OPERATIONAL ABSTRACTIONS OF NEURAL NETWORK CONCEPTS VIA TOPOLOGICAL REPRESENTATIONS

Anonymous authors
Paper under double-blind review

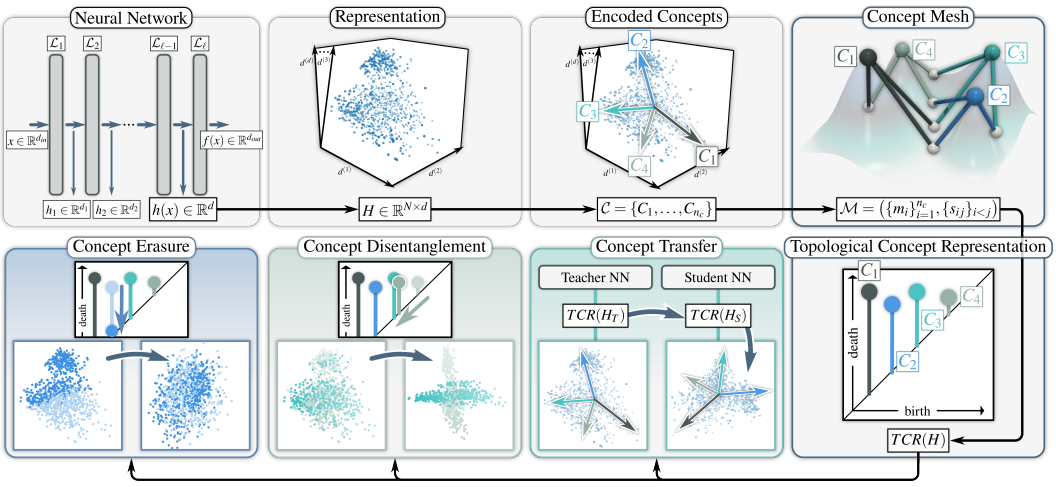


Figure 1: TCR provides a common post-hoc abstraction for concept organization in learned representations. It maps encoded concepts and their interactions to a concept mesh and summarizes its topological structure. Desired changes are specified on this abstraction and propagated back to the representation, enabling diverse interventions such as concept erasure, disentanglement and transfer.

ABSTRACT

Concept-based methods provide a semantic level for interpreting and manipulating learned representations, but existing editing approaches are typically specialized to particular interventions and do not provide a common and editable representation of concept organization. To achieve this, we introduce *Topological Concept Representations* (TCR), a post-hoc operational abstraction that jointly characterizes the concepts encoded in a learned representation and their relationships. TCR constructs an intermediate concept space from concept recoverability and interaction scores, and compactly encodes its organization through topology. Interventions are expressed through modifications of this abstraction that are propagated back to the underlying learned representation. This allows different concept-level operations to share the same optimization framework and separates the desired concept organization from the mechanism to achieve it. We establish stability and reparameterization-invariance properties of TCR and its connections to existing concept-editing formulations. We use TCR to disentangle concepts as a preprocessing step for existing erasure methods, improving worst-group accuracy by 21.89 on average at comparable concept leakage. We further use TCR to transfer concepts from teacher to student models, improving concept recoverability by up to 5.54 while also improving or maintaining competitive test top-1 accuracy.

1 INTRODUCTION

As computational resources and learning methods continue to advance, increasingly large and complex models have become trainable (Kaplan et al., 2020), giving rise to rich and complex learned

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.

representations (Ansuini et al., 2019). Their internal organization is often not directly interpretable, making it difficult to understand and manipulate them in terms of semantically meaningful properties we wish to reason about. A natural way to address this is to seek higher-level abstractions of these representations that would capture such properties in a compact way. This notably raises the question of which *operational abstractions* can be defined for such representations, that, in addition of capturing these properties, would offer a principled and unified interface to manipulate them.

The concept-based literature provides a natural semantic level for reasoning about learned representations by identifying human-interpretable properties encoded within them (Bau et al., 2017; Kim et al., 2018; Fel et al., 2023a). Beyond interpretation, such concepts can also be manipulated, notably by erasing unwanted information from a representation (Ravfogel et al., 2020; 2022; Belrose et al., 2023), preserving downstream task information during erasure (Holstege et al., 2024; 2025; Akbari et al., 2025) or transferring concept-related information across models (Gupta et al., 2023). However, these methods are typically developed as independent frameworks tailored to a particular intervention and optimization algorithm. Consequently, changing the intended operation or introducing additional requirements often requires deriving a new formulation and optimization procedure rather than simply specifying a different desired outcome. Specifically, they are not formulated within a common framework in which concepts and their interactions are jointly characterized and different concept-level objectives can be expressed, composed and optimized. Moreover, existing methods generally do not treat concept relationships as quantities that can be adjusted to influence the outcome of an intervention. This fragmentation motivates a common post-hoc representation in which concept organization itself can be described, compared and manipulated in a unified way.

Topological Data Analysis (TDA) provides natural tools for constructing such abstractions (Edelsbrunner & Harer, 2009; Zomorodian, 2010). The resulting objects compactly characterize the global organization of their input, while their multiscale nature accounts for relationships of varying strengths without committing to a single arbitrary threshold. They can also be compared in a principled manner and, beyond their descriptive role, can serve as *operational proxies*: when coupled with appropriate optimization procedures, modifications expressed on the abstraction can be translated into controlled modifications of the underlying input (Carriere et al., 2021).

We build on these properties to introduce *Topological Concept Representations* (TCR), a post-hoc operational abstraction of the concept organization encoded in learned representations. TCR first constructs an intermediate *concept space* in which individual concepts are represented together with their relationships, characterized by scalar scores derived from the learned representation. This space is then compactly encoded through a topological abstraction, which provides an interface for acting on the concept organization and, through it, on the underlying learned representation. Concretely, an intervention is specified by modifying this abstraction to define a target TCR, and optimizing a transformation of the learned representation to realize it. In this way, different operations are expressed through different target topologies within a shared optimization framework, thereby separating *what* concept organization is desired from *how* the representation is modified to achieve it.

We evaluate TCR on complementary concept editing and transfer tasks. For concept erasure, we consider both direct erasure with TCR and its use as a preprocessing step that disentangles unwanted concepts from task-relevant information before applying existing erasure methods; across datasets and baselines, this improves test WGA by 21.89 on average at comparable levels of concept leakage. For concept transfer, we use the TCR of a teacher representation as a target for a student and show that the transfer generalizes beyond the concepts explicitly included in the optimization, improving concept recoverability by up to 5.54 while improving or maintaining competitive test top-1 accuracy.

Contributions. Our main contributions are as follows:

- We introduce TCR, a post-hoc topological abstraction that jointly represents concepts and their relationships and provides a common interface for concept-level interventions.
- We establish theoretical properties of TCR such as stability, invariance to representation reparameterization and its connections to existing concept-editing formulations.
- We evaluate TCR across concept erasure, disentanglement and transfer, showing improved task preservation during erasure and improved transfer of concept recoverability.

Related work. We provide an extended discussion of concept representation and intervention methods as well as topological abstractions and persistence-based optimization in Appendix A.

2 PRELIMINARIES

Concepts and probes. Given a representation function $h : \mathcal{X} \to \mathbb{R}^d$ and a collection of N samples $\{x_1, \ldots, x_N\}$, we denote their representations by $H = [h(x_1), \ldots, h(x_N)]^\top \in \mathbb{R}^{N \times d}$. A concept C_i is described by a vector $\mathbf{c}_i = (c_{i,1}, \ldots, c_{i,N})^\top$, indicating its presence or strength for each sample. Such concept annotations may be provided explicitly or obtained through a concept discovery procedure. A common way to characterize how a concept is encoded in a representation is through a probe $q_i : \mathbb{R}^d \to \mathbb{R}$ predicting $c_{i,n}$ from $h(x_n)$. For instance, a linear probe $q_i(h) = \mathbf{w}_i^\top h + b_i$ associates the concept with a direction $\mathbf{w}_i$ in the representation space. The ability of such a probe to predict $\mathbf{c}_i$ can then be used to quantify how strongly the concept is encoded in the representation.

Persistence diagrams. Let a scalar field $f : V \to \mathbb{R}$ defined on the vertex set V of a simplicial complex $\mathcal{K}$. The topological features of f can be identified using the *lower-star filtration* and encoded in the persistence diagram (Edelsbrunner & Harer, 2009). Let $\mathbf{f} = \big(f(v_1), \ldots, f(v_{n_v})\big) \in \mathbb{R}^{|V|}$ be the vector of vertex values of f and let the *filtration map* $\mathcal{F} : \mathbb{R}^{|V|} \to \mathbb{R}^{|\mathcal{K}|}$ mapping a vector of vertex values to a $\mathbb{R}^{|\mathcal{K}|}$-dimensional vector of simplex values where $\mathcal{F}_\sigma(\mathbf{f}) = \max_{v \in \sigma} f(v)$ for each simplex $\sigma \in \mathcal{K}$, ensuring that $\mathcal{F}_\tau(\mathbf{f}) \leq \mathcal{F}_\sigma(\mathbf{f})$ for any $\tau \subseteq \sigma$. The lower-star filtration is the nested family of subcomplexes $\mathcal{K}_a = \{\sigma \in \mathcal{K} \mid \mathcal{F}_\sigma(\mathbf{f}) \leq a\}$ with $\mathcal{K}_a \subseteq \mathcal{K}_{a'}$ for any $a \leq a'$. Assuming f injective on the vertices of $\mathcal{K}$, obtained easily with symbolic perturbations (Edelsbrunner & Mucke, 1990), the filtration changes only at the $n_v = |V|$ distinct vertex values, represented by the finite sequence $\emptyset \subset \mathcal{K}_{f(v_1)} \subset \ldots \subset \mathcal{K}_{f(v_{n_v})} = \mathcal{K}$ (such that $f(v_1) < \ldots < f(v_{n_v})$) where each step adds a vertex v as well as its lower star (i.e. each simplex σ containing v such that $\mathcal{F}_\sigma(\mathbf{f}) = f(v)$).

Topological changes, such as the appearance and merging of connected components, occur at specific filtration values. When two components merge, the one created later *dies* according to the *Elder rule* (Edelsbrunner & Harer, 2009) and defines a *persistence pair* (v, v'), where v creates the component and v' merges it; $f(v)$ and $f(v')$ are respectively its *birth* and *death*, and $f(v') - f(v)$ its *persistence*, commonly interpreted as a measure of importance. The component born at the global minimum never dies, and we conventionally assign it the global maximum as death value. The *persistence diagram* $\mathcal{D}(f)$ is the multiset of points $\big(f(v), f(v')\big)$ associated with all persistence pairs.

Wasserstein distances. Two persistence diagrams can be compared using the Wasserstein distance, which seeks a minimum-cost matching between their points. Each off-diagonal point can either be matched to exactly one point of the other diagram, associating two topological features, or to the diagonal, corresponding to the insertion or deletion of a feature. In practice, for diagrams $\mathcal{D}(f_i)$ and $\mathcal{D}(f_j)$, each is augmented with the diagonal projections of the off-diagonal points of the other, e.g. $\mathcal{D}'(f_i) = \mathcal{D}(f_i) \cup \{\Delta(p_j) \mid p_j \in \mathcal{D}(f_j)\}$, where $\Delta(x, y) = ((x + y)/2, (x + y)/2)$. This accounts for the infinite multiplicity of the diagonal and ensures $|\mathcal{D}'(f_i)| = |\mathcal{D}'(f_j)|$, so that the distance can be computed as a bijective matching between the augmented diagrams.

Two points $p_i = (x_i, y_i) \in \mathcal{D}'(f_i)$ and $p_j = (x_j, y_j) \in \mathcal{D}'(f_j)$ are mapped with a cost of $d_q(p_i, p_j)^q = |x_i - x_j|^q + |y_i - y_j|^q = \|p_i - p_j\|_q^q$, with $q \geq 1$ and where $d_q(p_i, p_j)$ is set to zero if both p_i and p_j are diagonal points. The L^q-Wasserstein distance W_q is then defined as:

$$W_q\big(\mathcal{D}(f_i), \mathcal{D}(f_j)\big) = \min_{\phi \in \Phi} \Big(\sum_{p_i \in \mathcal{D}'(f_i)} d_q\big(p_i, \phi(p_i)\big)^q \Big)^{1/q}, \tag{1}$$

where Φ is the set of all possible assignments ϕ mapping each point $p_i \in \mathcal{D}'(f_i)$ to exactly one point $p_j \in \mathcal{D}'(f_j)$ being either an off-diagonal point or the diagonal projection of p_i (i.e. $\phi(p_i) = \Delta(p_i) = p_j \in \mathcal{D}'(f_j)$) indicating its deletion. W_q can be computed using traditional assignment problem solvers (Munkres, 1957; Bertsekas, 1981; Cuturi, 2013; Kerber et al., 2017).

Persistence-based topological optimization. Persistence diagrams can be incorporated into differentiable optimization objectives (Carriere et al., 2021). Let $\mathcal{P} : \mathbb{R}^{|\mathcal{K}|} \to \mathbb{R}^{|\mathcal{K}|}$ denote the *persistence map*, being a permutation of the simplex filtration values returned by $\mathcal{F}$ according to the persistence pairing, such that the birth and death values of each pair are placed consecutively (including zero-persistence pairs). Gradients with respect to birth and death values can therefore be propagated back to the simplices inducing them and, through $\mathcal{F}$, to the vertex values defining the filtration. A persistence-based objective can then be written as $\mathcal{L} = \mathcal{E} \circ \mathcal{P} \circ \mathcal{F}$, where $\mathcal{E}$ is a differentiable en-

ergy defined on the persistence representation, such as one derived from the Wasserstein distance of Eq. 1. Under mild regularity conditions, such objectives are differentiable almost everywhere, admit well-defined generalized subgradients at non-smooth points and can be optimized with stochastic subgradient methods with convergence guarantees to critical points (Carriere et al., 2021).

3 Topological Concept Representations

3.1 Overview and Problem Formulation

Given N samples $\{x_1, \dots, x_N\}$, let $H = [h(x_1), \dots, h(x_N)]^\top \in \mathbb{R}^{N \times d}$ denote their representations, such as those extracted at a selected model layer, and let $\mathcal{C} = \{C_1, \dots, C_{n_C}\}$ be a collection of concepts defined on these samples. Our objective is to construct an abstraction of H that jointly characterizes which concepts are encoded in the representation and how these concepts interact with each other, and that can subsequently be used as an operational proxy to modify H. To this end, we introduce *Topological Concept Representations* (TCR). We associate to each concept C_i a scalar value m_i characterizing how strongly C_i is encoded in H, and to each pair of concepts (C_i, C_j) an interaction value s_{ij}. These values define a scalar field on a *concept mesh* in which concepts form maxima and their interactions form saddle vertices. The persistence diagram of this scalar field, the TCR of H, summarizes both the presence of individual concepts and their organization with respect to the other concepts. Beyond being descriptive, the resulting abstraction permits to express a desired intervention by defining a *target TCR*, e.g., by reducing or increasing a concept maximum or a saddle connecting two concepts. Considering a parameterized transformation $T_\theta : \mathbb{R}^d \to \mathbb{R}^{d'}$, applied row-wise to H as $T_\theta(H) := [T_\theta(h_1), \dots, T_\theta(h_N)]^\top \in \mathbb{R}^{N \times d'}$, we optimize θ such that the TCR of $T_\theta(H)$ matches this target. Hence, different concept manipulations can be formulated as modifications of a common abstraction rather than through operation-specific objectives.

3.2 Constructing Topological Concept Representations

Given a representation $Z \in \mathbb{R}^{N \times d'}$, we first associate scalar values with individual concepts, then measure their pairwise interactions and combine them into a topological abstraction.

Concept scores. At the framework level, TCR represents simply a concept score by a scalar $m_i = S(Z, C_i)$ whose value increases with how much the concept C_i is encoded in the representation. While different choices of S can instantiate different notions of concept encoding, in this work we primarily quantify it through *recoverability*. Let $\mathbf{c}_i = (c_{i,1}, \dots, c_{i,N})^\top$ denote the sample-level scores of C_i, and let $Z = [z_1, \dots, z_N]^\top \in \mathbb{R}^{N \times d'}$ denote the representation on which the concept score is evaluated. Given a probe q_{ϕ_i} and a prediction loss ℓ, we define the concept score as:

$$m_i = \mathcal{L}_i^0 - \mathcal{L}_i^\star, \qquad \mathcal{L}_i^0 = \min_a \frac{1}{N} \sum_{n=1}^{N} \ell(c_{i,n}, a), \qquad \mathcal{L}_i^\star = \min_{\psi_i} \frac{1}{N} \sum_{n=1}^{N} \ell(c_{i,n}, q_{\psi_i}(z_n)), \quad (2)$$

where ψ_i denotes the fitted probe parameters and $\mathcal{L}_i^0$ is the loss of the optimal constant predictor, which does not use Z and thus provides a baseline for prediction without information from the representation. Thus, m_i is the improvement of the fitted probe over the optimal constant predictor where $m_i = 0$ indicates no improvement over the null predictor, while larger values indicate that C_i is more readily predicted from Z. We consider binary cross-entropy (BCE) with a logistic probe for binary concepts, and mean squared error (MSE) with a least-squares probe for binary or continuous concepts. The score m_i can optionally be normalized by $\mathcal{L}_i^0$, yielding a dimensionless score facilitating the joint treatment of concepts evaluated with different prediction losses, e.g., BCE for binary and MSE for continuous concepts. For MSE, it corresponds to an R^2-type explained-variance score. The MSE formulation additionally admits a convenient analytical form described in Appendix B.

