{2}\|\mathbf{x}_V - \boldsymbol{\mu}_{V,c}\|^2\big]}{\sum_{r=1}^{K_V} \exp\big[-\frac{1}{2}\|\mathbf{x}_V - \boldsymbol{\mu}_{V,r}\|^2\big]},$$

and hence

$$I_{\boldsymbol{\mu}}(X_V; V) = \log_2 K_V + \mathbb{E}_{\boldsymbol{\mu}}\big[\log_2 p(V \mid X_V)\big]. \tag{4}$$

The held-out balanced linear accuracy $\mathrm{Lin}(\mathbf{X}; V)$ used in Q_1 is an accessibility diagnostic computed on the realized matrix, not an estimator of $I_{\boldsymbol{\mu}}$.

A.4 Nonlinear dimensionality-reduction methods and settings

We evaluate t-SNE, UMAP, PaCMAP, and TriMap as representative nonlinear dimensionality-reduction methods for exploratory 2D visualization. All four construct a low-dimensional realization from geometric relations in the high-dimensional representation, but they differ in which relations their objectives emphasize. This variation provides a cross-method audit of whether the inferential phenomena studied in the paper depend on a particular mapping principle. None of the semantic variables used for post-hoc evaluation is supplied to these objectives.

t-SNE. t-SNE (van der Maaten & Hinton, 2008) converts high-dimensional pairwise similarities into affinities $\mathbf{P_X}$ and compares them with heavy-tailed low-dimensional affinities $\mathbf{Q_Z}$. Its objective is

$$\mathcal{L}_{\mathrm{tSNE}} = \mathrm{KL}(\mathbf{P_X} \,\|\, \mathbf{Q_Z}) \,,$$

which places particular emphasis on representing high-affinity relationships in the low-dimensional map.

UMAP. UMAP (McInnes et al., 2018) constructs a weighted nearest-neighbor relation from $\mathbf{X}$ and optimizes a low-dimensional fuzzy relation to approximate it. Schematically,

$$\mathcal{L}_{\mathrm{UMAP}} = \mathrm{CE}(\mathbf{F_X}, \mathbf{F_Z}) \,,$$

where the cross-entropy compares high- and low-dimensional fuzzy neighborhood relations.

PaCMAP. PaCMAP (Wang et al., 2021) optimizes a weighted objective over different types of point pairs,

$$\mathcal{L}_{\mathrm{PaCMAP}} = \mathcal{L}_{\mathrm{near}} + \mathcal{L}_{\mathrm{mid}} + \mathcal{L}_{\mathrm{far}},$$

with stage-dependent weighting of near, mid-near, and farther pairs. The objective therefore balances relations at several distance scales rather than matching a single neighborhood statistic.

TriMap. TriMap (Amid & Warmuth, 2019) represents high-dimensional relative-distance information through weighted triplets. Its objective can be written schematically as

$$\mathcal{L}_{\mathrm{TriMap}} = \sum_{(i,j,h)} w_{ijh}\, \varphi_{\mathrm{tri}}(\mathrm{dist}_{\mathbf{Z}}(i,j), \mathrm{dist}_{\mathbf{Z}}(i,h)) \,,$$

where w_{ijh} is the triplet weight, $\mathrm{dist}_{\mathbf{Z}}$ is Euclidean distance in the 2D realization, and φ_{tri} is the triplet loss; the loss favors low-dimensional realizations in which the ordering encoded by the high-dimensional triplet (i, j, h) is respected.

These objectives encode different geometric preferences, but they share the feature relevant to the post-hoc workflow:

$\boxed{\mathcal{L}_g = \mathcal{L}_g(\widetilde{\mathbf{Z}}; \mathbf{X}, \xi) \qquad \text{for every } g \in \mathcal{M},}$ with no semantic variable Y among its arguments. They can therefore expose meaningful structure in $\mathbf{X}$ without being defined to preserve, rank, or calibrate the information carried about an arbitrary semantic variable Y introduced only after the map is fixed.

The paper-facing dimensionality-reduction settings are centralized in Table 1. Parameters not listed there use the defaults of the pinned package version for the corresponding reproducibility stage. Reported TriMap fits use the release's deterministic compatibility policy: NumPy, Numba, and the internal Annoy index are seeded from the declared seed, and the TriMap fit is executed with one Numba worker before the previous thread count and NumPy state are restored.

Table 1: **Paper-facing dimensionality-reduction settings.** The table centralizes the mapper settings for the reported figures and numerical audits. Parameters explicitly set by the corresponding analysis are listed; independent illustrative fits run at the pinned package defaults are marked accordingly. "Audit k" refers to the downstream hard-kNN relation audit and is listed separately when it differs from the mapper neighborhood parameter.

Audit / figure	Mapper	Declared settings
Figure 1 MNIST illustration	UMAP	umap-learn 0.5.6 defaults; independent illustrative fit; no target-specific tuning.
Q_1 torus figure	t-SNE	$d_Z = 2$; perplexity 30; `init=pca`; `learning_rate=auto`; 1000 iterations; seed 42.
Q_1 matched rank reversal	t-SNE	$d_Z = 2$; perplexity 30; `init=random`; `learning_rate=auto`; 750 iterations; mapper seeds $\{0, 1, 2\}$.
Q_1 pinned cross-mapper	t-SNE	Redirect/Hide/Erase: perplexity 30, `init=pca`, `learning_rate=auto`, 1000 iterations, seed 42; Rank: `init=random`, 750 iterations, seed 0.
	UMAP	$d_Z = 2$; `n_neighbors=10`; `min_dist=.1`; Euclidean metric; seeds 42 for Redirect/Hide/Erase and 0 for Rank; remaining options use umap-learn 0.5.6 defaults.
	PaCMAP	$d_Z = 2$; `n_neighbors=10`; seeds 42 for Redirect/Hide/Erase and 0 for Rank; remaining options use PaCMAP 0.7.3 defaults.
	TriMap	`n_dims=2`; `n_inliers=10`; PCA initialization; declared seeds 42 for Redirect/Hide/Erase and 0 for Rank under the deterministic single-thread policy; remaining options use TriMap 1.1.4 defaults.
Q_2 natural + Q_2/Q_3 torus	UMAP	$d_Z = 2$; `n_neighbors=10`; `min_dist=.1`; `random_state=42`.
	t-SNE	$d_Z = 2$; perplexity 30; `init=pca`; `learning_rate=auto`; `random_state=42`; remaining options use scikit-learn defaults.
	PaCMAP	$d_Z = 2$; `n_neighbors=10`; `random_state=42`; remaining options use PaCMAP defaults.
	TriMap	`n_dims=2`; `n_inliers=10`; PCA initialization; seed 42 via the deterministic compatibility policy described above; remaining options use TriMap defaults.
Q_2 locality control	UMAP	$d_Z = 2$; `n_neighbors=10`; `min_dist=.1`; Euclidean metric; `n_epochs=200`; seeds $\{42, 43, 44\}$.
Q_3 natural source→2D profile	UMAP	`n_components=2`; `n_neighbors=10`; `min_dist=.1`; input/output metric Euclidean; `init=random`; `random_state=42`; `transform_seed=42`; `n_epochs=500`; `n_jobs=1`; `low_memory=True`.
Q_3 MNIST same-d diagnostic	UMAP	`n_components=` d_X; otherwise the same graph-construction and optimization settings as the Q_3 source→2D profile; fit independently from the same $\mathbf{X}$.
Q_3 fixed-native control	UMAP	$d_Z = 2$; `n_neighbors=15`; audit $k = 10$; `min_dist`$\in \{0, .01, .1, .5, .9\}$; `n_epochs=500`; seed 42. The native fuzzy-graph digest is required to be identical across the five realizations.
Bibtex multi-view	t-SNE	$d_Z = 2$; perplexity 30; `init=pca`; `learning_rate=auto`; 1000 iterations; Euclidean metric; Barnes–Hut method; seed 42; fit to all 7,395 SVD vectors before semantic selection.
Figure 4A raw-CIFAR illustration	UMAP	umap-learn 0.5.6 defaults; independent illustrative fit; no target-specific tuning.
Matched CIFAR	UMAP	$d_Z = 2$; `n_neighbors=10`; `min_dist=.1`; Euclidean metric; 200 epochs; map seeds $\{42, 43, 44\}$, averaged within encoder seed.

A.5 NATURAL-DATA PANEL

The natural audits treat each stored array as the declared high-dimensional representation for that condition; claims are conditional on those fixed representations rather than on the pretraining procedures that produced them.

Table 2: Complete inventory of the 13 natural dataset–representation conditions used in the high-dimensional-to-2D audits.

Dataset	n	d_X	K	Domain	Repr.	Add. preproc.
MNIST	5000	784	10	Vision	pixels	None
FMNIST	5000	784	10	Vision	pixels	None
KMNIST	5000	784	10	Vision	pixels	None
CIFAR-10 raw	5000	3072	10	Vision	pixels	None
CIFAR-10 DINOv2	5000	768	10	Vision	stored DINOv2	None
KMNIST DINOv2	5000	768	10	Vision	stored DINOv2	None
20NG	5000	768	20	Text	e5-base-v2	None
AG News	5000	768	4	Text	e5-base-v2	None
Emotion	2000	768	6	Text	e5-base-v2	None
PBMC3k	2638	50	8	scRNA	stored HVG+PCA	None
Pancreas	1327	50	9	scRNA	stored HVG+PCA	None
Paul15	2730	50	19	scRNA	stored HVG+PCA	None
Dermatology	358	34	6	Tabular	raw	standardized

For the natural-data audits, each stored dataset is capped at $n = 5000$. When a stored array contains more than 5,000 observations, we draw 5,000 observations uniformly without replacement using seed 42; smaller datasets are used in full. Experiment-specific analyses with different cohort construction are stated separately below.

For MNIST, Fashion-MNIST, and KMNIST, the declared source representation is the raw flattened 28×28 pixel vector ($d = 784$). The Q_2 cross-mapper audit and the Q_3 source-geometry audit use separate mapper realizations under their respective protocols, so their UMAP $\mathrm{D}_{\mathrm{comp}}$ values can differ slightly despite sharing the same source representation. Across the thirteen conditions shared by the two audits, the UMAP $\mathrm{D}_{\mathrm{comp}}$ values remain close but need not coincide because the fits use protocol-specific initialization and epoch settings (median absolute difference .0023; maximum .0129). The Q_2 locality control is a third, separate UMAP audit using seeds 42–44 and 200 epochs, so its UMAP $\mathrm{D}_{\mathrm{comp}}$ values should likewise not be identified with either single-seed audit. Figures 1 and 4A are independent default-UMAP illustrations rather than the numerical source for the Q_2/Q_3 audit values; Appendix B.7 separately defines the MNIST ink-mass target and establishes its source-space realizability.

Dataset provenance is as follows: MNIST (LeCun et al., 1998), Fashion-MNIST (Xiao et al., 2017), KMNIST (Clanuwat et al., 2018), CIFAR-10 (Krizhevsky, 2009), 20 Newsgroups (Lang, 1995), AG News (Zhang et al., 2015), Emotion (Saravia et al., 2018), Paul15 (Paul et al., 2015), and Dermatology (Ilter & Güvenir, 1998). The Dermatology audit uses 358 observations after removing the eight rows with missing values from the 366-instance source table. PBMC3k uses the public 10x Genomics PBMC3k release (10x Genomics, 2016). The stored Pancreas condition is consistent with the pancreas collection of Baron et al. (2016), but the released materials do not include an upstream sample identifier that would certify the exact 1,327-cell subset; we therefore treat that stored array as the declared input rather than claim an independently reconstructed release. DINOv2 rows use stored 768-dimensional features associated with Oquab et al. (2024); the released materials do not include the exact DINOv2 checkpoint/extraction script. Text rows use the named `e5-base-v2` representation from the E5 family (Wang et al., 2022); the released materials do not include the original tokenization/pooling script. The single-cell rows are stored 50-dimensional HVG+PCA representations. Accordingly, all natural claims are conditional on the fixed representations used in the audit, not on independent regeneration of the upstream feature-extraction pipelines.

A.6 Evaluation and reproducibility conventions

All relation comparisons use 10-nearest-neighbor sets computed from Euclidean distances, without symmetrization. Table 3 collects the shared analysis settings; experiment-specific exceptions are noted where they arise. When logistic regression does not converge within the iteration budget in any cross-validation fold, the record is marked as *partial*: the balanced-accuracy value is reported but not treated as a converged measurement. For repeated runs of the same condition, we report mean±SD across seeds. When SD is reported across datasets or dataset–method cells, it describes heterogeneity among the aggregated conditions, not a sampling confidence interval.

Table 3: **Shared analysis, readout, and training settings.** These settings cover the quantities reported in the manuscript and distinguish the two PH normalizations used by the paper.

Component	Declared settings
Canonical semantic relation	Directed hard-10NN with Euclidean distance and self-neighbors removed; $\mathrm{D}_{\mathrm{comp}}$ is the macro-average of row-wise total variation; exact-neighbor replacement uses the same directed kNN sets.
Canonical linear accessibility	Balanced accuracy; stratified 5-fold CV with shuffling; seed 42 unless a per-run seed is declared; logistic regression with L-BFGS, $C = 1$, and `max_iter=10000`. Any nonconverged fold makes the component *partial*.
Q_1 coarse map expression	Label-blind K-means with $L = 4$, `n_init=20`, and `random_state=42`.
Q_1 rank-reversal linear readout	Balanced accuracy; stratified 5-fold CV with shuffling and seed 42; L-BFGS logistic regression, $C = 1$, `max_iter=5000`.
Bibtex representation / readout	Row-ℓ_2 normalization; 50-dimensional TruncatedSVD with seed 42; ten supplied train/test partitions for linear access; L-BFGS logistic regression with $C = 1$ and `max_iter=2000`; directed Euclidean hard-10NN for relation diagnostics.
Q_3 class-local PH	One deterministic subset of at most 250 observations per class, reused across source and 2D in the 13-condition audit; the MNIST same-d diagnostic reuses that subset at the additional same-d stage. Each class is divided by its own within-class pairwise-distance q90 separately at each stage; Vietoris–Rips PH0/PH1/PH2 with Ripser 0.6.12 over $\mathbb{F}_2$; connectivity uses within-class hard-10NN.
Full-vector PH (fixed-native control)	At most 250 observations per class; one global pairwise-distance q90 scale per geometric space, estimated from at most 2,000 points with seed 42; source and mapped spaces are normalized independently; PH computed through dimension 2.
Matched-CIFAR representation learning	Stratified $n = 5000$ sample, data seed 42; PCA to 50 dimensions (seed 42) followed by standardization; MLP $50 \to 256 \to 128 \to 64$ with ℓ_2-normalized output; supervised-contrastive temperature 0.1; Adam learning rate 10^{-3}; batch size 256 with `drop_last=True`; 200 epochs; deterministic CPU training; encoder seeds 42–46. The reported linear readout uses stratified 3-fold CV and L-BFGS with `max_iter=2000`.

The reproducibility environment pins Python 3.10, NumPy 1.24.4, SciPy 1.11.4, scikit-learn 1.3.2, umap-learn 0.5.6, PaCMAP 0.7.3, TriMap 1.1.4, Numba 0.59.1, and Ripser 0.6.12. For stages whose original library build was not version-locked, reproducibility is tied to the released numerical record and script together with the declared settings above (a compatibility shim translates only the `max_iter`/`n_iter` keyword name across scikit-learn versions; no hyperparameter is changed). The Q1 cross-mapper breadth audit is executed directly in the pinned environment listed above.

Evidence coverage. The three questions use different experimental scopes by design. Q_1 uses fixed t-SNE Redirect/Hide/Erase contrasts, a 5×3 matched-cardinality t-SNE Rank stress test, and a pinned single-run breadth audit of Redirect, Rank, Hide, and Erase under t-SNE, UMAP, PaCMAP, and TriMap using predeclared settings and no hyperparameter search; the fixed MNIST UMAP illustration and Bibtex t-SNE multi-view case provide complementary natural examples. Q_2 audits the controlled torus and the natural relation profiles across UMAP, t-SNE, PaCMAP, and TriMap; Table 7 reports the complete $13 \times 4 = 52$-cell natural cross-mapper record. Q_3 uses the four-method torus separation profile, a 13-condition source-to-UMAP class-local PH audit, the exact-relation geometric witness, a UMAP-specific fixed-native control, and a complementary MNIST same-dimensional diagnostic. The four-mapper Q1 audit uses one map per method to assess mapper breadth, whereas the 5×3 t-SNE Rank stress test assesses seed robustness; the broader Q_3 natural PH audit remains UMAP-specific.