Concept interactions. Individual concept scores characterize how much a concept is encoded in a representation, but not how concepts relate to one another. We therefore associate every pair (C_i, C_j) with a normalized interaction score $\rho_{ij} \in [0, 1]$, where $\rho_{ij} = 0$ indicates no interaction according to the chosen criterion and $\rho_{ij} = 1$ indicates maximal interaction. While some interaction criteria are naturally symmetric, others characterize a directional relationship $r_{i \to j} \in [0, 1]$. Directional scores can be combined to associate each concept pair with a single interaction value: $\rho_{ij} = \Psi(r_{i \to j}, r_{j \to i})$, where $\Psi : [0, 1]^2 \to [0, 1]$ is a symmetric aggregation function, such as $\Psi(a, b) = (a + b)/2$.

We principally instantiate this construction through *interference*. For instance, a representation may strongly encode both an undesired concept and the task while relying on shared information to encode them, such that removing the former also reduces the score of the latter. Inspired by Naowarat et al. (2026), who characterize interference through the prediction error of one concept after erasing another, we measure directional interference through the relative decrease in the concept score of C_j after removing C_i. Let $\mathcal{E}_i(Z)$ denote the representation obtained by erasing concept C_i from Z, and let $m_j(Z)$ denote the score of C_j evaluated on Z. We define the directional interference as:

$$r_{i \to j} = 1 - \frac{m_j\big(\mathcal{E}_i(Z)\big)}{m_j(Z)}. \tag{3}$$

Intuitively, r_{ij} is high when removing C_i substantially reduces the recoverability score of C_j, and low when C_j remains largely unaffected by this intervention. We instantiate $\mathcal{E}_i$ with a differentiable implementation of LEACE (Belrose et al., 2023). Consequently, Eq. 3 characterizes interactions through the effect of an explicit intervention, rather than through their correlation alone. For MSE concept scores, interference admits an equivalent closed-form expression, described in Appendix B.

From concept scores and interactions to topology. To jointly represent concept scores and interactions, we first express both quantities on a common scalar scale. While m_i directly quantifies concept C_i, the interaction score $\rho_{ij} \in [0, 1]$ is normalized and therefore only specifies the relative interaction between C_i and C_j. We seek an absolute interaction level that increases with ρ_{ij}, vanishes when $\rho_{ij} = 0$, and cannot exceed the score of either concept: an interaction between two concepts should not be assigned a larger value than the extent to which either of them is itself represented. Accordingly, we rescale ρ_{ij} by the largest admissible value $\min(m_i, m_j)$ and define:

$$s_{ij} = \min(m_i, m_j)\rho_{ij}. \tag{4}$$

Thus, s_{ij} varies continuously from 0 for no interaction to $\min(m_i, m_j)$ for maximal interaction. Intuitively, the difference $m_i - s_{ij}$ represents how much C_i is *independent* from C_j.

We use these quantities to define a scalar field on a graph $\mathcal{M}$ that we refer to as the *concept mesh*. The graph contains one concept vertex v_i of value m_i for each C_i, together with a connector vertex v_{ij} of value s_{ij} connected to v_i and v_j. Since $s_{ij} \leq \min(m_i, m_j)$, concept vertices form maxima of this scalar field, while connector vertices encode their pairwise interactions below the corresponding concept values. We additionally introduce a common vertex with value below all other vertices to assign a finite reference level to the component that survives all pairwise merges.

We apply the upper-star counterpart of the filtration introduced in Sec. 2, processing vertices from highest to lowest values. Concept maxima create connected components and connector vertices merge them; following the Elder rule, we associate C_i with $p_i = (s_i, m_i)$, where s_i is the connector value at which its component merges with an older one. Finally, $TCR(Z) = \{p_i\}_{i=1}^{n_C}$ is the persistence diagram defining the Topological Concept Representation of Z. This representation enables Wasserstein comparisons and differentiable topological optimization, which we exploit next.

3.3 Editing Topological Concept Representations

Having constructed a topological description of the concept organization, we now turn it into an operational interface to specify interventions on a representation H. Let $Z_\theta = T_\theta(H)$ denote the representation obtained by applying a differentiable transformation $T_\theta : \mathbb{R}^d \to \mathbb{R}^{d'}$ row-wise to H. Starting from $TCR(H)$, an intervention is specified by defining a target $TCR^\star$ through changes to its concept maxima and interaction structure. Given a pair $(s^\star_{\mathcal{AB}}, m^\star_i)$ in $TCR^\star$, examples include:

- $m^\star_i \downarrow$ / $m^\star_i \uparrow$: weaken / strengthen concept C_i; while $m^\star_i = m_i$ preserves it;
- $s^\star_{\mathcal{AB}} \downarrow$ / $s^\star_{\mathcal{AB}} \uparrow$: disentangle / entangle the sets of concepts $\mathcal{A}$ and $\mathcal{B}$ that merged at $s^\star_{\mathcal{AB}}$;
- $TCR^\star = TCR(Z^\star)$: transfer the entire concept organization from a representation $Z^\star$.

These operations can be combined within the same target, e.g., suppressing an undesired concept while preserving the task and selected concepts. Importantly, the target specifies an *editing criterion*, rather than an editing algorithm. Concept erasure, for instance, can be expressed by setting $m^\star_i = 0$ while preserving protected concepts. More generally, varying the concept score S and target constraints allows TCR to capture a broad class of existing erasure criteria at the objective level, independently of the optimization used to enforce them; we further examine this relation in Sec. 3.4.

Topological objective. Given $TCR(Z_\theta)$ and $TCR^\star$, we measure how closely the transformed representation Z_θ realizes the desired concept organization using the L^q-Wasserstein distance (Sec. 2):

$$\mathcal{L}_{TCR}(\theta) = W_q\big(TCR(Z_\theta), TCR^\star\big)^q. \tag{5}$$

When concept correspondences are known (possibly only on a subset) this semantic information can be exploited by constraining the Wasserstein assignment to match the corresponding features.

Topological optimization. To realize the target, we optimize θ through persistence-based topological optimization (Sec. 2). For a given Z_θ, we refit the concept probes, reconstruct the concept mesh and its TCR. The gradient of Eq. 5 is propagated through the persistence map $\mathcal{P}$ to the filtration values of the concept mesh, and through the filtration map $\mathcal{F}$ to the concept maxima and interaction saddles involved in the matched persistence features. Since these values are differentiable functions of Z_θ, through the differentiable concept scores and interactions, the gradient further propagates through $Z_\theta = T_\theta(H)$ to update θ. The framework is agnostic to the parameterization of T_θ, we consider post-hoc linear or nonlinear transformations, or directly the network producing H.

Concept disentanglement. Given two sets of concepts $\mathcal{A}$ and $\mathcal{B}$, we define their structural disentanglement by lowering the persistence saddles at which components containing concepts from respectively each set merge. Several such merges may occur at different locations of the hierarchy, involving different subsets of $\mathcal{A}$ and $\mathcal{B}$; their saddle coordinates are jointly targeted while selected concept maxima can be preserved. As the TCR is recomputed during optimization, the selected saddles adapt to the current concept organization. This operation therefore edits the hierarchical connectivity between the two concept sets, rather than a prescribed individual pairwise interaction.

Concept transfer. Given two representations Z^A and Z^B, possibly produced by different models or having different dimensionalities, $TCR(Z^A)$ can serve as the target for Z^B. Thus, concept transfer replaces a manually specified target by the concept organization from another representation.

3.4 Properties of Topological Concept Representations

***(1)* Benefits of the persistence reduction.** Rather than reproducing every pairwise interaction of the concept mesh, TCR summarizes how such local interactions combine globally to form groups of concepts and how they organize across interaction levels, beyond what individual interaction scores reveal in isolation. Among the $\binom{n_C}{2}$ interactions, the ones being responsible for the concept hierarchy correspond exactly to the $n_C - 1$ saddles computed by persistence. TCR therefore provides a compact multiscale description of the concept organization by identifying the minimum sparse set of interactions through which this organization can be modified. Consequently, it is invariant to changes not affecting the concept hierarchy in contrast to concept meshes that can differ substantially entry-wise while still inducing the same concept hierarchy. Representing this organization as a persistence diagram provides established tools for Wasserstein comparison and differentiable optimization, together with stability and convergence properties, whereas operating directly on concept meshes instead requires additional choices of correspondence, comparison and optimization criteria.

***(2)* Relation to existing erasure objectives.** With appropriate choices of the concept scores, several existing erasure criteria, such as those of the baselines considered in our experiments, arise as special cases of TCR at the objective or constraint level. Details and proofs are given in Appendix E.

***(3)* Affine reparameterization invariance.** For the MSE–overlap instantiation, TCR depends on the intrinsic concept organization represented rather than on a particular affine parameterization of the feature space. Hence, TCR is invariant to invertible affine feature changes, see Appendix F.

Proposition 1 (Affine reparameterization invariance)**.** *Let $Z \in \mathbb{R}^{N\times d}$ be a representation of N samples and let $\widetilde{Z} = ZB^\top + \mathbf{1}a^\top$ with $B \in \mathbb{R}^{d\times d}$ invertible and $a \in \mathbb{R}^d$. Assume that TCR is constructed using the MSE concept scores of Eq. 7 (Appendix B) and the predictive-overlap interactions of Eq. 9. Then $TCR(\widetilde{Z}) = TCR(Z)$. The same result holds for normalized MSE concept scores.*

***(4)* Stability.** Finally, small perturbations of the representation induce proportionally small changes in its TCR under mild local regularity conditions. Proofs and discussions are in Appendix G.

Theorem 1 (Representation-to-TCR stability)**.** *Given $H \in \mathbb{R}^{N\times d}$. Assume that $\Sigma(H) \succ 0$ and that every concept represented in the TCR has strictly positive MSE recoverability, $m_i(H) > 0$ for every $i = 1, \dots, n_C$. Then there exist $\varepsilon > 0$ and $C > 0$ such that, for every centered representation H' satisfying $\|H - H'\|_F < \varepsilon$, $d_B\left(TCR(H), TCR(H')\right) \le C\,\|H - H'\|_F$ with $d_B = W_\infty$.*

4 EXPERIMENTS & RESULTS

We evaluate TCR on two applications: *concept disentanglement before erasure*, studying whether modifying concept interactions can facilitate the removal of undesired information while preserving task utility, and *concept transfer*, investigating whether the concept organization of a teacher model can be transferred to a student. We organize our experiments around the following questions:

(Q1) *How does TCR disentanglement preprocessing improves existing erasure methods at similar linear concept leakage and can TCR directly realize effective concept erasure?*

(Q2) *How do different experimental settings and data regimes, such as the level of erasure, spurious correlation and label imbalance, affect the task utility preservation TCR?*

(Q3) *Can TCR transfer concept information across models while maintaining task accuracy compared with standard knowledge-distillation methods?*

(Q4) *How does TCR transfer generalize beyond the concepts explicitly optimized and how is this transfer affected by concepts discovered at increasingly local and fine spatial scales?*

4.1 CONCEPT DISENTANGLEMENT BEFORE ERASURE

Experimental setup. We consider as baselines INLP (Ravfogel et al., 2020), LEACE (Belrose et al., 2023), RLACE (Ravfogel et al., 2022), SPLINCE (Holstege et al., 2025), LEOPARD (Saillenfest & Lemberger, 2025), KRaM (Chowdhury et al., 2023) and Obliviator (Akbari et al., 2025). We evaluate on Waterbirds (Sagawa et al., 2020), Bias in Bios (De-Arteaga et al., 2019), DIAL-Mention (Blodgett et al., 2016; Elazar & Goldberg, 2018) and Jigsaw (Borkan et al., 2019). We use frozen ImageNet-pretrained ResNet-50 (He et al., 2016) average-pooling features for Waterbirds and BERT-base (Devlin et al., 2019) final-layer `[CLS]` representations for the three text datasets. Each dataset defines a classification task together with an unwanted concept to erase. All methods operate post-hoc on these fixed pretrained representations reduced to 128 dimensions using PCA fitted on the training set before intervention. More details are provided in Appendix H.1.

(Q1) To compare task preservation after concept removal, we report the test worst-group accuracy (WGA), averaged over 10 seeds, of methods with an adjusted balanced linear concept leakage below 10% in Tab. 1 (metrics described formally in Appendix H.1). Disentangling the task and concept with TCR before erasure consistently improves task utility: TCR-DIS increases WGA for every erasure method on all four datasets. The improvement is particularly pronounced on Waterbirds and Bias in Bios, where the best baseline WGA increases from 27.88 to 74.66 and from 59.53 to

Table 1: Test task worst-group accuracy with an adjusted balanced linear concept leakage $< 10\%$.

Method	**Waterbirds**	**Bias in Bios SN**	**Dial Mention**	**Jigsaw**
Raw (WGA / Leakage)	*56.54 / 80.22*	*78.77 / 98.48*	*39.20 / 54.57*	*53.29 / 41.57*
INLP	27.88±0.00	59.47±0.00	52.05±0.00	55.64±0.00
LEACE	13.71±0.00	59.47±0.00	50.45±0.00	52.59±0.00
RLACE	14.55±0.95	59.53±0.92	50.80±0.29	53.27±0.37
SPLINCE	13.71±0.00	58.98±0.00	50.45±0.00	52.59±0.00
LEOPARD	13.71±0.00	59.47±0.00	49.95±0.19	52.59±0.00
KRaM	9.08±4.01	44.02±3.17	52.25±1.40	53.12±1.41
Obliviator	18.68±3.46	58.18±1.29	50.56±1.48	52.56±1.07
TCR-DIS + INLP	66.01±8.80	81.12±3.85	54.66±1.15	56.71±1.12
TCR-DIS + LEACE	74.45±1.31	81.61±1.26	58.04±0.74	60.58±0.92
TCR-DIS + RLACE	**74.66±0.96**	81.59±1.36	57.78±0.73	**61.05±0.72**
TCR-DIS + SPLINCE	74.17±1.24	81.53±0.97	58.03±0.70	60.56±0.99
TCR-DIS + LEOPARD	74.47±0.89	81.47±1.43	57.83±0.83	60.61±0.87
TCR-DIS + KRaM	55.64±11.32	57.48±5.85	54.23±1.56	56.92±1.27
TCR-DIS + Obliviator	69.30±7.99	83.50±1.17	**58.20±1.73**	60.01±1.21
TCR-ERA	71.17±1.42	**86.52±1.01**	55.70±0.89	57.37±1.37

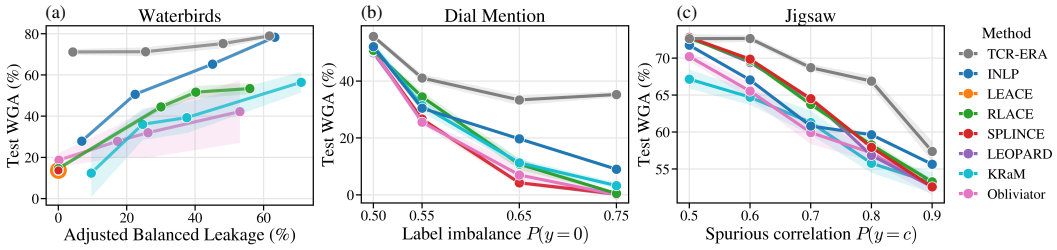


Figure 2: Erasure-utility trade-off showing the test WGA given the adjusted balanced linear leakage (a) and impact on test WGA of varying label imbalance (b) and spurious correlation (c).

83.50, respectively. This also shows that the interference saddle generalizes beyond the erasure operator used to define it, although constructed from LEACE, it facilitates erasure with all considered methods. Moreover, directly optimizing TCR for concept erasure (TCR-ERA) outperforms every standalone erasure baseline across the four datasets and achieves the highest WGA on Bias in Bios. Across all datasets, TCR-DIS improves the WGA of each baseline by 21.89 on average, with large gains on Waterbirds (+53.91) and Bias in Bios (+21.31), while TCR-ERA improves WGA by 23.43.

Note that, following prior work on linear-guardedness, e.g., LEACE (Belrose et al., 2023), concept leakage is measured through a linear probe. We therefore additionally evaluate stronger leakage probes, including oracle-conditioned linear and non-linear probes, in Appendix I.1.1. Additional analysis of global representational similarity before and after erasure is provided in Appendix I.1.3.

Takeaway. TCR disentanglement can effectively serve as preprocessing for a diverse set of erasure methods and systematically facilitates utility-preserving concept erasure. Moreover, TCR can also perform effective task-preserving concept erasure directly within the same framework.

(Q2) To assess utility preservation under different erasure strengths and data regimes, we vary the erasure constraint, label imbalance and spurious correlation. We first examine the erasure-utility trade-off across different erasure strengths. On Waterbirds (Fig. 2a), TCR retains 71.17 WGA at 4.26% adjusted balanced linear leakage, while INLP reaches 27.88 WGA at 6.88%. As the erasure constraint is relaxed, several baselines recover substantial task utility and the gap to TCR narrows, whereas TCR retains high WGA already under strong erasure. We then vary properties of the data distribution while maintaining strong concept erasure. The effect of increasing label imbalance is illustrated on Dial Mention (Fig. 2b), where the margin of TCR over the best baseline at each setting grows from 3.46 WGA points at $P(y = 0) = 0.50$ to 26.27 at $P(y = 0) = 0.75$. For increasing spurious correlation, on Jigsaw (Fig. 2c), TCR is comparable to the best baseline at $P(y = c) = 0.5$ and achieves the highest WGA at every stronger correlation, with a gain of up to 7.25 points at $P(y = c) = 0.8$. Further erasure-utility and data-regime analyses are reported in Appendix I.1.2.

Takeaway. TCR maintains a favorable erasure-utility trade-off across erasure strengths and consistently strong utility preservation relative to the baselines across data regimes, with its advantage particularly pronounced under strong erasure, label imbalance and increased spurious correlation.