Controlled Q_1 settings. For the primary fixed-center information calculations in Q_1, Eq. 4 is evaluated with 10^6 fresh Monte-Carlo draws for each fixed center configuration. The primary rank-reversal stress test uses $n = 2000$, data seeds $\{0, 1000, 2000, 3000, 4000\}$, and t-SNE mapper seeds $\{0, 1, 2\}$. The pinned cross-mapper breadth audit uses the fixed $n = 1500$, seed-42 Redirect/Hide/Erase source arrays and one matched-cardinality Rank dataset with $n = 2000$, data seed 0; each is mapped once by every audited mapper with the predeclared settings in Table 1. Its independent MI check uses 10^6 Monte-Carlo draws for the seed-42 Redirect/Hide channels and 5×10^5 draws for each seed-0 Rank channel. No mapper or readout hyperparameter is selected using the semantic outcomes.

Full natural cross-mapper audit. The complete audit contains $13 \times 4 = 52$ dataset–mapper cells and is reported directly in Table 7; the broader Q_3 geometry analysis is reported separately below.

B Q_1 Support: High-Dimensional Information and Visual Expression

This appendix supports the quantities and contrasts reported in Q_1. The shared controlled construction and fixed-center probability model are defined in Appendix A; the material below contains the Q_1-specific information properties, declared coarse expression readout, experiment-level calculations, the pinned four-mapper audit, and the MNIST/Bibtex natural-data illustrations.

B.1 Information under scaling and removal

The high-dimensional calculations in Q_1 use the fixed-center Gaussian-channel convention of Appendix A.3. We evaluate Eq. 4 by Monte-Carlo from the known generator using 10^6 fresh draws for each fixed center configuration. Because the torus, nuisance, and competing semantic blocks are jointly independent of the queried pair (X_V, V) by construction, adding them does not change the mutual information about that semantic target in the high-dimensional representation.

Two properties of this setup are used repeatedly:

- **Scaling preserves information.** A deterministic nonzero rescaling is bijective, so for any $\alpha \neq 0$,
$$I_{\boldsymbol{\mu}}(\alpha X_V; V) = I_{\boldsymbol{\mu}}(X_V; V)$$
(Cover & Thomas, 2006, Sec. 2.8). Rescaling a semantic block therefore changes its contribution to the metric structure seen by the dimensionality-reduction method without changing the information it carries.
- **Zeroing removes information.** Setting the only Y-dependent block to $\alpha = 0$ removes it from the representation entirely. Because the remaining blocks are jointly independent of Y by construction, the resulting representation is independent of Y and therefore
$$I_{\boldsymbol{\mu}}(X_{\text{syn}}^{(\text{Erase})}; Y) = 0.$$

The held-out linear-readout accuracies Lin reported alongside each experiment are separate accessibility diagnostics computed on the realized rows and are not estimators of mutual information.

B.2 COARSE MAP-EXPRESSION STATISTIC

The post-hoc workflow ultimately asks a visual question: how strongly do the semantic colors align with coarse spatial regions that a viewer can perceive in the finished map? We operationalize that judgment with a deliberately simple label-blind spatial partition rather than treating Mix as a perceptual model. For the controlled Q_1 analyses, the uncolored realized 2D coordinates $\mathbf{Z} \in \mathbb{R}^{n\times 2}$ are partitioned by

$$C_{\mathbf{Z}} = \mathrm{KMeans}_4(\mathbf{Z}),$$

using scikit-learn K-means with `n_clusters`=4, `n_init`=20, and `random_state`=42. We use $L = 4$ as the primary resolution because it is fine enough to resolve the four-class target while remaining coarse enough to evaluate the two-class target without forcing an excessively fragmented partition. The rule is fixed across the controlled Q_1 conditions and does not receive Y or Y'. A separate partition is fit for each realized map, but when multiple semantic variables are evaluated on the same coordinates, the fitted region assignments are reused unchanged. We use $\widehat{I}$ for the empirical plug-in mutual information computed from the same region-by-class counts.

Let $n_{\ell c}$ be the number of observations in region ℓ with semantic class c, and $n_\ell = \sum_c n_{\ell c}$. The implementation uses the point-weighted plug-in conditional entropy

$$\widehat{H}(Y \mid C_{\mathbf{Z}}) = -\sum_{\ell: n_\ell > 0} \frac{n_\ell}{n} \sum_{c: n_{\ell c} > 0} \frac{n_{\ell c}}{n_\ell} \log\left(\frac{n_{\ell c}}{n_\ell}\right), \tag{5}$$

and reports

$$\mathrm{Mix}(\mathbf{Z}; Y) = \frac{\widehat{H}(Y \mid C_{\mathbf{Z}})}{\log K}.$$

The implementation uses natural logarithms; because the numerator and denominator use the same base, the normalized value is base-invariant. For an empirically balanced K-class variable, $\widehat{H}(Y) = \log K$, and therefore

$$1 - \mathrm{Mix}(\mathbf{Z}; Y) = \frac{\widehat{I}(C_{\mathbf{Z}}; Y)}{\log K}.$$

Thus $1 - \mathrm{Mix}$ is the fraction of the empirical label entropy captured by the declared coarse spatial partition, while Mix is the normalized residual label uncertainty after the spatial region is known. More generally, under empirical independence of Y and $C_{\mathbf{Z}}$, $\mathrm{Mix} = \widehat{H}(Y)/\log K$, which equals 1 only for an exactly balanced K-class target. Thus the controlled values near one are interpreted as near-maximal coarse mixing for the realized class frequencies; Mix is a coarse partition readout rather than a perceptual-equivalence measure.

Under idealized conditions—the mapping and partition are fitted once and then held fixed, and a new observation is mapped without access to its label—the chain $Y \to X \to Z \to C_Z$ is Markov, so data processing gives $I(Y; C_Z) \le I(Y; Z) \le I(Y; X)$ (Cover & Thomas, 2006, Secs. 2.4, 2.8). This upper bound holds at the population level under a fixed fitted state; the empirical Mix remains an empirical readout rather than a population lower confidence bound on source information.

B.3 EXPERIMENT 1: INFORMATION PRESENT BUT NOT VISUALLY EXPRESSED

In the Redirect condition, both semantic channels are retained ($\alpha_Y = 0.1$, $\alpha_{Y'} = 10$): neither variable is removed, so by scaling invariance both channels preserve their original information. Using Eq. 4 on the fixed seed-42 center configuration, the Monte-Carlo estimates are

$$I_{\boldsymbol{\mu}}(X_{\mathrm{syn}}^{(\mathrm{Redirect})}; Y) \approx 1.966 \text{ bits}, \qquad I_{\boldsymbol{\mu}}(X_{\mathrm{syn}}^{(\mathrm{Redirect})}; Y') \approx 0.995 \text{ bit}.$$

Thus the representation carries nearly twice as many raw mutual-information bits about Y as about Y', and both variables are highly linearly accessible (Lin = .991 and .997 respectively). For Y alone, the key finding is already complete: nearly two bits of source information and near-perfect linear accessibility coexist with near-maximal coarse mixing ($\mathrm{Mix} = 0.997$). The same frozen map then supplies a second observation: Y' aligns strongly with the visible organization ($\mathrm{Mix} = 0.029$). Thus Experiment 1 establishes the absolute failure mode: substantial source information can remain weakly expressed even when another informative variable is clearly expressed on the same map. Experiment 2 complements this by matching cardinality and testing the stronger relative question of whether the source-information ordering itself can reverse.

B.4 Experiment 2: strict rank reversal at matched cardinality

The matched-cardinality variant gives both Y and Y' four classes and five signal dimensions, so both variables have the same maximum entropy of 2 bits. The Y channel uses a larger Gaussian center-radius parameter ($\delta_Y = 6$ versus $\delta_{Y'} = 3$), while the input scales are $\alpha_Y = 0.1$ and $\alpha_{Y'} = 10$.

Applying Eq. 4 independently to each of the five fixed center configurations generated with data seeds $0, 1000, 2000, 3000, 4000$ yields the following mean±SD of the five configuration-specific Monte-Carlo estimates:

$$1.985 \pm .021 \text{ bits} \quad \text{for } I_{\boldsymbol{\mu}}(X_{\text{syn}}^{(\text{Rank})}; Y), \qquad 1.605 \pm .227 \text{ bits} \quad \text{for } I_{\boldsymbol{\mu}}(X_{\text{syn}}^{(\text{Rank})}; Y'),$$

where the reported mean and standard deviation are taken across center configurations; the Monte-Carlo standard error within each fixed configuration is below .002 bit. The corresponding held-out balanced linear accuracies are $.995 \pm .008$ and $.887 \pm .062$, confirming the same ordering.

On the frozen 2D maps, this ordering reverses in every one of the fifteen 5×3 data/map realizations:

$$\text{Mix}(\mathbf{Z}; Y) = .9978 \pm .0006 > \text{Mix}(\mathbf{Z}; Y') = .325 \pm .119$$

(lower is clearer). The variable about which the representation carries more information consistently looks less expressed.

B.5 Experiment 3: retained versus absent information with nearly equal coarse expression

Hide and Erase differ only in whether the Y-dependent channel is retained at a nonzero scale or removed:

- **Hide** ($\alpha_Y = 0.1$): the Y-channel is retained under an invertible rescaling, so by scaling invariance the information is unchanged:

$$I_{\boldsymbol{\mu}}(X_{\text{syn}}^{(\text{Hide})}; Y) = I_{\boldsymbol{\mu}}(X_{\text{syn}}; Y) = 1.966 \text{ bits}.$$

- **Erase** ($\alpha_Y = 0$): the only Y-dependent block is removed, so the remaining representation is independent of Y and therefore

$$I_{\boldsymbol{\mu}}(X_{\text{syn}}^{(\text{Erase})}; Y) = 0.$$

Despite this sharp difference in the high-dimensional information state, the declared coarse Mix readout on the frozen 2D maps is nearly identical (0.9980 vs. 0.9976). Thus, under the declared coarse Mix readout, nearly equal expression can accompany opposite high-dimensional information states. This result concerns the declared coarse statistic; the complete embeddings may differ.

B.6 Pinned four-mapper audit of the controlled Q1 contrasts

To test whether the controlled Q_1 failures depend on a particular mapper, we evaluate each contrast once under all four audited nonlinear dimensionality-reduction methods using the pinned environment and the predeclared settings in Table 1. No semantic-target-specific hyperparameter search is performed. Redirect, Hide, and Erase use the fixed $n = 1500$, seed-42 source construction; the matched-cardinality Rank breadth run uses $n = 2000$ and data seed 0. This breadth audit uses one map per method to assess mapper breadth; the separate 5×3 t-SNE Rank stress test assesses seed robustness.

For the fixed Redirect/Hide source configuration, the Monte-Carlo estimates are $I_{\boldsymbol{\mu}}(X; Y) = 1.9655$ bits and $I_{\boldsymbol{\mu}}(X; Y') = 0.9945$ bits; Erase has $I_{\boldsymbol{\mu}}(X; Y) = 0$ by construction. For the matched Rank seed-0 configuration, $I_{\boldsymbol{\mu}}(X; Y) = 1.9878$ bits and $I_{\boldsymbol{\mu}}(X; Y') = 1.5585$ bits. Table 4 reports the corresponding coarse expression values.

Table 4: **Pinned single-run cross-mapper audit of all three controlled Q_1 contrasts.** Lower Mix means clearer alignment with the declared label-blind four-region partition. Redirect tests Experiment 1; Rank tests the matched-cardinality ordering reversal; Hide/Erase tests nearly equal coarse expression under opposite source-information states.

Mapper	Redir. Y	Redir. Y'	Rank Y	Rank Y'	Hide Y	Erase Y
t-SNE	.9973	.0288	.9982	.3531	.9980	.9976
UMAP	.9976	.0380	.9986	.3431	.9979	.9984
PaCMAP	.9986	.0218	.9973	.3566	.9984	.9986
TriMap	.9973	.0186	.9982	.3387	.9982	.9982

The qualitative conclusion is identical for every audited mapper. In Redirect, Y remains near-maximally mixed while the also informative Y' is clearly expressed, reproducing the absolute contrast under every audited mapper. In the matched Rank construction, $\mathrm{Mix}(\mathbf{Z}; Y) > \mathrm{Mix}(\mathbf{Z}; Y')$ for every mapper despite the opposite source-information ordering. In Hide versus Erase, both maps remain near-maximally mixed although the source information changes from 1.9655 bits to zero. These results establish mapper breadth across the four audited methods.

B.7 MNIST INK-MASS ILLUSTRATION

Figure 1 provides the descriptive visual comparison; this subsection defines the ink-mass target and establishes its source-space linear realizability. For raw MNIST pixels (LeCun et al., 1998), ink mass is the linear statistic $m_i = \mathbf{1}_{784}^\top \mathbf{x}_i$. We define ten ordered ink-mass classes using the 10th, 20th, …, 90th percentiles of ink mass over the full 60,000-image MNIST training cohort; values exactly on a threshold are assigned to the upper bin. Because the ink-mass classes are intervals of the linear statistic $m(\mathbf{x}) = \mathbf{1}^\top \mathbf{x}$, affine multiclass scores can be chosen whose decision boundaries coincide with the declared percentile thresholds. The ink-mass deciles are therefore exactly linearly realizable from the original pixel representation. Ink mass is not merely a recoding of digit identity: over the full 60,000-image cohort, $\widehat{I}(Y_{\mathrm{digit}}; Y_{\mathrm{ink}})/\widehat{H}(Y_{\mathrm{ink}}) = 0.101$ and $\widehat{H}(Y_{\mathrm{ink}} \mid Y_{\mathrm{digit}})/\widehat{H}(Y_{\mathrm{ink}}) = 0.899$.

B.8 BIBTEX NATURAL MULTI-VIEW CONSTRUCTION

Bibtex provides a descriptive natural multi-view example of the same post-hoc setting: two semantic readings of one fixed map differ substantially in visual organization while both remain nontrivially accessible in the source representation. Its role is to illustrate how one fixed map can support different semantic readings, rather than to supply a cross-target rank-reversal test.

We use the Mulan distribution of the Bibtex benchmark (Katakis et al., 2008), containing 7,395 papers with 1,836 input features and 159 tags. From these tags, we construct two semantic views—*broad topic* (3 classes) and *research mode* (5 classes)—by grouping related tags as shown in Table 5. A paper receives a class in a view only when its tag set intersects exactly one group; the common unambiguous subset contains 467 papers.

The representation and map are constructed without either semantic view:

1. Row-ℓ_2 normalize all 7,395 sparse feature vectors.
2. Fit a 50-dimensional truncated SVD with seed 42.
3. Fit t-SNE to all 7,395 SVD vectors using perplexity 30, PCA initialization, automatic learning rate, 1,000 iterations, Euclidean distance, and seed 42.
4. Only after the coordinates are fixed, retain the 467 common papers and attach the two semantic labelings post-hoc.

Table 5: **Bibtex semantic views: tag-group definitions.** Tag names follow the Mulan label order with the common `TAG_` prefix removed. A paper is assigned to a view only when its tags intersect exactly one listed group.

Broad topic (3 classes, 467 papers)		
Physics / complex sys.	physics, quantum, spin, phase, nonequilibrium, dynamics, montecarlo, granular	138
Bio / biosensing	agdetection, amperometry, antibody, apob, bettasplendens, compounds, electrochemistry, elisa, epitope, immunoassay, immunoelectrode, immunosensor, ldl, liposome, molecular, sequence	55
Software / systems	software, programming, architecture, requirements, system, systems, objectoriented, maintenance	274
Research mode (5 classes, 467 papers)		
Theory / formal	theory, formal, logic, mathematics	79
Modeling / simulation	model, modeling, models, simulation, simulations	177
Empirical / evaluation	empirical, evaluation, metrics	76
Design / engineering	design, engineering	97
Review / survey	review, survey, litreview	38

Result. Figure 5 and Table 6 show that broad topic aligns clearly with visible 2D regions, whereas research mode appears substantially more diffuse. Broad topic has linear access $.642 \pm .053$ and 2D silhouette .326, whereas research mode has linear access $.428 \pm .047$ and silhouette $-.011$. The neighbor replacement rate is identical ($R_{\text{nbr}} = .679$) because the geometry is fixed and only the semantic coloring changes; D_{comp} differs because it is label-dependent (.022 versus .052). Because the semantic views differ in cardinality and are constructed from tag groups rather than original Bibtex targets, their diagnostics should be read as two post-hoc semantic views of the same fixed representation and map.