4.2 Concept Transfer

Experimental setup. Following the distillation experimental protocol of Tian et al. (2020), we evaluate concept transfer on CIFAR-100 (Krizhevsky & Hinton, 2009), using a ResNet32×4 teacher and ResNet8×4 and ShuffleNetV2 students. We consider as baselines a student trained without distillation (Student), KD (Hinton et al., 2015), FitNet (Romero et al., 2015), SPKD (Tung & Mori, 2019) and TopKD (Kim et al., 2024). For TCR, concepts are discovered from the teacher penultimate representation, from the full images or from 2×2, 3×3 or 4×4 crops (Fel et al., 2023b), and their organization is transferred to the student during distillation. Further details are given in Appendix H.2.

(Q3) To evaluate how TCR transfers concept information while preserving task performance, we report in Tab. 2 test top-1 accuracy and concept recoverability (CR k_{20}), being the mean held-out R^2 of linear probes predicting 20 teacher-derived concept coefficients from the student representation, that can be seen as their preservation (detailed in Appendix H.2). For ResNet8×4, both TCR configurations outperform all baselines on both metrics: optimizing 10 concepts (k_{10}) achieves the highest

Table 2: Distillation test top-1 accuracy and concept recoverability.

Method	ResNet8×4		ShuffleNetV2	
	Top-1	CR k_{20}	Top-1	CR k_{20}
Student	72.35±0.24	41.50±0.18	72.58±0.23	52.65±0.23
KD	73.35±0.24	54.07±0.17	74.62±0.19	66.64±0.19
FitNet	73.19±0.23	41.05±0.25	74.25±0.15	53.52±0.45
SPKD	73.00±0.24	54.51±0.19	**76.04±0.15**	**67.62±0.23**
TopKD	73.20±0.28	54.18±0.25	74.98±0.28	66.58±0.14
TCR k_{10}	**73.68±0.47**	57.18±0.06	74.97±0.70	66.83±0.37
TCR k_{20}	73.55±0.28	**60.05±0.35**	75.02±0.54	66.67±0.28

top-1 accuracy (73.68) while k_{20} achieves the highest CR (60.05). Compared with the strongest baseline for each metric, this corresponds to gains of 0.33 for top-1 over KD and 5.54 for CR over SPKD, respectively. Averaged over all baselines, gains are respectively 0.60 and 9.55 points. For ShuffleNetV2, SPKD achieves the strongest results on both metrics while TCR remains close in CR.

Takeaway. TCR can transfer teacher-derived concept information to a student while preserving task performance, simultaneously improving concept recoverability and task accuracy (or maintaining it).

(Q4) To assess whether TCR transfers concept information beyond the concept bank used during optimization, we train TCR with $k_{\text{train}} \in \{2, 5, 10, 20\}$ concepts and evaluate CR using all $k_{\text{eval}} \in \{2, 5, 10, 20\}$ concept banks in Fig. 3. We report the CR gain over SPKD, the best baseline regarding CR, with 0 indicating equal recoverability. Since concept banks for different values of k are independently discovered, settings with $k_{\text{train}} \neq k_{\text{eval}}$ measure transfer to concepts that were not explicitly optimized. For full-image concepts (Fig. 3a), TCR improves over SPKD for every matched $k_{\text{train}} = k_{\text{eval}}$ setting, but not when evaluated on different bank sizes. This behavior changes as concepts become increasingly local: the mean gain when $k_{\text{train}} \neq k_{\text{eval}}$ becomes $+0.48$, $+1.16$ and $+1.39$ for the 2×2, 3×3 and 4×4 crop grids, respectively. Likewise, the mean gain across all training and evaluation banks increases from -0.35 for full-image concepts to $+1.05$, $+1.78$ and $+1.88$ at finer spatial scales. At the two finest scales, larger training banks yield particularly broad transfer; for example, training with k_{20} improves CR on the independently discovered k_{10} bank by 3.48 for the 3×3 grid and 4.94 for the 4×4 grid. These results indicate a transition from primarily bank-specific transfer for global concepts to broader cross-bank transfer as concepts become more spatially localized. One possible interpretation is that global concepts may induce more bank-specific structure, whereas local concepts expose more reusable structure that are shared across independently discovered banks. Part of this effect may also arise because localized concepts are less recoverable for the baselines, leaving more room for improvement.

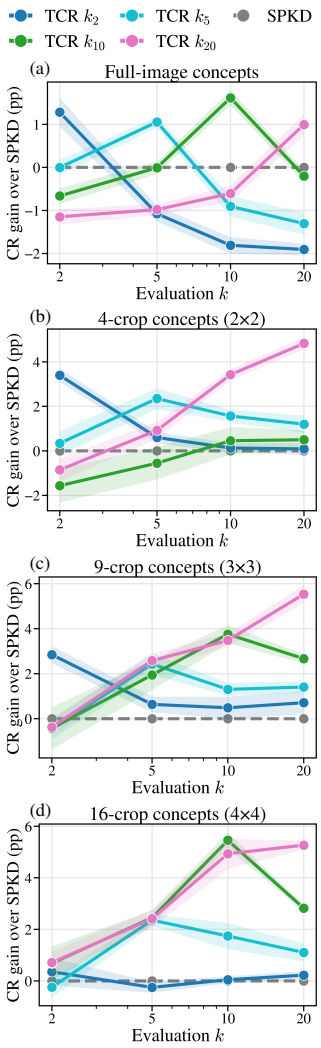


Figure 3: Impact of concept spatial scale on recoverability.

Takeaway. For global concepts, TCR primarily improves recoverability of the concept bank used during optimization. As concepts become increasingly local, its advantage over SPKD grows and extends to independently discovered concept banks of different sizes.

5 Conclusion

We introduced TCR, a topological abstraction that jointly represents concepts and their relationships and provides a common interface for concept-level interventions. Across concept erasure, disentanglement and transfer, our results demonstrate its ability to manipulate concept organization while preserving task-relevant information. For disentanglement, it improves WGA by 21.89 on average at comparable concept leakage and, for transfer, it improves concept recoverability by up to 5.54.

AI USE STATEMENT

In this work, we used generative AI tools for assisting in the writing of proofs, designing or providing feedback on research methodology or experiments, implementing methods, supporting qualitative and thematic data analysis and interpreting results. We have not used generative AI tools for generating synthetic data sets, helping developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, assisting with translation and cleaning and reformatting dataset. Additionally, we used generative AI tools for suggesting experimental parameters, creating or editing software code, summarize or analyse existing literature, brainstorming, sourcing/searching for information, editing a research paper to improve readability and identifying relevant literature. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our experiments use existing benchmark datasets, including human-generated text with demographic, dialectal and identity-related annotations or proxies. We do not collect new human-subject data. Such annotations and proxies may be imperfect and should not be interpreted as definitive characterizations of individuals or groups. While concept erasure can help reduce unwanted information in learned representations, removing a selected concept does not by itself guarantee fairness or complete removal of the corresponding information; we therefore evaluate leakage using complementary diagnostics. More generally, representation interventions can be applied to arbitrary concepts, and their use in high-stakes applications requires careful validation of both the information removed and the utility preserved.

REPRODUCIBILITY STATEMENT

We provide the code used to produce the results of this paper as supplementary material with the anonymous submission. Detailed experimental protocols, including datasets and splits, model architectures, hyperparameter selection and evaluation metrics, are provided in the appendix. Upon de-anonymization, we will make the code publicly available on GitHub.

REFERENCES

Ramin Akbari, Milad Afshari, and Vishnu Boddeti. Obliviator reveals the cost of nonlinear guardedness in concept erasure. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 144860–144892. Curran Associates, Inc., 2025. doi: 10.52202/085713-4847. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/d55f39791f04745b2e0c8abebf3dd5d7-Paper-Conference.pdf.

Alessio Ansuini, Alessandro Laio, Jakob H. Macke, and Davide Zoccolan. *Intrinsic dimension of data representations in deep neural networks*. Curran Associates Inc., Red Hook, NY, USA, 2019.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3319–3327, 2017. doi: 10.1109/CVPR.2017.354.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 66044–66063. Curran Associates, Inc., 2023. doi: 10.52202/075280-2884. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/d066d21c619d0a78c5b557fa3291a8f4-Paper-Conference.pdf.

Dimitri P. Bertsekas. A new algorithm for the assignment problem. *Mathematical Programming*, 21 (1):152–171, 1981. doi: 10.1007/BF01584237.

Su Lin Blodgett, Lisa Green, and Brendan T. O'Connor. Demographic dialectal variation in social media: A case study of african-american english. In Jian Su, Xavier Carreras, and Kevin Duh (eds.), *EMNLP*, pp. 1119–1130. The Association for Computational Linguistics, 2016. ISBN 978-1-945626-25-8. URL http://dblp.uni-trier.de/db/conf/emnlp/emnlp2016.html#BlodgettGO16.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In Sihem Amer-Yahia, Mohammad Mahdian, Ashish Goel, Geert-Jan Houben, Kristina Lerman, Julian J. McAuley, Ricardo Baeza-Yates, and Leila Zia (eds.), *WWW (Companion Volume)*, pp. 491–500. ACM, 2019. ISBN 978-1-4503-6675-5. URL http://dblp.uni-trier.de/db/conf/www/www2019c.html#BorkanDSTV19.

Peter Bubenik and Alex Elchesen. Universality of persistence diagrams and the bottleneck and wasserstein distances. *Comput. Geom. Theory Appl.*, 105(C), August 2022. ISSN 0925-7721. doi: 10.1016/j.comgeo.2022.101882. URL https://doi.org/10.1016/j.comgeo.2022.101882.

Mathieu Carriere, Frederic Chazal, Marc Glisse, Yuichi Ike, Hariprasad Kannan, and Yuhei Umeda. Optimizing persistent homology based functions. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 1294–1303. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/carriere21a.html.

Mathieu Carrière, Yuichi Ike, Théo Lacombe, and Naoki Nishikawa. Persistence-based topological optimization: a survey, 03 2026.

Mauricio Che, Fernando Galaz-García, Luis Guijarro, and Ingrid Amaranta Membrillo Solis. Metric geometry of spaces of persistence diagrams. *Journal of Applied and Computational Topology*, 8 (8):2197–2246, Dec 2024. ISSN 2367-1734. doi: 10.1007/s41468-024-00189-2. URL https://doi.org/10.1007/s41468-024-00189-2.

Chao Chen, Xiuyan Ni, Qinxun Bai, and Yusu Wang. A topological regularizer for classifiers via persistent homology. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89, pp. 2573–2582. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/chen19g.html.

Somnath Basu Roy Chowdhury, Nicholas Monath, Avinava Dubey, Amr Ahmed, and Snigdha Chaturvedi. Robust concept erasure via kernelized rate-distortion maximization. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Mattéo Clémot, Julie Digne, and Julien Tierny. Topological autoencoders++: Fast and accurate cycle-aware dimensionality reduction. *IEEE Transactions on Visualization and Computer Graphics*, 32(2):1622–1639, 2026. doi: 10.1109/TVCG.2025.3644671.

Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pp. 2292–2300, Red Hook, NY, USA, 2013. Curran Associates Inc.

Maria De-Arteaga, Alexey Romanov, Hanna M. Wallach, Jennifer T. Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Cem Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In danah boyd and Jamie H. Morgenstern (eds.), *FAT*, pp. 120–128. ACM, 2019. URL http://dblp.uni-trier.de/db/conf/fat/fat2019.html#De-ArteagaRWCBC19.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

H. Edelsbrunner and J. Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2009.

Herbert Edelsbrunner and Ernst P Mucke. Simulation of simplicity: a technique to cope with degenerate cases in geometric algorithms. *ACM Transactions on Graphics*, 9(1):66–104, 1990. doi: 10.1145/77635.77639.

Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological Persistence and Simplification. 28(4):511–533, 2002. doi: 10.1007/S00454-002-2885-2.

Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *EMNLP*, pp. 11–21. Association for Computational Linguistics, 2018. ISBN 978-1-948087-84-1. URL http://dblp.uni-trier.de/db/conf/emnlp/emnlp2018.html#ElazarG18.

Thomas Fel, Victor Boutin, Mazda Moayeri, Rémi Cadène, Louis Bethune, Léo Andéol, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023a. Curran Associates Inc.

Thomas Fel, Agustin Picard, Louis Béthune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2711–2721, June 2023b.

Amirata Ghorbani, James Wexler, James Zou, and Been Kim. *Towards automatic concept-based explanations*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Avani Gupta, Saurabh Saini, and P J Narayanan. Concept distillation: Leveraging human-centered explanations for model improvement. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 63724–63737. Curran Associates, Inc., 2023. doi: 10.52202/075280-2783. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/c9450295fd667740a39a68148fc17f6e-Paper-Conference.pdf.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, pp. 770–778. IEEE, June 2016. doi: 10.1109/CVPR.2016.90. URL http://ieeexplore.ieee.org/document/7780459.

Felix Hensel, Michael Moor, and Bastian Rieck. A survey of topological machine learning methods. *Frontiers in Artificial Intelligence*, Volume 4 - 2021, 2021. ISSN 2624-8212. doi: 10.3389/frai.2021.681108. URL https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2021.681108.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015. URL http://dblp.uni-trier.de/db/journals/corr/corr1503.html#HintonVD15.

Christoph Hofer, Roland Kwitt, Marc Niethammer, and Mandar Dixit. Connectivity-optimized representation learning via persistent homology. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2751–2760. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/hofer19a.html.

Christoph Hofer, Florian Graf, Marc Niethammer, and Roland Kwitt. Topologically densified distributions. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 4304–4313. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/hofer20a.html.

Floris Holstege, Bram Wouters, Noud Van Giersbergen, and Cees Diks. Removing spurious concepts from neural network representations via joint subspace estimation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of

Proceedings of Machine Learning Research, pp. 18568–18610. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/holstege24a.html.

Floris Holstege, Shauli Ravfogel, and Bram Wouters. Preserving task-relevant information under linear concept removal. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=t79rMQGk4S.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *CoRR*, abs/2001.08361, 2020. URL https://arxiv.org/pdf/2001.08361.pdf.

Michael Kerber, Dmitriy Morozov, and Arnur Nigmetov. Geometry helps to compare persistence diagrams. *ACM J. of Experimental Algorithmics*, 22:1–20, 2017. doi: 10.1145/3064175.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kim18d.html.

Jungeun Kim, Junwon You, Dongjin Lee, Ha Young Kim, and Jae-Hun Jung. Do topological characteristics help in knowledge distillation? In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24674–24693. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/kim24aj.html.

Mohamed Kissi, Mathieu Pont, Joshua A. Levine, and Julien Tierny. A practical solver for scalar data topological simplification. *IEEE Transactions on Visualization and Computer Graphics*, 31 (1):97–107, 2025. doi: 10.1109/TVCG.2024.3456345.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Michael Moor, Max Horn, Bastian Rieck, and Karsten M. Borgwardt. Topological autoencoders. In *ICML*, 2020.

James Munkres. Algorithms for the assignment and transportation problems. *J. of SIAM*, 5(1): 32–38, 1957. doi: https://doi.org/10.1137/0105003.

Burin Naowarat, Hao Tang, and Sharon Goldwater. A framework for analyzing concept representations in neural models. In Claire Bonial and Yevgeni Berzak (eds.), *Proceedings of the 30th Conference on Computational Natural Language Learning*, pp. 574–587, San Diego, California, USA, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-410-1. doi: 10.18653/v1/2026.conll-main.34. URL https://aclanthology.org/2026.conll-main.34/.

Arnur Nigmetov and Dmitriy Morozov. Topological optimization with big steps. *Discrete & Computational Geometry*, 1 2024. ISSN 0179-5376. doi: 10.1007/s00454-023-00613-x. URL https://www.osti.gov/biblio/2337698.

Mathieu Pont and Christoph Garth. Region-aware wasserstein distances of persistence diagrams and merge trees. *IEEE Transactions on Visualization and Computer Graphics*, 32(9):7756–7773, 2026. doi: 10.1109/TVCG.2026.3703539.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647/.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. Linear adversarial concept erasure. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18400–18421. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ravfogel22a.html.

Shauli Ravfogel, Yoav Goldberg, and Ryan Cotterell. Log-linear guardedness and its implications. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9413–9431, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.523. URL https://aclanthology.org/2023.acl-long.523/.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6550.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

Antoine Saillenfest and Pirmin Lemberger. Nonlinear concept erasure: A density matching approach. In Inês Lynce, Nello Murano, Mauro Vallati, Serena Villata, Federico Chesani, Michela Milano, Andrea Omicini, and Mehdi Dastani (eds.), *ECAI 2025 - 28th European Conference on Artificial Intelligence, 25-30 October 2025, Bologna, Italy - Including 14th Conference on Prestigious Applications of Intelligent Systems (PAIS 2025)*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pp. 2138–2145. IOS Press, 2025. doi: 10.3233/FAIA251053. URL https://doi.org/10.3233/FAIA251053.

Maxime Soler, Melanie Plainchault, Bruno Conche, and Julien Tierny. Lifted Wasserstein Matcher for Fast and Robust Topology Tracking . In *2018 IEEE 8th Symposium on Large Data Analysis and Visualization (LDAV)*, pp. 23–33, October 2018. doi: 10.1109/LDAV.2018.8739196. URL https://doi.ieeecomputersociety.org/10.1109/LDAV.2018.8739196.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *ICLR*. OpenReview.net, 2020. URL http://dblp.uni-trier.de/db/conf/iclr/iclr2020.html#TianKI20.

Fred Tung and Greg Mori. Similarity-Preserving Knowledge Distillation . In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1365–1374, Los Alamitos, CA, USA, November 2019. IEEE Computer Society. doi: 10.1109/ICCV.2019.00145. URL https://doi.ieeecomputersociety.org/10.1109/ICCV.2019.00145.