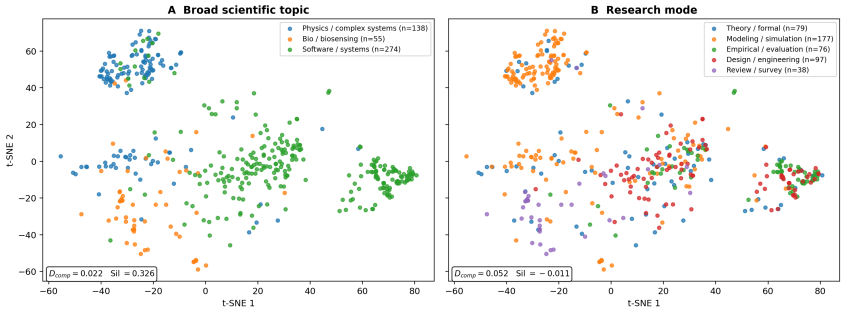


Figure 5: **One fixed Bibtex 2D map, two semantic readings.** A single label-free t-SNE map is fit to all 7,395 papers before either semantic view is attached. The panels show the same 467 papers on identical coordinates, colored by **A.** broad scientific topic and **B.** research mode. No point moves between panels.

Table 6: **Bibtex multiclass views.** Both columns use the same fixed 50-dimensional representation and the same label-free t-SNE coordinates. Linear access is mean±SD over the ten supplied train/test partitions; all remaining quantities refer to the fixed seed-42 map.

Metric	Broad topic	Research mode
Classes / papers	$3/467$	$5/467$
Linear access $\mathrm{Lin}(\mathbf{X}; Y)$	$.642 \pm .053$	$.428 \pm .047$
$\mathrm{D}_{\mathrm{comp}}$	.022	.052
R_{nbr}	.679	.679
2D silhouette	.326	$-.011$

C Q_2 SUPPORT: SEMANTIC RELATION RETENTION

This appendix supports Result 2. Shared notation, mapper settings, the natural-data inventory, and evaluation conventions are collected in Appendix A; the sections below provide the Q_2-specific relation definition, breadth results, replacement mechanism, and locality control.

C.1 POINTWISE SEMANTIC COMPOSITION AND AGGREGATION

For point i in space A, the semantic composition of its directed kNN neighborhood is the probability vector

$$\boldsymbol{\pi}_i^A = \frac{1}{k} \sum_{j \in \mathrm{NN}_A^k(i)} \mathbf{e}_{Y_j}, \qquad \mathbf{e}_r \in \mathbb{R}^K \text{ is the } r\text{th standard basis vector.}$$

This replaces exact neighbor identities by their semantic classes. With $U \sim \mathrm{Unif}([n])$ and $J \mid U \sim \mathrm{Unif}(\mathrm{NN}_A^k(U))$, averaging these pointwise compositions within conditioning class c gives the main-text relation,

$$\mathbb{E}[\boldsymbol{\pi}_U^A \mid Y_U = c] = \Pr(Y_J = \cdot \mid Y_U = c) = \mathbf{T}_A(c, \cdot).$$

For $A = \mathbf{X}$ and $A = \mathbf{Z}$, these are precisely $\mathbf{T}_{\mathbf{X}}$ and $\mathbf{T}_{\mathbf{Z}}$. The main claim concerns this class-conditioned relation, not preservation of every pointwise composition.

C.2 NATURAL-DATA $\mathrm{D}_{\mathrm{comp}}$ BREADTH

The complete 13-condition dataset–representation inventory and provenance notes are given in Appendix A.5. Table 7 reports the full $13 \times 4 = 52$-cell cross-mapper relation audit.

Table 7: **Complete natural-data Q_2 cross-mapper audit at $k = 10$.** Each mapper cell reports $\mathrm{D}_{\mathrm{comp}}$ / exact-neighbor overlap for the same dataset–method realization. Lower $\mathrm{D}_{\mathrm{comp}}$ indicates greater retention of the class-conditioned semantic relation; higher overlap indicates greater retention of exact neighbor identities. The pooled row summarizes all $13 \times 4 = 52$ displayed cells.

Dataset	UMAP	t-SNE	PaCMAP	TriMap
MNIST	.052/.324	.048/.446	.082/.239	.068/.257
FMNIST	.081/.337	.061/.444	.103/.264	.102/.275
KMNIST	.055/.305	.066/.414	.059/.236	.057/.251
CIFAR-10 raw	.265/.096	.254/.192	.277/.056	.273/.062
CIFAR-10 DINOv2	.029/.250	.030/.373	.032/.186	.033/.202
KMNIST DINOv2	.095/.242	.055/.371	.245/.146	.259/.150
20 Newsgroups	.072/.267	.067/.367	.093/.171	.098/.175
AG News	.015/.335	.013/.435	.012/.215	.012/.237
Emotion	.092/.178	.079/.282	.131/.097	.117/.137
PBMC3k	.088/.110	.083/.202	.062/.074	.065/.096
Pancreas	.025/.533	.019/.621	.025/.482	.030/.500
Paul15	.149/.208	.158/.304	.185/.151	.186/.167
Dermatology	.044/.516	.042/.551	.046/.445	.040/.496
Pooled mean±SD	$\mathrm{D}_{\mathrm{comp}} = .091 \pm .075$	/	overlap $= .278 \pm .143$	

C.3 Replacement decomposition and semantic substitutability

For observation i, define its pointwise exact-neighbor replacement fraction and the size of conditioning class c by

$$r_i := 1 - \frac{|\,\mathrm{NN}^k_{\mathbf{X}}(i) \cap \mathrm{NN}^k_{\mathbf{Z}}(i)|}{k}, \qquad n_c := |\{i : Y_i = c\}|.$$

For conditioning class c, let $\rho_c := n_c^{-1} \sum_{i:Y_i=c} r_i$ be the fraction of outgoing edges replaced. Then $R_{\mathrm{nbr}} = \sum_c (n_c/n)\rho_c$. When $\rho_c > 0$, let $\boldsymbol{\nu}_c^-$ and $\boldsymbol{\nu}_c^+$ be the destination-class distributions of removed and added edges. For $\rho_c = 0$, choose the same arbitrary probability vector for both so that the expressions below remain well defined. Equal out-degree gives the exact identity

$$\begin{aligned} \mathbf{T}_{\mathbf{Z}}(c,\cdot) - \mathbf{T}_{\mathbf{X}}(c,\cdot) &= \rho_c(\boldsymbol{\nu}_c^+ - \boldsymbol{\nu}_c^-), \\ \mathrm{TV}(\mathbf{T}_{\mathbf{Z}}(c,\cdot), \mathbf{T}_{\mathbf{X}}(c,\cdot)) &= \rho_c\, \mathrm{TV}(\boldsymbol{\nu}_c^+, \boldsymbol{\nu}_c^-). \end{aligned} \tag{6}$$

Thus large identity replacement can coexist with small semantic drift when added and removed neighbors are semantically substitutable at the measured resolution.

Corollary 1 (Semantic substitutability bound)**.** *If* $\mathrm{TV}(\boldsymbol{\nu}_c^+, \boldsymbol{\nu}_c^-) \le \varepsilon_c$ *for every conditioning class* c*, then*

$$\mathrm{D}_{\mathrm{comp}}(\mathbf{T}_{\mathbf{X}}, \mathbf{T}_{\mathbf{Z}}) \le \frac{1}{K} \sum_c \rho_c \varepsilon_c.$$

In particular, $\varepsilon_c \le \varepsilon$ *for all* c *implies* $\mathrm{D}_{\mathrm{comp}}(\mathbf{T}_{\mathbf{X}}, \mathbf{T}_{\mathbf{Z}}) \le \varepsilon$.

Locality control. To probe why $\mathbf{T}$ can remain stable beyond the replacement identity, we define *RankBand*: each source top-k neighborhood is replaced by the next k nearest observations, at source-distance ranks $k{+}1{:}2k$. RankBand therefore has exactly zero top-k identity overlap by construction, just like the global null, but keeps the replacements local in the source geometry. If locality rather than observation identity stabilizes the semantic relation, RankBand should produce much smaller $\mathrm{D}_{\mathrm{comp}}$ than a global zero-overlap replacement.

Table 8: Q_2 **locality control.** Zero-overlap controls at $k = 10$ for 20 Newsgroups and AG News; all entries are $\mathrm{D}_{\mathrm{comp}}$ (lower is closer). The global null draws replacements outside the original top-k. This separate locality audit uses UMAP seeds 42–44, with 20 randomized null draws per setting.

Dataset	UMAP $\mathrm{D}_{\mathrm{comp}}$	RankBand $\mathrm{D}_{\mathrm{comp}}$	Global-null $\mathrm{D}_{\mathrm{comp}}$
20 Newsgroups	.0767	.1079	.6498
AG News	.0124	.0280	.5814
Mean	.045	.068	.616

Table 8 shows that source-local replacement provides explanatory structure beyond exact identity overlap: local rank-band replacement preserves far more class composition than global replacement despite identical zero top-k overlap. Its UMAP column comes from the separate three-seed locality protocol in Table 1, not from the single-seed cross-mapper or Q_3 geometry realizations.

D Q_3 Support: Geometry After Relation Retention

This appendix supports Result 3. Its logic is conditional: after verifying whether the class-conditioned semantic relation remains close as in Q_2, we ask whether the *geometry* through which that relation is realized is also preserved—or whether the mapping can reorganize geometry while keeping the relation close. We proceed in six steps. Appendix D.1 defines the diagnostics and their interpretation; Appendix D.2 applies them to the controlled torus across all four mappers; Appendix D.3 gives an exact-$\mathrm{D}_{\mathrm{comp}} = 0$ witness; Appendix D.4 audits all thirteen natural conditions; Appendix D.5 holds UMAP's native fuzzy graph fixed while varying the realization; and Appendix D.6 provides a complementary same-dimensional diagnostic.

D.1 GEOMETRY DIAGNOSTICS

The diagnostics below complement $\mathbf{T}$ by quantifying geometric properties *conditional on* the semantic-relation analysis in Q_2. We measure geometry at two levels: *between classes* (how strongly separated are the semantic groups?) and *within classes* (what internal form does each group have?).

Notation. For a geometric space A, let $\mathbf{a}_i$ denote the coordinate vector of observation i. For each nonempty class c, define

$$n_c := |\{i : Y_i = c\}|, \qquad \bar{\mathbf{a}}_c := \frac{1}{n_c} \sum_{i:Y_i=c} \mathbf{a}_i, \qquad \bar{\mathbf{a}} := \frac{1}{n} \sum_{i=1}^{n} \mathbf{a}_i.$$

Between/within spread ratio (B/W). We define the macro between- and within-class spreads as

$$B(A;Y) = \frac{1}{K} \sum_{c=1}^{K} \|\bar{\mathbf{a}}_c - \bar{\mathbf{a}}\|_2^2, \qquad W(A;Y) = \frac{1}{K} \sum_{c=1}^{K} \frac{1}{n_c} \sum_{i:Y_i=c} \|\mathbf{a}_i - \bar{\mathbf{a}}_c\|_2^2.$$

B measures the average squared displacement of class centroids from the overall sample centroid, whereas W measures average within-class scatter, with classes weighted equally in both summaries. All reported conditions have $W(A;Y) > 0$, and we define

$$\mathrm{B/W}(A;Y) = \frac{B(A;Y)}{W(A;Y)}.$$

Large $\mathrm{B/W}$ indicates strong class-centroid separation relative to within-class scatter. Small $\mathrm{B/W}$ indicates weak centroid separation relative to within-class scatter and is compatible with disjoint multimodal or nonconvex class geometry. Under a uniform rescaling $A \mapsto \lambda A$ with $\lambda \neq 0$, both B and W scale by λ^2, so the ratio is unchanged.

Silhouette score. As a complementary separation measure, we report the standard silhouette score (Rousseeuw, 1987). For each observation, silhouette compares its average distance to same-class observations with its average distance to the nearest other class; values near $+1$ indicate strong pointwise separation, values near zero indicate overlap, and negative values indicate that another class is geometrically closer. Silhouette captures pointwise class separation, whereas $\mathrm{B/W}$ captures class-centroid separation relative to within-class scatter; together they provide complementary views of between-class geometry.

Within-class geometry: persistent homology (PH). Separation metrics do not describe the internal geometric organization of a class. We therefore summarize within-class structure with Vietoris–Rips persistent homology (Edelsbrunner et al., 2002). PH0 describes how connected components merge across the filtration; its total persistence aggregates the finite merger lifetimes rather than counting sub-clusters or connected components at any fixed scale. PH1 and PH2 summarize one- and two-dimensional homological features of the Vietoris–Rips filtration, respectively. These quantities describe homological features of the Vietoris–Rips filtration rather than literal geometric structures in the underlying data-generating support. Changes in total persistence therefore quantify reorganization of the persistence profile across filtration scales without identifying which specific topological features changed.

For class c and homological degree q, let $\mathcal{D}_q(A, c)$ denote the persistence intervals. We summarize the finite lifetimes by total persistence,

$$\mathrm{TP}_q(A, c) = \sum_{(b,d)\in\mathcal{D}_q(A,c):\, d<\infty} (d - b), \qquad \overline{\mathrm{TP}}_q(A) = \frac{1}{K} \sum_{c=1}^{K} \mathrm{TP}_q(A, c).$$

Larger values indicate more total finite persistence at that degree; these scalar sums do not by themselves identify which individual topological features changed.

Class-local normalization. Direct persistence values are sensitive to scale. For the primary source-to-2D audit, each class is normalized separately at each stage: within-class pairwise distances are divided by the 90th percentile of that class's own pairwise distances before computing PH. This removes class translation and uniform class-wise rescaling as explanations for changes in the persistence summaries.

Implementation. We compute PH with Ripser 0.6.12 over $\mathbb{F}_2$ (`coeff=2`). For every class in the 13-condition audit, one deterministic subset of at most 250 observations is reused across source and 2D stages. For the MNIST same-dimensional diagnostic, the same subset is additionally reused at the same-d stage. Infinite PH0 bars are retained as counts but excluded from finite-lifetime sums.

D.2 Controlled four-method geometry profile

Purpose. To assess mapper breadth for the geometry change documented in Q_3, we evaluate the same controlled Baseline under all four audited nonlinear dimensionality-reduction methods.

Result. On torus Baseline, the class-conditioned semantic relation remains close across all four methods ($\mathrm{D}_{\mathrm{comp}} = 0.0046$ for each mapper), while between-class separation changes substantially. In this controlled Baseline, all four 2D maps have class-pure hard-10NN neighborhoods, so $\mathbf{T}_Z = \mathbf{I}_4$ for every mapper. The identical $\mathrm{D}_{\mathrm{comp}} = 0.0046$ values therefore arise from comparing the same slightly impure source relation $\mathbf{T}_X$ with the same class-pure destination relation, despite substantial differences in exact neighbor identity. Table 9 reports the complete profile. In every method, the 2D map increases both silhouette and B/W relative to the source despite replacing 52–68% of exact outgoing neighbors.

Table 9: **A retained class-conditioned relation can appear through different geometries.** On torus Baseline, $\mathrm{D}_{\mathrm{comp}} = 0.0046$ for each audited mapper while separation changes substantially. R_{nbr} is the fraction of outgoing kNN edges replaced by the mapping.

Space	R_{nbr}	$\mathrm{D}_{\mathrm{comp}}$	Sil.	B/W
$\mathbf{X}_{\mathrm{syn}}$ (source)	–	–	.182	.513
UMAP	.602	.0046	.523	3.45
t-SNE	.522	.0046	.376	1.47
PaCMAP	.676	.0046	.318	1.13
TriMap	.661	.0046	.281	1.12

D.3 Exact semantic relation, different connectivity and separation

Purpose. The controlled profiles show large geometric changes when the semantic relation is close. We sharpen the premise with a witness consisting of two geometric realizations with *exactly* the same class-conditioned semantic relation and distinct geometry.