Johanna Vielhaben, Stefan Bluecher, and Nils Strodthoff. Multi-dimensional concept discovery (MCD): A unifying framework with completeness guarantees. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=KxBQPz7HKh.

Afra J. Zomorodian. Topology for computing. In Mikhail J. Atallah and Marina Blanton (eds.), *Algorithms and Theory of Computation Handbook (Second Edition)*, chapter 3, pp. 82–112. CRC Press, Boca Raton, 2010. doi: https://doi.org/10.1017/CBO9780511546945.

APPENDICES

These appendices provide additional details including extended related work, method details, mathematical proofs, detailed experimental setup and additional experiments and results to further illustrate and validate the effectiveness of TCR. The contents are organized as follows:

- Appendix A: Extended Related Work
- Additional Method Details
  - Appendix B: MSE Formulation
  - Appendix C: TCR Implementation Details and Computational Cost
- Mathematical Proofs
  - Appendix D and E: Existing Concept Erasure Criteria as Special Cases of TCR
  - Appendix F: Affine Reparameterization Invariance of TCR
  - Appendix G: Stability of TCR
- Appendix H: Detailed Experimental Setup
  - Appendix H.1: Setup for Concept Disentanglement Before Erasure
  - Appendix H.2: Setup for Concept Transfer
- Appendix I: Additional Experiments & Results
  - Appendix I.1: Concept Disentanglement Before Erasure
    - Appendix I.1.1: Beyond-linear Concept Leakage
    - Appendix I.1.2: Utility Preservation Across Erasure Levels and Data Regimes
    - Appendix I.1.3: Representation Similarity After Intervention
- Appendix J: Limitations

A EXTENDED RELATED WORK

Concept discovery. Concept-based interpretability aims to characterize learned representations through high-level semantic properties that are more meaningful to humans than individual activation dimensions (Kim et al., 2018; Ghorbani et al., 2019). A concept is often represented as a direction in the activation space of a selected layer, for instance obtained from a linear classifier trained to distinguish examples associated with the concept (Kim et al., 2018) or from multidimensional concept subspaces (Vielhaben et al., 2023). Concepts can also be discovered in an unsupervised manner, through clustering (Ghorbani et al., 2019), matrix factorizations such as NMF (Fel et al., 2023b) or sparse autoencoders (Huben et al., 2024). These approaches provide complementary mechanisms for extracting semantically meaningful properties encoded by neural representations.

Concept-based interventions. Beyond their use for interpretation, concepts can also serve as targets for post-hoc interventions on learned representations. A prominent line of work considers concept erasure, which aims to remove information associated with an undesired concept from a fixed representation through linear (Ravfogel et al., 2020; 2022; Belrose et al., 2023) or non-linear formulations (Chowdhury et al., 2023; Saillenfest & Lemberger, 2025). Other methods additionally seek to preserve information relevant to a downstream task (Holstege et al., 2024; 2025; Akbari et al., 2025). While effective for their respective objectives, these approaches are generally tailored to specific interventions, with new edits often requiring new objectives or optimization procedures. More fundamentally, they do not provide a common representation in which concept organization can be described and manipulated independently of the editing algorithm.

Topological abstractions. Topological Data Analysis (TDA) provides concise abstractions of the global organization of the features present within complex data (Edelsbrunner & Harer, 2009; Hensel et al., 2021). In particular, such methods track how topological features appear and disappear across a filtration of the input, thereby characterizing structure over multiple scales (Edelsbrunner et al., 2002). These features can be summarized compactly through persistence diagrams, while richer abstractions such as merge trees additionally encode how connected components progressively merge. Their ability to jointly represent individual features and their global organization makes such abstractions a natural candidate for describing structured relationships between concepts.

Persistence-based topological optimization. Beyond their descriptive role, topological abstractions can be incorporated into differentiable objectives to constrain learned representations. Such losses have been used, for instance, to regularize decision boundaries (Chen et al., 2019; Hofer et al., 2020) or preserve input topology during dimensionality reduction (Hofer et al., 2019; Moor et al., 2020; Clémot et al., 2026). Carriere et al. (2021) introduced a generic framework for optimizing persistence-based objectives, later extended to improve gradient density and scalability (Nigmetov & Morozov, 2024; Kissi et al., 2025; Carrière et al., 2026). TCR differs from these approaches in that it is not used to regularize the representation itself, but to encode its concept organization and to provide a common way to specify desired changes to the concepts of the underlying representation.

B Analytical MSE Formulation

This section details the closed-form MSE concept scores and predictive-overlap interactions used as an efficient analytical construction of TCR.

Concept scores. Assume that the representation and concept scores are centered, and let $\Sigma_Z = \mathrm{Cov}(Z)$ and $\mathbf{g}_i = \mathrm{Cov}(Z, \mathbf{c}_i)$. For a linear least-squares probe with coefficient vector $\boldsymbol{\beta}_i$, the population MSE is:

$$\mathbb{E}\Big[\big(c_i - \boldsymbol{\beta}_i^\top Z\big)^2\Big] = \mathrm{Var}(\mathbf{c}_i) - 2\boldsymbol{\beta}_i^\top \mathbf{g}_i + \boldsymbol{\beta}_i^\top \Sigma_Z \boldsymbol{\beta}_i. \tag{6}$$

Its minimum-norm solution is $\boldsymbol{\beta}_i^\star = \Sigma_Z^+ \mathbf{g}_i$, where $(\cdot)^+$ denotes the Moore–Penrose pseudoinverse. Substitution into Eq. 6 shows that the improvement over the optimal constant predictor is:

$$m_i = \mathbf{g}_i^\top \Sigma_Z^+ \mathbf{g}_i = A_{ii}, \qquad A_{ij} := \mathbf{g}_i^\top \Sigma_Z^+ \mathbf{g}_j. \tag{7}$$

When the score is normalized by the constant-predictor loss, it is instead $A_{ii}/\,\mathrm{Var}(\mathbf{c}_i)$, i.e., the explained-variance fraction. Thus, the diagonal of the matrix A contains all MSE concept scores.

Predictive overlap. As an efficient analytical alternative, we additionally consider *predictive overlap* as concept interactions. The off-diagonal entries of the same matrix quantify how strongly two concepts rely on aligned predictive directions. We define their normalized predictive overlap as:

$$\rho_{ij}^{\mathrm{overlap}} = \frac{A_{ij}^2}{A_{ii} A_{jj}}. \tag{8}$$

To interpret this quantity, define the whitened predictive direction $\mathbf{v}_i = (\Sigma_Z^{1/2})^+ \mathbf{g}_i$. Then $A_{ij} = \mathbf{v}_i^\top \mathbf{v}_j$ and $A_{ii} = \|\mathbf{v}_i\|_2^2$, so Eq. 8 becomes:

$$\rho_{ij}^{\mathrm{overlap}} = \frac{(\mathbf{v}_i^\top \mathbf{v}_j)^2}{\|\mathbf{v}_i\|_2^2 \|\mathbf{v}_j\|_2^2}. \tag{9}$$

It is therefore the squared cosine between the predictive directions of the two concepts and belongs to $[0, 1]$ whenever A_{ii} and A_{jj} are nonzero. A value of zero indicates orthogonal predictive directions, whereas a value of one indicates collinear directions, irrespective of their sign. Consequently, a single matrix A yields both the concept scores through its diagonal and every pairwise predictive overlap through its normalized off-diagonal entries.

C Implementation Details and Computational Cost

Differentiable construction. When optimizing a representation, we reconstruct TCR from the current transformed features at each optimization step. For BCE recoverability, the linear concept probes are therefore refitted differentiably, such that gradients propagate through the probe optimization. For interference interactions, removing a source concept is performed with differentiable LEACE and the probe of the target concept is subsequently refitted on the erased representation. Hence, interference measures the recoverability remaining after erasure, rather than the degradation of a fixed pre-erasure predictor. When differentiating a saddle $s_{ij} = \min(m_i, m_j)\rho_{ij}$, we stop gradients through the $\min(m_i, m_j)$ factor while retaining its value in the forward pass. This prevents interaction optimization from trivially decreasing a saddle by decreasing one of its incident concept scores instead of modifying their interaction.

Analytical least-squares construction. The MSE formulation of Appendix B allows all concept scores to be computed jointly. We first accumulate the sufficient statistics required to compute the representation covariance and concept cross-covariances, namely sums of features, feature outer products and feature–concept products. Given n_C concepts, we collect the resulting cross-covariances in $G = [\mathbf{g}_1, \ldots, \mathbf{g}_{n_C}]$ and perform a single eigendecomposition of the $d \times d$ representation covariance Σ_Z. The same decomposition yields all least-squares predictive directions and concept scores simultaneously. When predictive overlap is used, all pairwise interactions are subsequently obtained from their $n_C \times n_C$ predictive Gram matrix. The covariance decomposition is performed in double precision while numerically null eigendirections are discarded.

Computational cost. Computing the sufficient statistics requires $O(Nd^2 + Ndn_C)$ operations, followed by one $O(d^3)$ covariance decomposition. Computing only the concept scores then scales linearly with n_C up to matrix products, whereas forming all predictive-overlap interactions additionally costs $O(dn_C^2)$. Thus, although a complete concept mesh contains $O(n_C^2)$ pairwise interactions, the analytical construction does not require $O(n_C^2)$ matrix decompositions. By contrast, interference explicitly evaluates both directional effects for each concept pair and therefore requires $O(n_C^2)$ erased-recoverability evaluations. In our disentanglement experiments this construction involves only the task and concept nodes, while the larger concept sets used for transfer use the analytical least-squares construction. We further evaluate computation of TCR in Appendix I.1.

Online construction for representation transfer. When TCR is optimized jointly with a network, constructing it from the complete training set after every minibatch would be unnecessarily expensive. We instead maintain a bounded feature bank and update the statistics required by the analytical formulation as new features are produced. The resulting TCR therefore uses a larger sample of the current representation than the instantaneous minibatch while keeping its memory footprint bounded. The precise bank size and TCR instantiation used for concept transfer are reported in the corresponding experimental setup.

Choice of recoverability. BCE and MSE are different instantiations of concept recoverability rather than numerical approximations of one another. BCE is natural for the binary supervised concepts considered in our post-hoc erasure experiments and supports the intervention-based interference measure. MSE instead measures explained variance under linear least squares, naturally accommodates continuous concept coefficients and admits the shared analytical construction above. Likewise, interference and predictive overlap provide complementary interaction measures: the former measures the effect of explicitly erasing one concept on the recoverability of another, whereas the latter measures the alignment of their whitened predictive directions. We therefore use BCE with interference for concept disentanglement and normalized MSE for concept transfer.

D Concept Signatures

The persistence coordinates of a TCR feature describe the score of a concept and the interaction level at which it merges with an older component, but they do not encode additional concept-specific information. We therefore optionally associate each feature with a concept signature and extend the Wasserstein ground metric to jointly compare persistence coordinates and signatures. This formalism is distinct from the hard semantic matching used in the main method (Sec. 3.3) and is introduced primarily to express the objective-level specializations studied in Appendix E.

Related constructions have previously augmented persistence-diagram comparisons with information beyond birth and death. This has been done by incorporating spatial information associated with persistence pairs (Soler et al., 2018), properties of their associated regions (Pont & Garth, 2026) or, more generally, by allowing persistence-diagram points to live in general metric spaces (Bubenik & Elchesen, 2022; Che et al., 2024). Concept signatures follow a similar principle, but use auxiliary information describing the represented concepts rather than spatial or regional information.

D.1 Signature-Augmented Ground Metric

Recall that a TCR feature associated with concept C_i is represented by the persistence point $p_i = (s_i, m_i) \in \mathbb{R}^2$. We additionally associate it with a signature $\boldsymbol{\sigma}_i \in \mathcal{S}$, where $(\mathcal{S}, d_\sigma)$ is a metric space describing the information carried by the signatures. Given two signature-augmented persis-

tence features $(p_i, \boldsymbol{\sigma}_i)$ and $(p_j, \boldsymbol{\sigma}_j)$, we define their signature-augmented ground metric as:

$$\widetilde{d}_{q,\lambda_\sigma}\big((p_i, \boldsymbol{\sigma}_i), (p_j, \boldsymbol{\sigma}_j)\big) = \left(|s_i - s_j|^q + |m_i - m_j|^q + \lambda_\sigma^q d_\sigma(\boldsymbol{\sigma}_i, \boldsymbol{\sigma}_j)^q\right)^{1/q}, \tag{10}$$

where $q \geq 1$ and $\lambda_\sigma \geq 0$ controls the relative importance of the signature information. Setting $\lambda_\sigma = 0$ recovers the standard ground metric used in Sec. 2. When signatures are vectors in a common Euclidean space, a direct choice is $\mathcal{S} = \mathbb{R}^r$ with $d_\sigma(\boldsymbol{\sigma}_i, \boldsymbol{\sigma}_j) = \|\boldsymbol{\sigma}_i - \boldsymbol{\sigma}_j\|_q$. In this case, Eq. 10 is simply the ℓ^q distance between the augmented vectors $(s_i, m_i, \lambda_\sigma \boldsymbol{\sigma}_i)$ and $(s_j, m_j, \lambda_\sigma \boldsymbol{\sigma}_j)$.

Proposition 2. *Let $q \geq 1$, $\lambda_\sigma > 0$ and let d_σ be a metric on $\mathcal{S}$. Then $\widetilde{d}_{q,\lambda_\sigma}$ is a metric on $\mathbb{R}^2 \times \mathcal{S}$.*

Proof. Non-negativity, symmetry and identity of indiscernibles follow directly from the corresponding properties of the absolute value and d_σ. For the triangle inequality, let $u, v, w \in \mathbb{R}^2 \times \mathcal{S}$. The two persistence-coordinate differences and the signature distance each satisfy their corresponding triangle inequalities. Applying these inequalities component-wise, followed by Minkowski's inequality to the resulting three non-negative terms, gives $\widetilde{d}_{q,\lambda_\sigma}(u, w) \leq \widetilde{d}_{q,\lambda_\sigma}(u, v) + \widetilde{d}_{q,\lambda_\sigma}(v, w)$. Hence $\widetilde{d}_{q,\lambda_\sigma}$ is a metric. □

The proposition allows signatures to be more general than vectors, although vector-valued signatures are sufficient for the constructions considered here. The corresponding signature-aware Wasserstein distance is obtained by replacing the ground metric d_q by $\widetilde{d}_{q,\lambda_\sigma}$ in Eq. 1. Thus, signatures modify the cost of matching two off-diagonal features but leave the assignment formulation itself unchanged.

Matching to the diagonal. A signature is intended to characterize a persistence feature, not to prevent that feature from being removed or inserted through a diagonal match. We therefore let the diagonal projection of a signature-augmented point inherit its signature. For $p_i = (s_i, m_i)$, let $\Delta(p_i) = ((s_i + m_i)/2, (s_i + m_i)/2)$; its signature-augmented projection is $(\Delta(p_i), \boldsymbol{\sigma}_i)$. Consequently, matching a feature to its diagonal projection has no signature contribution and retains the usual persistence deletion cost:

$$\widetilde{d}_{q,\lambda_\sigma}\big((p_i, \boldsymbol{\sigma}_i), (\Delta(p_i), \boldsymbol{\sigma}_i)\big) = d_q\big(p_i, \Delta(p_i)\big). \tag{11}$$

As in the standard augmented-diagram formulation, the cost between two diagonal dummy points is set to zero. Hence signatures only affect correspondences between actual persistence features, while feature insertion and deletion retain their standard topological interpretation.

E Existing Concept Erasure Criteria as Special Cases of TCR

The editing formulation of TCR separates the criterion used to characterize a concept from the procedure used to modify the representation. In the main manuscript, this criterion is represented by the concept score $m_i = S(Z, C_i)$ and an intervention is specified by a target value $m_i^\star$, possibly together with targets on the interaction structure. Concept signatures from Appendix D further allow the target to constrain information associated with a concept beyond its persistence coordinates. Together, these ingredients allow several existing concept-erasure criteria to be recovered as particular TCR editing objectives. Our goal is not to reproduce their optimization algorithms, which may rely on closed-form projections, adversarial optimization, sequential subspace construction or eigendecompositions, but rather the objectives or constraints that they seek to enforce.

We group these connections according to the quantity used to characterize concept information: covariance and predictive recoverability for linear erasure, distributional and kernel dependence for nonlinear erasure and rate-distortion criteria. Most of these connections arise in a feature-wise specialization of TCR, in which interaction coordinates are fixed and only the concept maxima and, when needed, their signatures are edited. This is sufficient to recover the corresponding erasure criteria, while the saddle structure of TCR provides additional information that is not required by these special cases.