Construction. We build two point sets with three semantic classes and identical directed hard-10NN class-conditioned relations,

$$\mathbf{T}^{(1)} = \mathbf{T}^{(2)} = \mathbf{I}_3, \qquad \mathrm{D}_{\mathrm{comp}}\Big(\mathbf{T}^{(1)}, \mathbf{T}^{(2)}\Big) = 0.$$

In both realizations, every point's 10 nearest neighbors belong to the same semantic class. Realization 1 has one connected component per class, whereas Realization 2 splits every class into two equal disconnected components. Thus the semantic relation is identical while connectivity and separation are not. Table 10 reports the resulting contrast.

Table 10: **Exact relation, non-unique geometry.** Both realizations have $\mathbf{T}^{(1)} = \mathbf{T}^{(2)} = \mathbf{I}_3$ and hence cross-realization $\mathrm{D}_{\mathrm{comp}} = 0$. Components/class and Largest report hard-10NN connectivity within each semantic class; B/W is the between/within spread ratio.

Realization	Components/class	Largest	$\mathrm{D}_{\mathrm{comp}}$	Sil.	B/W
1	1	1.00	.000	.974	1612.90
2	2	.50	.000	.244	6.46×10^{-6}

Implication. Even at $\mathrm{D}_{\mathrm{comp}} = 0$, the measured semantic relation does not determine the geometry. The same relation can be realized through one connected hard-10NN component per class or

through two disconnected hard-10NN components per class with a distinct separation profile. This exact witness is the discrete counterpart of the main-text MNIST lesson: semantically supported clear islands in 2D need not imply corresponding island geometry in the source space.

D.4 CLASS-LOCAL PH AUDIT

Purpose. The controlled torus and exact witness establish that geometry can change despite relation retention. We next ask whether the same conditional pattern appears across the thirteen natural dataset–representation conditions in Table 2.

Setup. Under the separate Q_3 source-to-2D geometry protocol, we compare the same observations in the declared high-dimensional representation and a UMAP 2D realization. B/W summarizes class-centroid separation relative to within-class scatter; PH0/PH1/PH2 use the class-local normalization defined above to summarize internal persistence profiles independently of uniform class-wise scale changes.

Result. Table 11 reports the full source-to-2D profile.

Table 11: **Source-to-UMAP-2D geometry profile with class-local PH normalization.** Each row uses the canonical semantic view and the same observations in the source representation and in a UMAP 2D realization fit directly from that source. $\mathrm{D}_{\mathrm{comp}}$ and R_{nbr} compare hard-10NN relations; silhouette and B/W summarize between-class separation. PH0/PH1/PH2 are mean per-class total persistence after dividing each class, separately in each space, by the 90th percentile of that class's within-class pairwise distances.

Data	d_X	$\mathrm{D}_{\mathrm{comp}}$	R_{nbr}	Sil.: src→2D	B/W: src→2D	PH0: src→2D	PH1: src→2D	PH2: src→2D
MNIST	784	.049	.675	$.042 \to .312$	$.258 \to 4.20$	$122.7 \to 10.80$	$4.82 \to .420$	$.933 \to .0009$
FMNIST	784	.082	.664	$.044 \to .150$	$.659 \to 4.35$	$99.48 \to 8.67$	$2.12 \to .448$	$.247 \to 3.6 \times 10^{-5}$
KMNIST	784	.052	.696	$.009 \to .084$	$.155 \to 1.65$	$137.1 \to 4.86$	$3.25 \to .197$	$.596 \to 5.7 \times 10^{-5}$
CIFAR-10 raw	3072	.262	.905	$-.054 \to -.129$	$.077 \to .186$	$113.7 \to 9.90$	$.765 \to .817$	$.068 \to .0018$
CIFAR-10 DINOv2	768	.027	.752	$.066 \to .462$	$.281 \to 11.19$	$168.1 \to 13.77$	$3.93 \to .576$	$1.15 \to .0010$
KMNIST DINOv2	768	.082	.750	$-.004 \to -.115$	$.097 \to .389$	$123.4 \to 7.38$	$2.68 \to .404$	$.395 \to .0012$
20 Newsgroups	768	.074	.731	$.014 \to .023$	$.140 \to 1.13$	$174.3 \to 12.50$	$3.36 \to .229$	$1.03 \to 4.9 \times 10^{-4}$
AG News	768	.013	.668	$.029 \to .340$	$.069 \to 3.04$	$190.8 \to 8.97$	$5.49 \to .459$	$2.50 \to 1.1 \times 10^{-4}$
Emotion	768	.092	.813	$-.001 \to -.070$	$.038 \to .294$	$150.3 \to 10.42$	$3.95 \to .887$	$1.16 \to .0000$
PBMC3k	50	.092	.886	$.128 \to .561$	$1.89 \to 17.12$	$100.9 \to 12.31$	$1.69 \to .774$	$.288 \to 3.7 \times 10^{-4}$
Pancreas	50	.023	.470	$.340 \to .345$	$2.38 \to 3.26$	$42.14 \to 4.58$	$.690 \to .134$	$.049 \to .0010$
Paul15	50	.156	.795	$.009 \to -.024$	$1.04 \to 13.58$	$68.12 \to 8.97$	$1.37 \to .356$	$.250 \to .0008$
Dermatology	34	.035	.457	$.237 \to .723$	$1.55 \to 73.60$	$27.93 \to 4.92$	$1.10 \to .193$	$.146 \to 3.5 \times 10^{-5}$

Under the Q_3 UMAP protocol, eleven of the thirteen conditions have $\mathrm{D}_{\mathrm{comp}} < 0.10$, indicating a well-retained semantic relation at the declared resolution. In all eleven, B/W increases in 2D, whereas silhouette changes less uniformly. The magnitude of the between-class change is heterogeneous: silhouette changes only slightly for 20 Newsgroups and Pancreas and decreases for KMNIST DINOv2 and Emotion. The two diagnostics therefore capture complementary aspects of between-class geometry: B/W tracks class-centroid separation relative to within-class scatter, while silhouette tracks pointwise separation. In the same eleven conditions, the reported PH0/PH1/PH2 total-persistence summaries decrease under the class-local normalization. This pattern documents substantial reorganization of the aggregate class-local persistence profiles without assigning the change to specific topological features. The retention set here is specific to the Q_3 UMAP realizations; the cross-mapper Q_2 audit also reveals mapper sensitivity—for example, KMNIST DINOv2 has $\mathrm{D}_{\mathrm{comp}} = .095/.055/.245/.259$ under UMAP/t-SNE/PaCMAP/TriMap, respectively. Raw-pixel CIFAR-10 ($\mathrm{D}_{\mathrm{comp}} = 0.262$) and Paul15 (0.156) define complementary low-retention boundary regimes; raw-pixel CIFAR-10 is discussed further in Section 4.

D.5 COMPLEMENTARY FIXED-NATIVE REALIZATION CONTROL

Purpose. To isolate geometric freedom in the low-dimensional realization after fixing UMAP's native fuzzy relation, this control holds the native fuzzy graph fixed while varying the realization.

Setup. For MNIST, UMAP is run with $n_{\mathrm{neighbors}} = 15$ and `min_dist`$\in \{0, .01, .1, .5, .9\}$. The native fuzzy graph is verified to be identical across all five realizations. The PH summaries in this

control use one global 90th-percentile pairwise-distance scale per geometric space, with the source and each 2D realization normalized independently; this differs from the class-local normalization used in the primary audit.

Result. Figure 6 and Table 12 show that, even with an identical native fuzzy graph, varying `min_dist` substantially changes silhouette, B/W, and PH1 total persistence. Tighter packing amplifies separation and reduces the reported internal-persistence summary, whereas looser packing attenuates the effect. Thus the realization step itself has substantial geometric freedom. The control fixes the native fuzzy relation while allowing the output hard-kNN relation $\mathbf{T}_{\mathbf{Z}}$ to vary, complementing the exact-$\mathrm{D}_{\mathrm{comp}} = 0$ witness above.

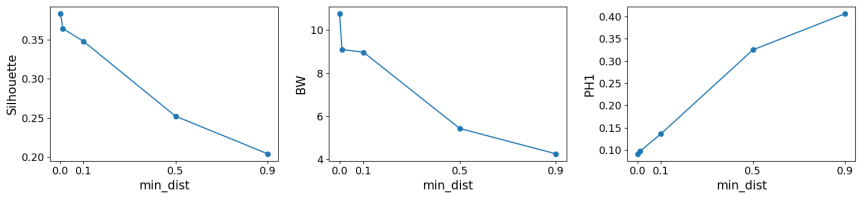


Figure 6: **Complementary fixed-native realization control (MNIST).** UMAP's fuzzy graph is identical across the plotted realizations while `min_dist` varies. The panels show silhouette, B/W, and mean per-class PH1 total persistence under the global per-space normalization used for this control; the full diagnostics are reported in Table 12.

Table 12: Q_3 **fixed-native control: geometry-profile diagnostics (MNIST).** $\overline{\mathrm{PH1}}$ and $\overline{\mathrm{PH2}}$ are mean per-class total persistence after one global 90th-percentile pairwise-distance normalization per geometric space. All 2D rows share the same native UMAP graph.

Data	Space	$\mathrm{D}_{\mathrm{comp}}$	R_{nbr}	Sil.	B/W	$\overline{\mathrm{PH1}}$	$\overline{\mathrm{PH2}}$
MNIST	$\mathbf{X}$	–	–	.042	.258	4.439	.870
	$\mathbf{Z}$, min_dist=.00	.048	.680	.383	10.76	.092	.0001
	$\mathbf{Z}$, min_dist=.01	.048	.683	.364	9.10	.098	.0000
	$\mathbf{Z}$, min_dist=.10	.046	.693	.348	8.97	.136	.0001
	$\mathbf{Z}$, min_dist=.50	.061	.761	.252	5.43	.325	.0004
	$\mathbf{Z}$, min_dist=.90	.093	.811	.204	4.26	.406	.0002

D.6 Same-dimensional UMAP diagnostic

Purpose. To test whether analogous changes in geometry summaries can occur without changing the ambient dimension, we include a complementary same-dimensional diagnostic.

Setup. We illustrate this diagnostic on MNIST by fitting UMAP directly from the declared source representation $\mathbf{X} \in \mathbb{R}^{n \times d_X}$ with $n_{\mathrm{components}} = d_X$, using the same observations and graph-construction parameters as the corresponding 2D fit. The d_X-dimensional and 2D realizations are parallel fits from $\mathbf{X}$, not a sequential pipeline.

Result. The reported geometry summaries can change substantially even when the ambient dimension is unchanged. On MNIST, mean per-class PH1 total persistence changes from 4.825 in the 784-dimensional source representation to 0.018 in the 784-dimensional UMAP realization. This complementary diagnostic shows that substantial changes in the reported summaries can arise even when the ambient dimension is unchanged.

Table 13: **Supervised, in-sample matched CIFAR five-class targets.** Under the same deterministic CPU training protocol, both author-declared coherent and deliberately arbitrary pairings attain favorable values on several diagnostics with respect to their own five-class training targets; the table reports descriptive multi-seed summaries rather than a statistical equivalence test. The final row evaluates the same 2D coordinates against the original ten CIFAR-10 classes, a distinct target from the preceding silhouette row.

Metric	Coherent pairing	Arbitrary pairing
Linear access to 5-class target $\mathrm{Lin}(\mathbf{X}; Y_5)$	0.99984 ± 0.00036	0.98500 ± 0.02056
$\mathrm{D}_{\mathrm{comp}}$ for 5-class target	0.000088 ± 0.000197	0.001500 ± 0.002105
Exact-neighbor replacement	0.671 ± 0.090	0.698 ± 0.104
2D silhouette on 5-class training target	0.803 ± 0.034	0.755 ± 0.113
2D silhouette on original 10 classes	0.0021 ± 0.0012	0.0027 ± 0.0020

E Limits and Positioning

The preceding appendices separate information, semantic-relation retention, and geometry. This final appendix records an outer boundary case for internal embedding diagnostics and positions the three questions relative to neighboring work.

E.1 Outer boundary: matched CIFAR

Matched CIFAR is a supervised, in-sample existence construction: the supervised-contrastive encoder (Khosla et al., 2020) is trained and audited on the same stratified sample of 5,000 CIFAR-10 observations (Krizhevsky, 2009). The experiment compares two deterministic five-class coarsenings of the original ten-class target.

We use *coherent* as an author-declared semantic contrast; scientific meaning is intentionally treated as evidence external to this construction. The coherent pairing is

$$\{\text{cat,dog}\},\ \{\text{deer,horse}\},\ \{\text{automobile,truck}\},\ \{\text{airplane,ship}\},\ \{\text{bird,frog}\},$$

whereas the deliberately *arbitrary* pairing is

$$\{\text{airplane,frog}\},\ \{\text{automobile,dog}\},\ \{\text{bird,truck}\},\ \{\text{cat,horse}\},\ \{\text{deer,ship}\}.$$

Let Y_{10} denote the original CIFAR-10 class and let $Y_5 = h(Y_{10})$ denote the declared five-class coarsening defined above. Both targets are therefore deterministic functions of the original CIFAR-10 class label and share the same label provenance. Because the stratified 5,000-image cohort contains 500 observations from each original class, each five-class target is exactly balanced with 1,000 observations per group and satisfies $\widehat{I}(Y_{10}; Y_5) = \widehat{H}(Y_5) = \log_2 5$ under the empirical distribution of this cohort. The experiment isolates the evidential role of internal embedding diagnostics relative to externally supplied semantic meaning. Once an upstream transformation f induces a strong grouping, clean downstream geometry and favorable internal diagnostics faithfully display that grouping, while its scientific meaning requires complementary external evidence. The same evidential distinction applies to unsupervised f, where the map displays the regularity emphasized by the representation while scientific relevance remains a separate evidential question.

The matched training pipeline is identical across the two targets. Raw pixels are reduced by PCA to 50 dimensions and standardized; a $50 \to 256 \to 128 \to 64$ MLP produces ℓ_2-normalized representations and is trained with the supervised-contrastive objective (temperature 0.1), Adam with learning rate 10^{-3}, batch size 256, and 200 epochs. The reported release run uses deterministic CPU training. Encoder seeds are 42–46. Each representation is mapped with UMAP using $n_{\mathrm{neighbors}} = 10$, `min_dist`$= .1$, Euclidean distance, 200 epochs, and map seeds 42–44. UMAP-seed repetitions are averaged within encoder seed before reporting mean$\pm$SD across the five encoder seeds. Any held-out split used for the linear accessibility readout applies only to the probe; representation learning remains in-sample because the encoder has already been trained using the five-class targets on this cohort. Table 13 reports the resulting multi-seed diagnostics.

Table 14: **Novelty map across neighboring questions.** Columns summarize class/cluster-fidelity work by Jeon et al. (2022; 2024), Chari & Pachter (2023), Lause et al. (2024), and inverse/recoverability work by Bergam et al. (2026); Draganov & Dohn (2023). ✓ = directly targeted, ∼ = conceptually related but not the same inferential target, and – = not the primary target.

	Class/cluster	Chari–P.	Lause et al.	Inverse / recov.	**Ours**
Same fixed Z compared under multiple post-hoc semantic readings	∼	∼	–	∼	✓
High-dimensional predictive/information accessibility versus 2D semantic expression	∼	∼	∼	–	✓
High-dimensional information ordering can reverse in visible expression	–	∼	–	–	✓
Similar visible expression, different semantic-information states	–	∼	–	–	✓
Directed neighbor-class destination distribution, including off-diagonals	∼	∼	∼	–	✓
Geometry conditional on verified semantic-relation retention	–	–	–	∼	✓

E.2 Novelty map across neighboring questions

Table 14 compares our three questions against the inferential targets of neighboring work. Prior results characterize geometric fidelity, cluster correspondence, label-level robustness, and limits on what can be recovered from low-dimensional embeddings. These directions address important parts of the high-dimensional-to-2D relationship, but they do not jointly separate the backward inferential chain studied here: whether source information is reflected in visible expression (Q_1), whether the semantic relation carried by neighborhoods survives even when exact neighbors do not (Q_2), and whether a surviving relation implies preservation of the geometry through which it is realized (Q_3). Our contribution is to show that these successive inferential targets can come apart.

Table 14 summarizes these distinctions.

The distinction is therefore not simply between faithful and distorted embeddings. Prior work studies how geometric, cluster-level, or label-level structure survives or can be recovered; we separate three successive inferential targets—source information, retained semantic relations, and the geometry through which those relations appear—and show that each can decouple from the one before it. Recent documentation of practical misuse (Jeon et al., 2026) further motivates why keeping these targets separate matters in applied analysis.