E.1 Objective-Level Specialization of TCR

Proposition 3 (Feature-wise specialization)**.** *Consider n semantically matched signature-augmented persistence features of the current and target TCR respectively given by $(p_i(Z), \boldsymbol{\sigma}_i(Z))$*

and $(p_i^\star, \boldsymbol{\sigma}_i^\star)$, with $p_i(Z) = (s_i, m_i(Z))$ and $p_i^\star = (s_i, m_i^\star)$. Assume that the interaction coordinate s_i is identical in both features and does not contribute to the intervention. Under the semantic matching introduced in Sec. 3.3 and the signature-augmented ground metric of Eq. 10, the TCR objective reduces to:

$$\mathcal{L}_{TCR}(Z) = \sum_{i=1}^{n} \left(|m_i(Z) - m_i^\star|^q + \lambda_\sigma^q d_\sigma(\boldsymbol{\sigma}_i(Z), \boldsymbol{\sigma}_i^\star)^q \right). \tag{12}$$

Proof. Since $(p_i(Z), \boldsymbol{\sigma}_i(Z))$ is constrained to match $(p_i^\star, \boldsymbol{\sigma}_i^\star)$, its contribution to the L^q-Wasserstein objective is the q-th power of their signature-augmented ground distance. The interaction coordinates being equal, the term $|s_i - s_i|^q$ vanishes and the contribution reduces to $|m_i(Z) - m_i^\star|^q + \lambda_\sigma^q d_\sigma(\boldsymbol{\sigma}_i(Z), \boldsymbol{\sigma}_i^\star)^q$. Summing over the matched features gives Eq. 12. □

Setting $\lambda_\sigma = 0$ recovers the extrema-only specialization, for which Eq. 12 reduces to $\sum_i |m_i(Z) - m_i^\star|^q$. Consequently, once a meaningful score and, when needed, a meaningful concept signature have been chosen, their target values directly specify the corresponding TCR editing objective. In the following, we instantiate this construction with quantities having independent statistical interpretations, rather than defining features from arbitrary pre-existing losses. The resulting connections therefore arise from expressing covariance, predictive information, distributional discrepancy, statistical dependence and rate-distortion within the same editing interface.

The specializations considered below are summarized compactly as follows. Each row specifies the TCR scores and target values associated with an existing erasure criterion:

$$\begin{cases}
m_c(P) = \|P\mathbf{g}_c\|_2,\ \boldsymbol{\sigma}_c(P) = P\mathbf{g}_c & m_c^\star = 0, \quad \boldsymbol{\sigma}_c^\star = \mathbf{0} & \equiv \text{LEACE} \\
m_a(P) = \|P\mathbf{g}_a\|_2,\ \boldsymbol{\sigma}_a(P) = P\mathbf{g}_a,\ a \in \{c, y\} & m_c^\star = 0, \quad \boldsymbol{\sigma}_c^\star = \mathbf{0}, & \equiv \text{SPLINCE} \\
 & m_y^\star = \|\mathbf{g}_y\|_2, \quad \boldsymbol{\sigma}_y^\star = \mathbf{g}_y & \\
m_c(P) = \mathcal{L}_c^0 - \min_\psi \mathcal{L}_{\text{BCE}}(c, q_\psi(PH)) & m_c^\star = 0 & \equiv \text{RLACE} \\
 & & \sim \text{INLP} \\
m_c(P) = \left(\sum_{a<b} \widehat{\text{MMD}}^2 \left(\mathbb{P}_{z|c=a}, \mathbb{P}_{z|c=b} \right) \right)^{1/2} & m_c^\star = 0 & \equiv \text{LEOPARD} \\
m_S(Z) = \text{HSIC}(Z, S),\ m_A(Z) = \text{HSIC}(Z, A) & m_S^\star = 0, & \approx \text{Obliviator} \\
 & m_A^\star > m_A(H),\ A \in \{X, X_i, Y\} & \\
m_c(Z) = U - R(Z \mid K),\ m_R(Z) = \lambda R(Z) & m_c^\star = 0, \quad m_R^\star = \lambda R(H) & \equiv \text{KRaM}
\end{cases}$$

Here, $\equiv$ denotes an exact correspondence at the stated criterion or objective level under the specified restrictions, $\approx$ denotes a correspondence in the optimized quantities and direction of intervention without equality of the complete objective, and $\sim$ denotes a structural or procedural correspondence. These relations do not imply equivalence of the complete optimization algorithms or of the transformations they select. The following subsections detail each case.

E.2 Linear and Predictive Erasure

Covariance-based erasure: LEACE and SPLINCE. LEACE (Belrose et al., 2023) characterizes perfect linear erasure through the absence of cross-covariance between the edited representation and the concept, while selecting among valid erasers by minimizing representation distortion. SPLINCE (Holstege et al., 2025) additionally requires the cross-covariance with a target task to remain unchanged. These constraints can be recovered from covariance-based concept features and signatures. Let $h \in \mathbb{R}^d$ be a centered random representation, let c and y denote centered scalar concept and task variables and define $\mathbf{g}_c = \text{Cov}(h, c)$ and $\mathbf{g}_y = \text{Cov}(h, y)$. For a linear edit $z = Ph$, their cross-covariances are:

$$\text{Cov}(z, c) = P\mathbf{g}_c, \qquad \text{Cov}(z, y) = P\mathbf{g}_y. \tag{13}$$

For this covariance-based specialization, we consider the TCR concept score $m_a(P) = \|P\mathbf{g}_a\|_2$ for both the concept $a = c$ and the task $a = y$. We associate each score with the covariance signature $\boldsymbol{\sigma}_a(P) = P\mathbf{g}_a$ and use the Euclidean signature distance.

Proposition 4 (LEACE and SPLINCE as covariance-based TCR specializations)**.** *Consider the feature-wise specialization of Prop. 3 with the covariance scores and signatures defined above and $\lambda_\sigma > 0$. For concept erasure, let the target concept feature satisfy $m_c^\star = 0$ and $\boldsymbol{\sigma}_c^\star = \mathbf{0}$. Then its TCR contribution is zero if and only if the LEACE linear-guardedness constraint $P\mathbf{g}_c = 0$ holds. If a task feature is additionally assigned the targets $m_y^\star = \|\mathbf{g}_y\|_2$ and $\boldsymbol{\sigma}_y^\star = \mathbf{g}_y$, the joint TCR contribution is zero if and only if:*

$$P\mathbf{g}_c = 0, \qquad P\mathbf{g}_y = \mathbf{g}_y, \tag{14}$$

which are the concept-erasure and task-covariance-preservation constraints of SPLINCE. The same result holds for vector-valued concepts or tasks by replacing the covariance vectors with the corresponding cross-covariance matrices and the Euclidean norm with the Frobenius norm.

Proof. For the concept feature, $m_c(P) = 0$ and $\boldsymbol{\sigma}_c(P) = \mathbf{0}$ are both equivalent to $P\mathbf{g}_c = \mathbf{0}$. Since every term in Eq. 12 is non-negative, its concept contribution is therefore zero if and only if the LEACE linear-erasure constraint holds.

For the task feature, the signature contribution vanishes if and only if $P\mathbf{g}_y = \mathbf{g}_y$. This equality also implies $m_y(P) = \|P\mathbf{g}_y\|_2 = \|\mathbf{g}_y\|_2 = m_y^\star$, so the persistence-coordinate contribution vanishes as well. Hence, the complete TCR contribution is zero if and only if both conditions in Eq. equation 14 hold. The matrix-valued case follows identically using the Frobenius norm. □

Preserving only the scalar covariance norm of the task is in general weaker than the SPLINCE requirement $P\mathbf{g}_y = \mathbf{g}_y$. The covariance signature makes this distinction explicit: preserving $\boldsymbol{\sigma}_y$ imposes the covariance-vector equality itself, while its associated covariance score is then preserved automatically. Thus, concept signatures allow the complete SPLINCE constraint to be expressed without replacing vector preservation by a weaker scalar constraint. LEACE and SPLINCE additionally select a particular transformation among those satisfying their respective constraints through a minimum-distortion criterion. The propositions above establish the erasure and preservation constraints represented by the TCR editing objective; reproducing the selected transformation also requires imposing the same distortion criterion and transformation family.

Predictive erasure: RLACE and INLP. RLACE (Ravfogel et al., 2022) formulates linear concept erasure as an adversarial game in which a linear predictor seeks to recover the concept while a rank-constrained linear transformation seeks to prevent this prediction. This criterion is directly related to the BCE recoverability score used by TCR. Let $c \in \{0, 1\}$ be a binary concept and, for a linear edit $z = Ph$, let:

$$m_c^{\mathrm{BCE}}(P) = \mathcal{L}_c^0 - \min_\psi \mathcal{L}_{\mathrm{BCE}}(c, q_\psi(Ph)), \tag{15}$$

where q_ψ is a linear logistic probe and $\mathcal{L}_c^0$ is the BCE of the optimal constant predictor, as in Eq. 2. Assuming that the probe family contains this constant predictor, $m_c^{\mathrm{BCE}}(P) \geq 0$.

Proposition 5 (RLACE as a predictive-recoverability TCR specialization)**.** *Let $\mathcal{A}$ be a common feasible family of linear edits for TCR and the adversarial erasure problem. Consider a feature-wise TCR specialization with the score of Eq. 15, target $m_c^\star = 0$ and no signature contribution. For any $q > 0$, minimizing the corresponding TCR contribution over $P \in \mathcal{A}$ is equivalent to maximizing the loss of the optimal linear logistic predictor:*

$$\arg\min_{P \in \mathcal{A}} \left| m_c^{\mathrm{BCE}}(P) \right|^q = \arg\max_{P \in \mathcal{A}} \min_\psi \mathcal{L}_{\mathrm{BCE}}(c, q_\psi(Ph)). \tag{16}$$

Consequently, choosing the same relaxed rank-constrained transformation family as RLACE recovers its concept-erasure criterion at the objective level.

Proof. Since $m_c^{\mathrm{BCE}}(P) \geq 0$ and $q > 0$, minimizing $|m_c^{\mathrm{BCE}}(P)|^q$ is equivalent to minimizing $m_c^{\mathrm{BCE}}(P)$. The null loss $\mathcal{L}_c^0$ does not depend on P, hence from Eq. 15 this is equivalent to maximizing $\min_\psi \mathcal{L}_{\mathrm{BCE}}(c, q_\psi(Ph))$, which gives Eq. 16. □

The connection also appears directly in the optimization procedure. At an TCR iteration, the probe is fitted on the current representation, approximately minimizing $\mathcal{L}_{\mathrm{BCE}}(c, q_\psi(Ph))$. The subsequent gradient update of P toward $m_c^\star = 0$ reduces the recoverability in Eq. 15, and therefore increases the loss attained by this fitted predictor. After updating the representation, rebuilding the TCR refits

the probe and provides a new best response to the current transformation. Thus, the repeated probe-fitting and representation-update steps can be viewed as the two sides of the same predictor–eraser adversarial game as RLACE. The two methods differ in how this game is optimized and in the constraints imposed on the transformation. This correspondence concerns a restricted specialization of TCR. RLACE restricts the intervention to a fixed-rank orthogonal projection and considers a single adversarial predictive-erasure objective. In contrast, TCR can use different concept scores and transformation families, jointly represent several concept and task features and additionally optimize their interactions through the target topology. The RLACE objective is therefore recovered when restricting TCR to a single BCE-based concept feature with a zero target, no interaction constraints and the same projection family; these restrictions are not inherent to the general TCR formulation.

INLP (Ravfogel et al., 2020) targets the same general notion of linear guardedness through a different sequential construction. At each iteration, it fits a linear concept classifier on the current representation, projects the representation onto the classifier nullspace and refits a new classifier on the projected representation. Each projection therefore exactly removes the direction used by the current classifier, while the repetition searches for additional predictive directions that remain after the previous ones have been removed. In contrast, BCE-based TCR uses the recoverability score of Eq. 15 as its editing criterion and replaces each hard nullspace projection by a gradient update that reduces the current concept recoverability before refitting the probe. If guardedness is evaluated with the same linear probe family and BCE criterion, both procedures target the condition $m_c^{\mathrm{BCE}} = 0$. INLP can therefore be viewed as a hard sequential counterpart of this predictive-erasure specialization of TCR, rather than as a distinct scalar objective reproduced by TCR. The original INLP formulation uses classification accuracy to operationalize guardedness, so this correspondence concerns the shared linear-predictive constraint rather than their optimization procedures.

E.3 Distributional and Nonlinear Dependence

The previous specializations characterize concept information through covariance or linear predictability. Nonlinear erasure methods instead consider stronger distributional or statistical dependencies between the edited representation and the concept. We next show that such criteria can similarly be expressed through TCR concept scores.

Distribution matching: LEOPARD. LEOPARD (Saillenfest & Lemberger, 2025) seeks a rank-constrained orthogonal projection for which the class-conditional distributions of a discrete concept become indistinguishable. Let $c \in \{1, \dots, K\}$ and $z = Ph$. We define the following distributional TCR concept score:

$$m_c^{\mathrm{MMD}}(P) = \left(\sum_{a<b} \widehat{\mathrm{MMD}}^2 \left(\mathbb{P}_{z|c=a}, \mathbb{P}_{z|c=b} \right) \right)^{1/2}. \tag{17}$$

For a binary concept, this is the empirical MMD between its two class-conditional distributions.

Proposition 6 (LEOPARD as a distribution-matching TCR specialization)**.** *Consider the feature-wise specialization of Prop. 3 with the score of Eq. 17, target $m_c^\star = 0$, no signature contribution and $q = 2$. Then the corresponding TCR contribution is exactly the LEOPARD erasure loss:*

$$\left| m_c^{\mathrm{MMD}}(P) \right|^2 = \sum_{a<b} \widehat{\mathrm{MMD}}^2 \left(\mathbb{P}_{z|c=a}, \mathbb{P}_{z|c=b} \right). \tag{18}$$

Restricting $P = UU^\top$ to the same rank-r parameterization and adding the LEOPARD projection penalty $\gamma \|U^\top U - I_r\|_F^2$ therefore recovers its complete training objective.

Proof. With $m_c^\star = 0$, no signature contribution and $q = 2$, Eq. 12 gives the contribution $|m_c^{\mathrm{MMD}}(P)|^2$. Substituting Eq. 17 yields Eq. 18. The remaining LEOPARD term depends only on the transformation parameterization and is recovered by imposing its orthogonality penalty. □

Thus, LEOPARD corresponds to a distributional choice of the TCR concept score: instead of measuring whether the concept is linearly predictable, it measures whether its class-conditional representation distributions can still be distinguished. With a characteristic kernel, a zero population MMD implies equality of these distributions. The adjustable projection rank and the orthogonalization used by LEOPARD concern how this criterion is optimized, not the criterion itself.

Kernel dependence: Obliviator. Obliviator (Akbari et al., 2025) extends the erasure criterion from linear predictability to nonlinear statistical dependence. Let S denote the unwanted concept, Y the task, X the original representation and X_i the representation provided as input to the encoder at iteration i. For its output Z_i, Obliviator minimizes an empirical RKHS objective that, up to the common HSIC normalization, can be written as:

$$\mathrm{HSIC}(Z_i, S) - \tau_x \,\mathrm{HSIC}(Z_i, X) - \tau_{x_i} \,\mathrm{HSIC}(Z_i, X_i) - \tau_y \,\mathrm{HSIC}(Z_i, Y), \tag{19}$$

where the task term is omitted in the unsupervised setting. The first term directly defines a nonlinear TCR concept score $m_S(Z) = \mathrm{HSIC}(Z, S)$. Targeting $m_S^\star = 0$ therefore recovers Obliviator's erasure criterion: with characteristic kernels, a vanishing population HSIC corresponds to statistical independence between the edited representation and the unwanted concept.

The remaining terms follow the same dependence measure but encourage information about X, X_i and Y to remain observable in the edited representation. They can consequently be represented in TCR as additional features with scores $m_A(Z) = \mathrm{HSIC}(Z, A)$, for $A \in \{X, X_i, Y\}$, whose target values specify the desired level of preservation. This recovers the same erasure–utility quantities used by Obliviator, although its encoder objective maximizes the utility HSIC terms directly rather than matching them to fixed target values.

The correspondence concerns the criterion encoded by these dependence quantities rather than the complete Obliviator procedure. Obliviator alternates encoder optimization with an additional RKHS disentanglement step obtained from a constrained eigenvalue problem, which exposes utility-related information while remaining insensitive to S before training the next encoder. In contrast, TCR expresses the desired decrease and preservation of these concept scores through its target representation and optimizes the chosen transformation accordingly.

E.4 Rate-Distortion Erasure

Kernelized rate-distortion: KRaM. KRaM (Chowdhury et al., 2023) formulates concept erasure through a kernelized rate-distortion criterion. Let $Z = T_\theta(H) \in \mathbb{R}^{N \times d}$ denote the edited representation and let $K \in \mathbb{R}^{N \times N}$ encode similarities between the concept labels of the samples. KRaM considers the standard rate-distortion $R(Z)$ and its kernelized counterpart:

$$R(Z \mid K) = \frac{1}{2} \log_2 \det \left(I + \frac{d}{N\epsilon^2} (ZZ^\top \odot K) \right). \tag{20}$$

Maximizing $R(Z \mid K)$ encourages samples that are similar according to the concept kernel to become dissimilar in the edited representation. Since this can simultaneously expand the overall representation space, KRaM additionally constrains its standard coding rate around the original value $b = R(H)$, leading to the relaxed objective:

$$\max_\theta \; R(Z \mid K) - \lambda \, |R(Z) - b| \,. \tag{21}$$

Both terms can be expressed through TCR features. Under the normalization assumed by KRaM, let U denote the constant upper bound of $R(Z \mid K)$. We define a concept feature and a rate-preservation feature with scores and targets:

$$m_c(Z) = U - R(Z \mid K), \qquad m_c^\star = 0, \qquad m_R(Z) = \lambda R(Z), \qquad m_R^\star = \lambda b. \tag{22}$$

The first feature measures the gap to maximal kernelized rate-distortion and therefore vanishes as the concept-erasure criterion of KRaM is optimized. The second preserves the coding rate of the original representation. With $q = 1$ and no signature contribution, their joint TCR contribution is:

$$\begin{aligned} \mathcal{L}_{TCR}(Z) &= |m_c(Z) - m_c^\star| + |m_R(Z) - m_R^\star| \\ &= U - R(Z \mid K) + \lambda \, |R(Z) - b| \,. \end{aligned} \tag{23}$$

Since U is independent of the transformation, minimizing Eq. 23 is equivalent to maximizing the KRaM objective of Eq. 21. KRaM can therefore be expressed as a two-feature rate-distortion specialization of TCR, with one feature driven toward concept erasure and the other preserving the global representation rate. The correspondence concerns the optimized criterion: KRaM additionally normalizes the transformed representations because the rate-distortion quantities are scale-sensitive and uses a nonlinear erasure function, whereas TCR is not restricted to this transformation or optimization procedure.

F AFFINE REPARAMETERIZATION INVARIANCE OF TCR

We show that the analytical MSE–overlap instantiation of TCR is invariant to any invertible affine change of coordinates of the representation. This includes, for example, permutations, rotations and nonsingular rescalings of the representation dimensions.

Proof of Prop. 1. Let Z_c and $\widetilde{Z}_c$ denote the centered representations. Centering removes the translation a, so $\widetilde{Z}_c = Z_c B^\top$. Since B is invertible, right multiplication by $B^\top$ does not change the column space and therefore $\mathrm{Col}(\widetilde{Z}_c) = \mathrm{Col}(Z_c)$. Let Π_Z denote the orthogonal projector onto this common column space. For concept C_i, let $\bar{\mathbf{c}}_i$ be its centered annotation vector. Since the annotations are centered, the optimal constant predictor is zero, with mean-squared error:

$$\mathrm{MSE}_{\mathrm{null}} = \frac{1}{N} \|\bar{\mathbf{c}}_i\|_2^2 .$$

The predictions of a linear probe belong to $\mathrm{Col}(Z_c)$. Hence, the fitted values of the optimal least-squares probe are given by the orthogonal projection $\widehat{\mathbf{c}}_i = \Pi_Z \bar{\mathbf{c}}_i$, and its mean-squared error is:

$$\mathrm{MSE}_{\mathrm{probe}} = \frac{1}{N} \|\bar{\mathbf{c}}_i - \Pi_Z \bar{\mathbf{c}}_i\|_2^2 .$$

By definition of an orthogonal projection, $\Pi_Z \bar{\mathbf{c}}_i$ and the residual $\bar{\mathbf{c}}_i - \Pi_Z \bar{\mathbf{c}}_i$ are orthogonal. Let $p := \Pi_Z \bar{\mathbf{c}}_i$ and $r := \bar{\mathbf{c}}_i - \Pi_Z \bar{\mathbf{c}}_i$, so that $\bar{\mathbf{c}}_i = p + r$. Since p and r are orthogonal, $p^\top r = 0$. Therefore, $\|\bar{\mathbf{c}}_i\|_2^2 = \|p + r\|_2^2 = (p + r)^\top (p + r) = p^\top p + 2p^\top r + r^\top r = \|p\|_2^2 + \|r\|_2^2$. Hence:

$$\|\bar{\mathbf{c}}_i\|_2^2 = \|\Pi_Z \bar{\mathbf{c}}_i\|_2^2 + \|\bar{\mathbf{c}}_i - \Pi_Z \bar{\mathbf{c}}_i\|_2^2 .$$

The MSE concept score is the gain of the best linear probe over the best constant predictor. Thus:

$$m_i = \mathrm{MSE}_{\mathrm{null}} - \mathrm{MSE}_{\mathrm{probe}} = \frac{1}{N} \|\Pi_Z \bar{\mathbf{c}}_i\|_2^2 .$$

Since Π_Z is the same for Z and $\widetilde{Z}$, m_i is identical for both representations Z and $\widetilde{Z}$ for every concept. Likewise, the predictive Gram matrix can be expressed using the same projected annotations. Using:

$$\Sigma_Z = \frac{1}{N} Z_c^\top Z_c, \qquad g_i = \frac{1}{N} Z_c^\top \bar{\mathbf{c}}_i,$$

and the original definition $A_{ij} = g_i^\top \Sigma_Z^+ g_j$, we obtain:

$$\begin{aligned} A_{ij} &= \left(\frac{1}{N} Z_c^\top \bar{\mathbf{c}}_i\right)^\top \left(\frac{1}{N} Z_c^\top Z_c\right)^+ \left(\frac{1}{N} Z_c^\top \bar{\mathbf{c}}_j\right) \\ &= \frac{1}{N} \bar{\mathbf{c}}_i^\top Z_c \left(Z_c^\top Z_c\right)^+ Z_c^\top \bar{\mathbf{c}}_j. \end{aligned}$$

The matrix $\Pi_Z = Z_c (Z_c^\top Z_c)^+ Z_c^\top$ is the orthogonal projector onto $\mathrm{Col}(Z_c)$. Hence:

$$A_{ij} = \frac{1}{N} \bar{\mathbf{c}}_i^\top \Pi_Z \bar{\mathbf{c}}_j.$$

We have $\bar{\mathbf{c}}_i^\top \Pi_Z \bar{\mathbf{c}}_j = \bar{\mathbf{c}}_i^\top \Pi_Z^\top \Pi_Z \bar{\mathbf{c}}_j = (\Pi_Z \bar{\mathbf{c}}_i)^\top (\Pi_Z \bar{\mathbf{c}}_j) = \langle \Pi_Z \bar{\mathbf{c}}_i, \Pi_Z \bar{\mathbf{c}}_j \rangle$ since Π_Z is an orthogonal projector making $\Pi_Z^\top = \Pi_Z$ and $\Pi_Z^2 = \Pi_Z$. Hence:

$$A_{ij} = \frac{1}{N} \langle \Pi_Z \bar{\mathbf{c}}_i, \Pi_Z \bar{\mathbf{c}}_j \rangle .$$

Thus every A_{ij} is unchanged. Consequently, whenever the overlap interaction is defined:

$$\rho_{ij}^{\text{overlap}} = \frac{A_{ij}^2}{A_{ii}A_{jj}},$$

it is also unchanged. Since the saddle value is $s_{ij} = \min(m_i, m_j)\rho_{ij}^{\text{overlap}}$, all concept and saddle values of the scalar field are identical for Z and $\widetilde{Z}$.

The two representations therefore induce the same scalar field on the same concept mesh and hence the same filtration, persistence pairs and persistence diagram. This proves $TCR(\widetilde{Z}) = TCR(Z)$. If the scores are normalized by $\mathrm{Var}(\mathbf{c}_i)$, the conclusion is unchanged because this quantity depends only on the concept annotations. □

Remark 1 (Scope of the result). The proposition concerns the MSE–overlap instantiation of TCR, for which the construction is determined by least-squares predictions. It does not require Z or its covariance matrix to have full rank. It establishes invariance to invertible affine changes of representation coordinates, not to arbitrary non-invertible or nonlinear transformations.

G Stability of TCR

We show that, away from degenerate cases, small perturbations of a representation induce small perturbations of its TCR. We consider the normalized MSE formulation with predictive-overlap interactions from Sec. 3.2. The result is local: this is sufficient to state that TCR varies continuously and at most linearly under sufficiently small representation perturbations.

G.1 Representation-to-TCR Stability

Let $H \in \mathbb{R}^{N\times d}$ be a centered representation of N samples, and let $\mathbf{c}_1, \ldots, \mathbf{c}_{n_C} \in \mathbb{R}^N$ be fixed, centered, non-constant concept annotation vectors. Set $n := N - 1$. For any centered representation $Z \in \mathbb{R}^{N\times d}$, define the empirical covariance and concept cross-covariances as:

$$\Sigma(Z) = \frac{1}{n}Z^\top Z, \qquad \mathbf{g}_i(Z) = \frac{1}{n}Z^\top \mathbf{c}_i.$$

Whenever $\Sigma(Z)$ is invertible, define the predictive Gram entries as:

$$A_{ij}(Z) = \mathbf{g}_i(Z)^\top \Sigma(Z)^{-1}\mathbf{g}_j(Z).$$

For each concept, define its variance as:

$$\sigma_i^2 = \frac{1}{n}\|\mathbf{c}_i\|_2^2 > 0.$$

The normalized MSE concept scores, predictive-overlap interactions and corresponding saddle values are respectively defined as:

$$m_i(Z) = \frac{A_{ii}(Z)}{\sigma_i^2}, \qquad \rho_{ij}(Z) = \frac{A_{ij}(Z)^2}{A_{ii}(Z)A_{jj}(Z)}, \qquad s_{ij}(Z) = \min(m_i(Z), m_j(Z))\rho_{ij}(Z).$$

We denote by f_Z the resulting scalar field on the fixed concept mesh $\mathcal{M}$, by $TCR(Z)$ its persistence diagram and by d_B the Bottleneck distance, being the $q \to \infty$ limit of the L^q-Wasserstein distance, measuring the largest cost of a matched pair instead of accumulating all matching costs.

Proof of Theorem 1. The proof follows from the fact that the scalar values defining the concept mesh are locally Lipschitz functions of the representation, followed by stability of persistence diagrams.

First, the covariance and cross-covariance maps defined above are polynomial and linear in Z, respectively, and are therefore smooth. Since $\Sigma(H)$ is positive definite, continuity implies that there exists $\varepsilon_1 > 0$ such that $\Sigma(Z)$ remains positive definite whenever $\|Z - H\|_F < \varepsilon_1$. The matrix inverse is smooth on the set of invertible matrices. Hence every $A_{ij}(Z)$ is smooth, and therefore locally Lipschitz, around H.

Because each σ_i^2 is a fixed positive constant, every maximum $m_i(Z)$ is also locally Lipschitz. Moreover, $m_i(H) > 0$ implies $A_{ii}(H) > 0$. Since there are finitely many concepts and every $A_{ii}(Z)$ is continuous, there exists $\varepsilon_2 > 0$ such that all $A_{ii}(Z)$ remain strictly positive whenever $\|Z - H\|_F < \varepsilon_2$. We may therefore choose $\varepsilon \leq \min(\varepsilon_1, \varepsilon_2)$, so that both the covariance inverse and every denominator appearing in ρ_{ij} are well-defined throughout the ball. Consequently, every $\rho_{ij}(Z)$ is locally Lipschitz as well.

Finally, the map $(a, b) \mapsto \min(a, b)$ is Lipschitz, so every $s_{ij}(Z)$ is locally Lipschitz. Since the concept mesh contains finitely many maxima and saddles, their local Lipschitz constants can be bounded by a single constant $C > 0$ on a possibly smaller ball of radius ε. Hence:

$$\|f_H - f_{H'}\|_\infty \leq C \|H - H'\|_F \qquad \text{whenever } \|H - H'\|_F < \varepsilon. \tag{24}$$

The two scalar fields are defined on the same concept mesh. Although TCR uses an upper-star filtration, applying the usual persistence stability theorem to the negated scalar fields (Edelsbrunner & Harer, 2009) gives:

$$d_B(TCR(H), TCR(H')) \leq \|f_H - f_{H'}\|_\infty.$$

Combining this inequality with Eq. 24 proves Theorem 1. □

Remark 2 (Scope of the result). The result is local and deliberately excludes degenerate representations. The radius ε is chosen small enough such that the covariance remains invertible and all concept recoverabilities remain strictly positive whenever $|H - H'|_F < \varepsilon$. Within this neighborhood, all quantities defining the TCR scalar field are smooth or Lipschitz functions of the representation. Handling rank changes requires controlling the Moore–Penrose pseudoinverse and is not needed for the local stability statement above. A more quantitative version could replace this neighborhood by explicit lower bounds on the smallest eigenvalue and on concept recoverability, together with an explicit upper bound on $|H - H'|_F$. Importantly, the proof does not assume that the persistence pairing remains fixed, since the persistence stability theorem directly controls such changes.

G.2 Network-to-TCR Stability

The representation-level result immediately yields a network-level statement whenever the selected representation is locally Lipschitz with respect to the network parameters. Let H_θ denote the matrix collecting the selected representations of the same N samples for a network with parameters θ.

Corollary 1 (Network-to-TCR stability)**.** *Assume the hypotheses of Theorem 1 hold at H_θ, with stability radius ε, and assume that the representation map is locally Lipschitz in the parameters: there exist $L_\theta > 0$ and $\eta > 0$ such that $\|H_\theta - H_{\theta'}\|_F \leq L_\theta \|\theta - \theta'\|$ whenever $\|\theta - \theta'\| < \eta$. Then there exists $C > 0$ such that, for every θ' satisfying:*

$$\|\theta - \theta'\| < \min\left(\eta, \frac{\varepsilon}{L_\theta}\right),$$

we have:

$$d_B(TCR(H_\theta), TCR(H_{\theta'})) \leq C L_\theta \|\theta - \theta'\|.$$

Proof. The parameter restriction ensures that $\|H_\theta - H_{\theta'}\|_F < \varepsilon$, so $H_{\theta'}$ lies inside the representation-level stability ball. Applying Theorem 1 and then the assumed local Lipschitz bound for the representation map proves the result. □

H Detailed Experimental Setup

H.1 Setup for Concept Disentanglement Before Erasure

Datasets and representations. We evaluate four binary classification datasets spanning image and text modalities. Tab. 3 summarizes their task and concept labels, split sizes, fixed representations, and dimensions before and after PCA for the erasure experiments. Tab. 4 summarizes their group size, task-concept agreement $P(y = c)$ and label imbalance $P(y = 0)$.

Waterbirds (Sagawa et al., 2020) is constructed by segmenting bird foregrounds from the CUB dataset and compositing them onto land or water scenes from Places. The task label distinguishes

landbirds from waterbirds, while the concept records the background type. In the training set, 95% of each bird type appears on its associated background; validation and test instead allocate each bird type equally across both backgrounds. The resulting shift makes background strongly predictive during training while retaining rare landbird-on-water and waterbird-on-land groups, which expose reliance on the spurious background cue during evaluation.

Bias in Bios SN (De-Arteaga et al., 2019) is obtained by retaining the surgeon and nurse examples from Bias in Bios, a corpus of 397,340 English online biographies collected from Common Crawl and labeled with one of 28 occupations and a binary gender inferred from gendered pronouns. We preserve the original train, validation, and test splits after filtering. The task is to predict which of the two professions is associated with a biography, while the concept is gender. This subset preserves the pronounced profession–gender imbalance in the source corpus and therefore tests whether gender information can be erased without removing profession-relevant content.

Dial Mention (Blodgett et al., 2016; Elazar & Goldberg, 2018) is built from Twitter messages in the DIAL corpus. Its task is to determine whether a tweet originally contained an `@mention`, thereby distinguishing conversational from non-conversational tweets; all mention tokens are removed before the tweet is encoded, so they cannot directly reveal the label. The concept distinguishes the African-American English (AAE) and Standard American English (SAE) dialect categories, which the original benchmark uses as proxies for non-Hispanic Black and non-Hispanic White authors, respectively. The source construction retains only dialect assignments whose estimated confidence exceeds 80%, selecting tweets for which the dialect proxy is comparatively unambiguous while preserving dialect-linked lexical variation in the encoded text.

Jigsaw (Borkan et al., 2019) contains public online comments from the Civil Comments platform, crowd-annotated for both toxicity and mentioned identities. We assign the positive task label when the aggregated toxicity score is greater than 0.5. For the concept, we retain comments for which at least one selected identity annotation is available and aggregate eight source attributes: female, male, transgender, other gender, heterosexual, homosexual/gay/lesbian, bisexual, and other sexual orientation. The concept is positive when any of these attributes has a nonzero annotation and negative otherwise; we use the compact name "Gender" in tables and results.

Table 3: Datasets and fixed representations used for concept disentanglement before erasure.

Dataset	**Task** y	**Concept** c	**Train**	**Val**	**Test**	**Representation**	**Original dim.**	**PCA dim.**
Waterbirds	Bird type	Background	4,795	1,199	5,794	ResNet-50 avgpool	2,048	128
Bias in Bios SN	Profession (surgeon/nurse)	Gender	21,145	3,256	8,136	BERT `[CLS]`	768	128
Dial Mention	Mention detection	Dialect	56,140	8,000	8,000	BERT `[CLS]`	768	128
Jigsaw	Toxicity classification	Gender	24,508	2,724	6,808	BERT `[CLS]`	768	128

Table 4: Group composition of each dataset split. The four group sizes are indexed by the task and concept labels (y, c), with their percentage of the split shown in parentheses. Agreement is $P(y = c)$ and label imbalance is reported as $P(y = 0)$.

Dataset	**Split**	**Group size** n_{yc}				**Agreement** $P(y=c)$	**Label imbalance** $P(y=0)$
		n_{00}	n_{01}	n_{10}	n_{11}		
Waterbirds	Train	3,498 (72.95%)	184 (3.84%)	56 (1.17%)	1,057 (22.04%)	94.99%	76.79%
	Val	467 (38.95%)	466 (38.87%)	133 (11.09%)	133 (11.09%)	50.04%	77.81%
	Test	2,255 (38.92%)	2,255 (38.92%)	642 (11.08%)	642 (11.08%)	50.00%	77.84%
Bias in Bios SN	Train	7,521 (35.57%)	1,308 (6.19%)	1,127 (5.33%)	11,189 (52.92%)	88.48%	41.75%
	Val	1,158 (35.57%)	202 (6.20%)	174 (5.34%)	1,722 (52.89%)	88.45%	41.77%
	Test	2,894 (35.57%)	504 (6.19%)	434 (5.33%)	4,304 (52.90%)	88.47%	41.76%
Dial Mention	Train	25,263 (45.00%)	2,807 (5.00%)	2,807 (5.00%)	25,263 (45.00%)	90.00%	50.00%
	Val	2,000 (25.00%)	2,000 (25.00%)	2,000 (25.00%)	2,000 (25.00%)	50.00%	50.00%
	Test	2,000 (25.00%)	2,000 (25.00%)	2,000 (25.00%)	2,000 (25.00%)	50.00%	50.00%
Jigsaw	Train	11,029 (45.00%)	1,225 (5.00%)	1,225 (5.00%)	11,029 (45.00%)	90.00%	50.00%
	Val	681 (25.00%)	681 (25.00%)	681 (25.00%)	681 (25.00%)	50.00%	50.00%
	Test	1,702 (25.00%)	1,702 (25.00%)	1,702 (25.00%)	1,702 (25.00%)	50.00%	50.00%

We extract one fixed representation per example: the average-pooling output of an ImageNet-pretrained ResNet-50 for Waterbirds, and the final-layer `[CLS]` representation of BERT-base for the text datasets. Before applying either TCR or an erasure baseline, we fit PCA on the training representations and use the same transformation for validation and test data. Each dataset is therefore represented in its own 128-dimensional PCA space. The feature extractors remain frozen throughout, so methods differ only in how they transform these representations post hoc.

Evaluation protocol. After each representation intervention, we independently fit linear logistic-regression probes for the task y and concept c on the transformed training representations. Inputs are standardized using training statistics and the ℓ_2 regularization parameter is selected from $C \in \{0.01, 0.1, 1, 10, 100\}$ by validation accuracy. The selected probes are kept fixed when evaluating the test set. Task utility is measured by accuracy and worst-group accuracy (WGA), where the four groups correspond to the joint task–concept labels (y, c) and WGA is the minimum task accuracy over these groups. Our primary concept-leakage metric is the adjusted balanced accuracy of the linear concept probe, defined as $\mathrm{Leak} = 2\,\mathrm{BalAcc} - 1$, where BalAcc averages the concept-class recalls. Thus, 0 corresponds to chance-level balanced prediction and 1 to perfect concept prediction, independently of concept-class imbalance. The leakage probe itself is unweighted; "balanced" refers to the evaluation metric. Unless stated otherwise, the strong-erasure setting requires $\mathrm{Leak} < 0.1$. Additional oracle-conditioned and non-linear leakage diagnostics are described in Appendix I.1.1.

Model selection. For methods with tunable erasure hyperparameters, candidates under a leakage constraint t are considered feasible only when both their training and validation adjusted balanced linear leakage are at most t. Among feasible candidates, we maximize the mean of validation task accuracy and validation WGA. Test metrics are used only after model selection.

TCR configuration. For both TCR-DIS and TCR-ERA, we use BCE recoverability with differentiable linear logistic probes, with $C = 0.01$, Newton updates and convergence-based stopping. The representation transformation is a two-layer ReLU MLP mapping the 128-dimensional PCA representation through a hidden dimension of 256 back to 128 dimensions. TCR-DIS additionally uses interference saddles between the concept and task nodes and is optimized for 200 iterations. Its learning rate is selected on validation after applying LEACE to the transformed representation, after which the same learned disentangling transformation is evaluated with all considered erasure methods. In contrast, TCR-ERA directly suppresses the concept score while preserving the task score and does not use a saddle interaction. Its learning rate, task-preservation weight and number of optimization iterations are selected jointly using the same leakage-constrained validation criterion used for model selection, considering both SGD and Adam candidates.

H.2 Setup for Concept Transfer

Dataset and architectures. We use the standard CIFAR-100 (Krizhevsky & Hinton, 2009) split, containing 50,000 training and 10,000 test images from 100 classes. For hyperparameter selection, we hold out a deterministic class-stratified 10% of the training set, resulting in 45,000 training and 5,000 validation images. After selecting the hyperparameters, we discard this validation split, retrain on all 50,000 training images, and report performance on the 10,000 test images. Following the protocol of CKD Tian et al. (2020), we use their pretrained ResNet32×4 teacher and consider ResNet8×4 and ShuffleNetV2 students. The teacher is frozen throughout distillation, while the students are trained from scratch. Their architectures are summarized in Tab. 5.

Distillation training protocol and losses. We follow the CIFAR-100 training protocol of CKD (Tian et al., 2020): all students are trained for 240 epochs with a batch size of 64 using SGD with momentum 0.9 and weight decay 5×10^{-4}. The initial learning rate is 0.05 for ResNet8×4 and

Table 5: Teacher and student architectures used for concept transfer.

Role	Architecture	Penultimate dim.	Parameters	Training
Teacher	ResNet32×4	256	7,433,860	CKD pretrained checkpoint, frozen
Student	ResNet8×4	256	1,233,540	From scratch
Student	ShuffleNetV2	1,024	1,355,528	From scratch

Table 6: Weights used in the concept transfer experiments. A dash indicates that no auxiliary loss is used.

Method	γ	α	β	**Auxiliary loss**
Student	1.0	0	–	–
KD	0.1	0.9	–	–
FitNet	1.0	0	100	Hint regression
SPKD	1.0	0	3000	Similarity preservation
TopKD	0.1	0.9	4.0	Topology distillation
TCR k_{10}	0.1	0.9	1.0	TCR matching
TCR k_{20}	0.1	0.9	0.8	TCR matching

Table 7: Validation-selected auxiliary-loss weights β for TCR.

	k_2	k_5	k_{10}	k_{20}
Full image	30.0	20.0	10.0	1.0
4 crops	5.0	2.0	0.2	0.9
9 crops	6.0	1.3	1.0	0.8
16 crops	0.9	1.0	2.0	2.0

0.01 for ShuffleNetV2 and is multiplied by 0.1 after epochs 150, 180 and 210. We use random cropping and horizontal flipping during training and the standard CIFAR-100 normalization. For logit distillation, the temperature is fixed to $T = 4$. The weights used by each method are summarized in Tab. 6, where the total objective is written as $\gamma\mathcal{L}_{\mathrm{CE}} + \alpha\mathcal{L}_{\mathrm{KD}} + \beta\mathcal{L}_{\mathrm{aux}}$. FitNet uses the CKD hint layer 2 with a learned convolutional regressor and a mean-squared hint loss, while SPKD matches pairwise similarities of the last convolutional feature maps.

For TCR, the auxiliary-loss weight is selected using the held-out CIFAR-100 validation split described above; the selected configuration is then retrained from scratch using the full training set before test evaluation. We apply the same procedure to TopKD (Kim et al., 2024). TopKD is not part of the CKD benchmark and its loss coefficients are tuned in the original work for a different teacher–student configuration. We therefore retain the same $\gamma = 0.1$ and $\alpha = 0.9$ CE/KD weighting as our KD baseline and select only the weight of its topology loss on validation data. This ensures that improvements are compared under the same underlying CE/KD objective rather than being confounded by different loss weighting. Final results are averaged over five random seeds.

Concept Distillation (Gupta et al., 2023) leverages teacher-defined concepts to (de)sensitize a student to selected concepts rather than transferring their organization between representations and is therefore not directly applicable to our setting.

TCR construction. For concept transfer, we use the analytical normalized-MSE formulation of Appendix B with least-squares probes and the continuous Dictionary Learning coefficients as concept targets. During training, the student TCR is estimated from a rolling bank of 2,048 augmented training representations, whose least-squares sufficient statistics are updated incrementally. The first epoch initializes this bank before TCR matching is applied. Concept identities and coefficients are obtained once from the deterministic frozen-teacher concept discovery, and the resulting teacher TCR remains a fixed reference throughout student training.

Concept discovery and recoverability. We discover concepts from the frozen teacher penultimate representation using Dictionary Learning with $K \in \{2, 5, 10, 20\}$ components, non-negative sparse codes and a fixed discovery seed. Each value of K defines an independently fitted dictionary: the $K = 2, 5, 10$ and 20 concept banks are therefore not nested subsets of a common dictionary. This distinction is important when evaluating generalization across concept banks in Sec. 4.2. The same frozen teacher-defined dictionaries are used when evaluating all distillation methods, so concept recoverability always refers to an identical target concept representation.

In addition to concepts discovered from full images, we consider concepts discovered at progressively finer spatial scales. Following the crop-based concept discovery strategy of CRAFT (Fel et al., 2023b), neighboring crops use a stride approximately equal to 80% of their width. We construct deterministic 2×2, 3×3 and 4×4 grids, corresponding respectively to 4, 9 and 16 crops per image. For the 32×32 CIFAR-100 images, the resulting crop sizes are 18×18, 12×12 and 9×9 pixels, with strides 14, 9 and 7 pixels, respectively. Each crop is resized to 32×32 before extracting its teacher representation. The dictionary is fitted on these crop representations and subsequently used to encode the full-image teacher representations, providing concept coefficients at increasingly local spatial scales. Together with the full-image setting, this yields four concept-discovery scales. The main transfer comparison in Tab. 2 uses the 3×3 (9-crop) concept discovery, while the other scales are used to evaluate the effect of concept locality and generalization across independently discovered concept banks.

We quantify concept recoverability (CR) by measuring how accurately the teacher-derived concept coefficients can be linearly decoded from a model representation on unseen images. For a dictionary containing K concepts, let $C \in \mathbb{R}^{N\times K}$ denote the corresponding teacher concept coefficients and $H \in \mathbb{R}^{N\times d}$ the representation whose recoverability we evaluate. Using the CIFAR-100 training set, we fit a linear least-squares decoder from H_{train} to C_{train}: $\widehat{C}_{\text{train}} = H_{\text{train}}W + b$, where W and b are estimated exclusively from the training samples. The same frozen decoder is then applied to the test representation, $\widehat{C}_{\text{test}} = H_{\text{test}}W + b$.

For a test image i and concept j, $\widehat{c}_{ij}$ therefore denotes the coefficient predicted by this linear decoder, while c_{ij} denotes the corresponding coefficient obtained by encoding the teacher representation with the fixed concept dictionary. For each concept j, we compute the test coefficient of determination:

$$R_j^2 = 1 - \frac{\sum_{i\in\text{test}}(c_{ij} - \widehat{c}_{ij})^2}{\sum_{i\in\text{test}}(c_{ij} - \overline{c}_{j,\text{train}})^2},$$

where $\overline{c}_{j,\text{train}}$ is the mean teacher coefficient of concept j over the training set and therefore defines a constant predictor that uses no test information. An R_j^2 of 1 indicates perfect linear recovery of the teacher coefficient, while 0 corresponds to the performance of this training-mean predictor; negative values indicate worse generalization than the constant predictor. We finally define CR k_K as the mean test R^2 over the K concepts and report it as a percentage. Thus, CR measures how much of the teacher-defined concept representation remains linearly accessible from the evaluated representation on unseen data: a high CR indicates that a linear decoder fitted to predict the teacher's concept coefficients on the training set continues to recover those same concepts accurately on previously unseen images, independently of the model's task accuracy.

I Additional Experiments & Results

I.1 Concept Disentanglement Before Erasure

Runtime. We additionally report the execution time of each intervention in Tab. 8. As expected, using TCR-DIS as a preprocessing step introduces an additional optimization cost before applying the corresponding erasure method. For lightweight methods such as LEACE and SPLINCE, the resulting runtime nevertheless remains below 90 seconds on all datasets, while combining TCR-DIS with more computationally intensive methods such as RLACE can be substantially more expensive, reaching 1001.34 seconds on DIAL-Mention. Direct TCR-ERA avoids this additional erasure stage and requires between 4.05 and 142.63 seconds across the four datasets. In particular, it is faster

Table 8: Execution time (seconds) for standalone erasure methods, TCR-DIS followed by erasure, and direct TCR-ERA.

Method	Waterbirds (s)	Bias in Bios SN (s)	Dial Mention (s)	Jigsaw (s)
INLP	1.35±0.41	5.57±1.01	28.05±1.90	3.07±0.65
LEACE	0.03±0.06	**0.03±0.06**	**0.04±0.07**	**0.02±0.05**
RLACE	38.91±1.13	65.28±2.56	208.60±3.70	96.77±6.67
SPLINCE	**0.03±0.02**	0.15±0.20	0.30±0.33	0.10±0.03
LEOPARD	10.65±0.61	47.28±0.36	132.32±0.56	53.18±0.21
KRaM	5.39±1.30	36.58±4.44	59.58±28.59	45.26±5.67
Obliviator	1.67±0.14	93.14±0.29	8.62±0.27	11.95±0.32
TCR-DIS + INLP	48.18±4.60	42.36±1.95	277.02±34.47	35.52±1.55
TCR-DIS + LEACE	47.33±4.36	41.06±2.05	88.46±16.05	32.74±1.54
TCR-DIS + RLACE	84.51±5.65	188.65±7.93	1001.34±679.06	113.47±2.43
TCR-DIS + SPLINCE	47.47±4.56	41.12±2.05	88.82±16.22	32.81±1.53
TCR-DIS + LEOPARD	57.39±4.37	88.70±2.02	226.28±20.29	86.53±1.52
TCR-DIS + KRaM	51.92±4.24	56.00±1.99	127.32±16.13	52.25±4.41
TCR-DIS + Obliviator	49.01±4.49	53.01±1.95	118.78±15.91	54.06±1.43
TCR-ERA	4.05±0.47	10.42±1.14	142.63±1.55	8.24±0.48

than RLACE on every dataset and faster than LEOPARD on three out of four datasets, showing that direct erasure with TCR can remain computationally competitive with iterative erasure baselines.

Test accuracy. In addition of the test WGA shown in Tab. 1, test accuracy is shown in Tab. 9. The test task accuracy provides a complementary view to WGA. On Waterbirds and Bias in Bios, the gains from TCR-DIS also translate into substantially higher overall accuracy, whereas on DIAL-Mention and Jigsaw the accuracy remains comparable or can slightly decrease despite consistent improvements in WGA. This difference highlights that aggregate accuracy can mask group-specific degradation, while WGA directly measures utility on the most affected group. Notably, TCR-ERA achieves higher task accuracy than every standalone erasure baseline on all four datasets and obtains the highest accuracy among all erasure methods on Waterbirds, DIAL-Mention and Jigsaw.

I.1.1 Beyond-linear Concept Leakage

The experiments of Sec. 4.1 use an adjusted balanced linear leakage threshold measured through a linear probe trained with a convex loss. This follows the linear-guardedness perspective of prior concept-erasure work, for instance, LEACE guarantees that, after erasure, no linear predictor can improve over the optimal constant predictor under convex losses (Belrose et al., 2023). This criterion is therefore broader than a particular choice of logistic regression, but does not imply that the edited representation is statistically independent of the concept or guarded against arbitrary non-linear predictors. Indeed, direct linear-guardedness need not imply that a downstream model cannot infer the concept indirectly through other preserved information (Ravfogel et al., 2023). This distinction is particularly relevant when task and concept labels are correlated, since preserving task-predictive information may necessarily retain some predictability of the concept.

To assess concept leakage beyond the criterion used for model selection, we therefore report two stronger diagnostics: *oracle-conditioned leakage*, which fits separate balanced linear concept probes within each task class, and *non-linear leakage*, measured with a non-linear concept probe.

For the *oracle-conditioned leakage*, we fit two separate linear concept probes, one on samples with $y = 0$ and one on samples with $y = 1$. Each probe is an ℓ_2-regularized logistic regression trained with balanced concept-class weights within the corresponding task train set; its regularization strength is selected on the matching validation set by balanced accuracy. At evaluation time, each sample is routed to the probe corresponding to its ground-truth task label y, and leakage is reported as adjusted balanced accuracy, $2\,\mathrm{BalAcc} - 1$. This asks whether c remains linearly predictable once y is held fixed, rather than allowing predictability of y itself to explain concept leakage.

For *non-linear leakage*, we fit an unweighted ReLU MLP concept probe on the edited representation. Inputs are standardized using training statistics, and the probe is optimized with cross-entropy and Adam. We select the architecture among one, two, or three hidden layers of width 128, together with the learning rate and training epoch, according to validation accuracy, the learning rate is searched around 10^{-3} on the $\{1, 2, 5\} \times 10^k$ lattice. We train for at most 200 epochs with batch size 256. The selected probe is then evaluated on the test set using the same adjusted balanced leakage metric.

The oracle-conditioned results (Tab. 10) show that low global linear leakage does not generally imply low leakage after conditioning on the task. Nevertheless, TCR-DIS often reduces this stronger conditional leakage relative to the corresponding erasure baseline, with particularly large reductions on Jigsaw and consistent improvements on Waterbirds and Dial Mention. Bias in Bios remains difficult, most methods retain high conditional leakage despite satisfying the linear leakage constraint.

Non-linear leakage (Tab. 11) remains substantial for many methods and is often not reduced by TCR-DIS. This confirms that our main results should be interpreted as improvements in task preservation at matched *linear* concept leakage, rather than as guarantees of complete concept removal. KRaM achieve lower non-linear leakage on several datasets, but at a different utility-erasure tradeoff, indeed it often has lower wga Tab. 1 and test accuracy Tab. 9 than the other methods.

I.1.2 Utility Preservation Across Erasure Levels and Data Regimes

Fig. 4 complements the analyses of Fig. 2 with the corresponding experiments on other dataset and data-regime combinations. Panels (a), (c) and (f) reproduce the main-text results for direct comparison; we discuss the complementary results below.

Table 9: Test task accuracy with an adjusted balanced linear concept leakage < 0.1.

Method	Waterbirds	Bias in Bios SN	Dial Mention	Jigsaw
Raw	*88.95*	*96.76*	*64.91*	*71.56*
INLP	85.17±0.00	71.66±0.00	63.49±0.00	65.82±0.00
LEACE	85.67±0.00	76.55±0.00	65.06±0.00	67.58±0.00
RLACE	85.64±0.25	76.62±0.33	64.78±0.21	67.30±0.17
SPLINCE	85.66±0.00	76.50±0.00	65.06±0.00	67.51±0.00
LEOPARD	85.66±0.00	76.55±0.00	64.85±0.02	67.58±0.00
KRaM	78.44±0.77	68.23±0.93	59.00±0.40	62.20±0.79
Obliviator	84.70±0.46	75.08±0.63	62.87±0.61	64.39±0.63
TCR-DIS + INLP	88.57±1.32	87.18±2.07	62.72±0.81	59.31±1.92
TCR-DIS + LEACE	88.47±0.55	**94.68±0.21**	64.99±0.28	64.98±0.85
TCR-DIS + RLACE	88.98±0.56	94.59±0.29	63.88±0.61	65.19±0.69
TCR-DIS + SPLINCE	88.64±0.98	94.60±0.27	64.99±0.31	64.89±0.84
TCR-DIS + LEOPARD	88.94±0.53	94.60±0.31	64.98±0.36	65.01±0.79
TCR-DIS + KRaM	85.14±1.82	74.41±5.91	61.19±0.85	62.42±1.21
TCR-DIS + Obliviator	88.13±1.50	89.06±1.39	64.90±0.56	64.78±0.66
TCR-ERA	**89.69±0.38**	92.69±0.30	**66.44±0.31**	**67.66±0.35**

Table 10: Adjusted balanced oracle-conditioned test concept leakage.

Method	Waterbirds	Bias in Bios SN	Dial Mention	Jigsaw
Raw	*85.71*	*98.62*	*61.78*	*51.88*
INLP	60.58±0.00	77.59±0.00	34.47±0.00	29.99±0.00
LEACE	72.35±0.00	95.09±0.00	38.13±0.00	33.37±0.00
RLACE	72.37±0.26	94.99±0.10	38.30±0.20	33.55±0.49
SPLINCE	72.14±0.00	94.56±0.00	38.28±0.00	33.31±0.00
LEOPARD	72.59±0.00	94.99±0.00	35.84±0.26	33.34±0.00
KRaM	**44.58±2.70**	**58.73±2.74**	23.60±1.27	22.85±1.56
Obliviator	63.26±0.96	86.43±1.78	30.83±1.10	24.29±1.54
TCR-DIS + INLP	56.41±1.84	89.23±1.32	28.31±1.58	**5.49±2.74**
TCR-DIS + LEACE	60.76±3.15	95.77±0.65	30.44±1.17	10.24±1.25
TCR-DIS + RLACE	59.56±1.39	95.79±0.67	28.41±1.80	10.34±1.37
TCR-DIS + SPLINCE	61.01±1.81	95.97±0.47	30.18±1.41	9.96±1.23
TCR-DIS + LEOPARD	59.03±1.36	95.74±0.65	29.40±1.24	10.41±1.37
TCR-DIS + KRaM	53.00±3.82	65.52±5.20	**21.98±1.76**	11.35±2.34
TCR-DIS + Obliviator	59.62±1.26	96.53±0.37	28.85±1.23	12.87±1.78
TCR-ERA	51.93±1.78	91.82±0.89	34.10±1.65	24.82±1.20

Table 11: Adjusted balanced non-linear test concept leakage.

Method	Waterbirds	Bias in Bios SN	Dial Mention	Jigsaw
Raw	*79.83*	*98.89*	*55.99*	*42.77*
INLP	71.01±1.60	93.64±0.21	52.99±0.65	31.37±0.95
LEACE	74.60±0.92	98.05±0.21	55.14±0.48	35.95±0.91
RLACE	74.54±1.79	98.01±0.27	54.96±0.58	36.44±1.11
SPLINCE	75.14±1.42	97.45±0.26	56.18±0.45	36.91±1.03
LEOPARD	74.07±1.06	98.02±0.18	53.45±0.47	35.44±1.24
KRaM	**40.71±4.02**	76.64±2.24	**36.37±1.30**	**11.37±2.96**
Obliviator	68.36±0.96	93.38±0.52	50.00±1.29	21.47±1.35
TCR-DIS + INLP	76.84±1.18	98.30±0.14	55.07±0.74	32.62±1.67
TCR-DIS + LEACE	78.63±1.86	98.59±0.12	56.49±0.61	37.83±1.90
TCR-DIS + RLACE	78.35±1.46	98.55±0.16	55.64±0.73	38.20±1.11
TCR-DIS + SPLINCE	74.28±3.48	98.60±0.17	57.29±0.63	37.45±1.53
TCR-DIS + LEOPARD	78.64±1.46	98.62±0.16	56.15±0.80	38.15±1.23
TCR-DIS + KRaM	63.46±2.05	**76.44±3.92**	37.14±1.22	15.71±2.20
TCR-DIS + Obliviator	75.02±2.08	98.48±0.24	55.06±0.53	33.59±1.66
TCR-ERA	65.75±2.14	97.75±0.22	53.02±0.78	28.73±1.24

On Bias in Bios (Fig. 4b), TCR exhibits a particularly favorable trade-off under strong erasure, reaching 86.52 WGA at only 1.88% adjusted balanced linear leakage, compared with 59.53 for the best baseline under the strong-erasure constraint. Its WGA then remains nearly constant as the constraint is relaxed, reaching 84.31 at 60.52% leakage. In contrast, several baselines recover task utility primarily by retaining substantially more concept information; for instance, INLP approaches TCR only at approximately 46% leakage (84.52 versus 85.50 WGA).

The Jigsaw label-imbalance experiment (Fig. 4d) shows a similar robustness advantage. TCR achieves the highest WGA at every imbalance level, with its margin over the best baseline increasing from 1.73 points at $P(y{=}0) = 0.50$ to 18.04 at 0.55 and 26.29 at 0.65. Even at $P(y{=}0) = 0.75$, TCR retains 40.42 WGA compared with 22.03 for the strongest baseline, indicating substantially better utility preservation as the task-label distribution becomes increasingly imbalanced.

Finally, on DIAL-Mention (Fig. 4e), TCR is essentially tied with the strongest baseline when task and concept are independent at $P(y{=}c) = 0.50$ (68.18 versus 68.17 WGA), but obtains the highest WGA at every stronger spurious correlation. Its advantage ranges from 2.92 to 4.23 points for $P(y{=}c) \in [0.6, 0.9]$, complementing the analogous trend observed on Jigsaw in the main text.

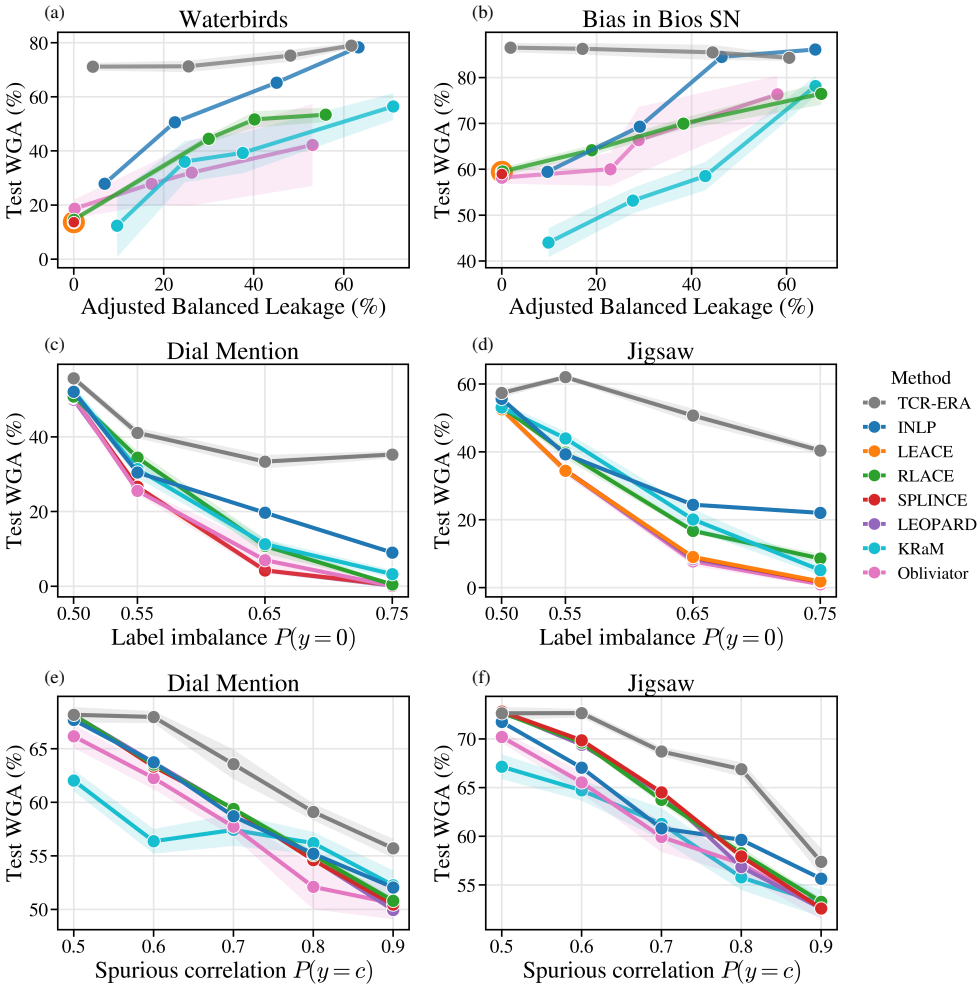


Figure 4: Extended erasure-utility and data-regime analysis. Erasure-utility trade-off, showing test WGA against adjusted balanced linear leakage, on Waterbirds (a) and Bias in Bios (b); impact of label imbalance on DIAL-Mention (c) and Jigsaw (d); and of spurious correlation on DIAL-Mention (e) and Jigsaw (f). Panels (a), (c) and (f) correspond to the main-text results for direct comparison.

Together, these additional experiments support the same behavior observed in the main analysis: TCR remains consistently competitive with the strongest baselines across the explored settings, while its utility-preservation advantage becomes most pronounced when strong erasure, label imbalance or spurious dependence make task and concept information harder to separate.

I.1.3 Representation Similarity After Intervention

Beyond task utility and concept leakage, we evaluate how strongly each intervention modifies the representation globally. This is important because improved task preservation after erasure could otherwise result from extensively reorganizing the representation around the specific task–concept pair. We quantify this effect using linear centered kernel alignment (CKA) (Kornblith et al., 2019), a standard measure for comparing neural representations that assesses their similarity.

For a test representation $H \in \mathbb{R}^{N\times d}$ before intervention and the corresponding representation $Z \in \mathbb{R}^{N\times d'}$ after intervention, let H' and Z' denote the matrices obtained by centering each feature over the same test samples. We compute linear CKA as:

$$\mathrm{CKA}(H, Z) = \frac{\|Z'^{\top} H'\|_F^2}{\|H'^{\top} H'\|_F \, \|Z'^{\top} Z'\|_F}.$$

Linear CKA is invariant to isotropic scaling and orthogonal transformations of either representation, and therefore does not penalize such changes of coordinates, while remaining sensitive to more general changes in representational geometry (Kornblith et al., 2019). Values closer to 1 indicate greater global similarity between the original and edited representations. Unlike a learned reconstruction or alignment score, CKA requires no fitted mapping; we therefore compute it directly between the original and final test representations. For composed TCR-DIS methods, Z denotes the representation

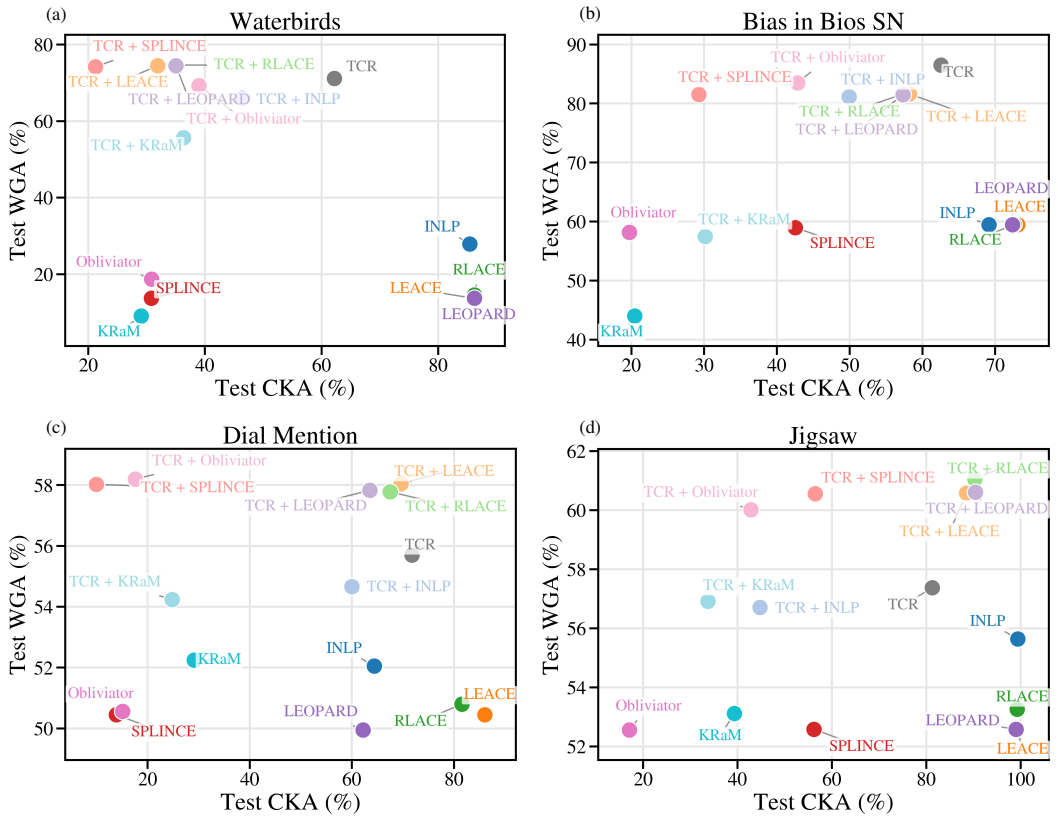


Figure 5: Task utility against representational similarity after concept erasure, showing test WGA against linear CKA between the original and intervened test representations on Waterbirds (a), Bias in Bios (b), DIAL-Mention (c) and Jigsaw (d). Higher CKA indicates greater similarity.

after both disentanglement and the subsequent erasure method. We compute CKA independently for each run and report its mean across seeds as a percentage.

Fig. 5 reports test WGA against the linear CKA between the original and intervened representations. The results first show that preserving global representation geometry and preserving task utility are distinct objectives: high CKA does not necessarily translate into high WGA. On Waterbirds, INLP, LEACE, RLACE and LEOPARD retain between 85.40% and 86.22% CKA while reaching only 13.71–27.88 WGA. Similarly, on Jigsaw, these methods retain approximately 99% CKA but obtain only 52.59–55.64 WGA. Thus, concept erasure can leave most of the global representational structure unchanged while substantially degrading the task-relevant information required by the worst-performing groups.

Conversely, TCR-DIS often induces a larger representational reorganization while preserving substantially more task utility. This is particularly pronounced on Waterbirds, where TCR-DIS variants reach 66.01–74.66 WGA despite lower CKA, and on Bias in Bios, where the variants based on INLP, LEACE, RLACE, SPLINCE and LEOPARD reach approximately 81–82 WGA. The same pattern is milder on Jigsaw: combining TCR-DIS with LEACE, RLACE or LEOPARD reduces CKA from approximately 99% to 89–90%, while increasing WGA from 52.59–53.27 to 60.58–61.05. Hence, the utility gains of disentanglement do not arise from keeping the edited representation as close as possible to its original geometry; modifying this geometry can instead facilitate the separation of task-relevant information from the concept being erased.

The amount of representational change also depends strongly on the form of the intervention. Across datasets, the non-linear baselines KRaM and Obliviator tend to exhibit among the lowest CKA values, whereas the linear projection-based INLP, LEACE and RLACE generally remain substantially closer to the original representation. On Waterbirds and Jigsaw, for instance, the latter retain approximately 85–99% CKA despite substantially lower WGA. At the same time, TCR does not systematically decrease CKA: TCR-DIS increases both CKA and WGA over Obliviator on Bias in Bios (19.71% to 42.83% CKA and 58.18 to 83.50 WGA) and Jigsaw (17.06% to 42.79% CKA and 52.56 to 60.01 WGA), with similar behavior for KRaM on Waterbirds. Direct TCR-ERA also retains substantial similarity to the original representation, with CKA ranging from 62.18% on Waterbirds to 81.22% on Jigsaw, while achieving strong WGA across the four datasets. Overall, the degree of geometric reorganization appears to depend on both the intervention and the subsequent erasure operator rather than being systematically larger for TCR.

Finally, CKA measures global representational similarity rather than preservation of all information in the original representation. A lower CKA therefore indicates a change in representation geometry, but does not by itself imply that unrelated information has been discarded. We consequently interpret these results as showing that improvements from TCR can involve meaningful representational reorganization, while task utility can remain preserved despite lower global similarity.

J LIMITATIONS

TCR depends on the concepts used to construct it and on the probes and interaction criterion used to characterize them. Different choices may therefore induce different concept organizations, and concepts absent from the chosen set are not represented. In particular, our interference criterion is operational: it measures how erasing one concept affects the recoverability of another and should not be interpreted as establishing a causal dependence between them. The persistence reduction is intentionally lossy. It retains the multiscale connectivity induced by pairwise interactions while discarding variations of the concept mesh that do not alter this hierarchy. We do not claim that topology is the unique useful abstraction of concept organization; alternative reductions or the full concept mesh may preserve complementary information. Moreover, our current construction relies on pairwise interactions and 0-dimensional topology and therefore does not explicitly represent higher-order relationships between concepts. Although the persistence diagram contains only $n_C - 1$ hierarchy-defining saddles, constructing the complete concept mesh requires quadratically many pairwise interaction evaluations, which may become costly for very large concept sets. Finally, our experiments focus primarily on penultimate representations and mostly linear concept probes. Nonlinear concept information may remain after linear interventions, and the estimated concept organization may also change under distribution shift. Extending TCR to larger concept collections, multiple network layers and richer interaction models provides directions for future work.