IsoAct: Structure-Preserving Post-hoc Debiasing via Isometric Actions

Anonymous Authors[1]

Abstract

Learning representations that are robust to nuisance factors is essential for reliable deployment, but remains challenging when the representations lie on non-Euclidean manifolds. Existing debiasing methods typically assume Euclidean representations or apply linear interventions, which can distort the geometry of structured latent spaces. This creates a mismatch between the Euclidean assumptions of the editing operation and the geometric constraints that manifold-valued representations must satisfy. We propose IsoAct, a post-hoc representation-editing framework that reduces nuisance information by composing manifold-native action primitives while maintaining downstream utility and manifold validity. The framework unifies debiasing on hyperspherical, hyperbolic, and SE(3) representations by parameterizing valid transformations within each geometry. We evaluate the approach across synthetic and real-world datasets using debiasing capability, task preservation, and manifold-constraint violation. Across these settings, IsoAct yields a favorable nuisance–utility–validity trade-off relative to Euclidean editing baselines, reducing nuisance recoverability while maintaining competitive task preservation and low manifold-constraint violation. [1]

1. Introduction

Nuisance-robust representations are important for reliable deployment, fairness, and domain robustness. Post-hoc debiasing and concept-erasure methods address this problem by editing pretrained representations after feature extraction, avoiding the cost of retraining the upstream model (Bolukbasi et al., 2016; Ravfogel et al., 2020; 2022; Belrose et al., 2023). However, this setting becomes more challenging when representations lie on non-Euclidean manifolds rather than unconstrained Euclidean spaces. CLIP-style vision-language embeddings are commonly normalized onto hyperspheres (Radford et al., 2021), hyperbolic image-text models such as MERU place embeddings on Lorentz hyperbolic manifolds (Desai et al., 2023), and kinase–ligand structural datasets such as KLIFS contain ligand poses that can be represented as rigid motions in $\mathrm{SE}(3)$ (Kooistra et al., 2016). For such manifold-valued representations, debiasing requires reducing nuisance information while preserving downstream utility and manifold validity.

Existing post-hoc methods typically realize debiasing through Euclidean editing operations, including projection-based subspace removal (Bolukbasi et al., 2016; Ravfogel et al., 2020), linear concept erasure (Ravfogel et al., 2022; Belrose et al., 2023), feature intervention (Jung et al., 2024), and ambient-space edits (Zhao et al., 2025; Gerych et al., 2026). These approaches are simple and effective baselines, and several preserve selected constraints in specific settings. Their limitation in our setting is that the editing operation is not generally defined as a native transformation of the target manifold. As a result, an edit that is valid in Euclidean feature space does not generally preserve unit norm on the hypersphere, Lorentz validity in hyperbolic space, or the orthogonality and determinant constraints of a rotation matrix of SE(3). Projection or retraction restores validity after editing, but provides no guarantee of preserving the nuisance–utility trade-off intended by the original edit. Manifold-aware and geometry-inspired debiasing methods provide important alternatives (Kumar et al., 2021; Zhong et al., 2026; Rathore, 2026). Still, they are often tied to a particular geometry, generative setting, representation space, or training-time objective, leaving open how to combine post-hoc nuisance estimation with manifold-native editing across target geometries.

This gap highlights the central mismatch addressed in this paper. The issue is not the use of Euclidean probe features themselves, but how the estimated nuisance directions are executed as representation edits. Nuisance probes and subspace-estimation methods naturally operate on Euclidean probe features, where nuisance-related directions are estimated from nuisance labels (Ravfogel et al., 2020). The representation itself, however, is a point on a target manifold

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

[1]Code: https://anonymous.4open.science/r/IsoAct-84C6/

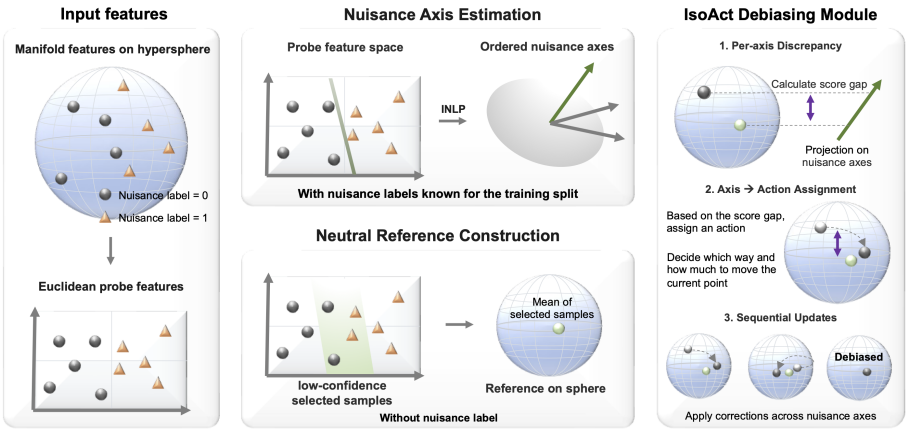

Figure 1. Overview of the IsoAct framework. IsoAct first estimates nuisance axes in Euclidean probe space, then constructs a neutral manifold reference from low-confidence samples. In inference, it applies sequential geometry-specific actions to reduce nuisance-axis discrepancies relative to the neutral reference, producing debiased representations while preserving manifold validity.

$\mathcal{M}$, and admissible edits should respect the geometry of $\mathcal{M}$. This suggests a geometric version of post-hoc debiasing that retains the practical probe-based estimation pipeline while executing the edit through manifold-native action primitives.

Building on this principle, we propose IsoAct, a post-hoc representation-editing framework that reduces nuisance information through manifold-native isometric action primitives. IsoAct estimates ordered nuisance axes in Euclidean probe features, constructs a neutral reference for the target nuisance score, and uses an axis-to-action policy to assign each axis to an admissible action family on the target manifold. The selected actions are composed with sample-dependent magnitudes, so the edit adjusts nuisance-related geometry while each primitive preserves manifold validity. The admissible action family is determined by the target geometry, taking the form of rotations on hyperspherical representations, Lorentz transformations in hyperbolic space, and rigid-motion updates on $\mathrm{SE}(3)$.

We evaluate IsoAct across synthetic and real-world settings on hyperspherical, Lorentz hyperbolic, and $\mathrm{SE}(3)$ representations (Radford et al., 2021; Desai et al., 2023; Kooistra et al., 2016). Using debiasing capability, utility preservation, and manifold-constraint violation as evaluation axes, we find that IsoAct reduces nuisance leakage while maintaining downstream utility and manifold validity. Together, these results highlight the importance of assessing post-hoc debiasing for manifold-valued representations as a joint nuisance–utility–validity problem, and show that manifold-native action primitives provide a practical mechanism for improving this trade-off. Accordingly, this work formulates validity-aware post-hoc debiasing for manifold-valued representations, introduces IsoAct as a manifold-native editing framework, and instantiates it on hyperspherical, Lorentz hyperbolic, and $\mathrm{SE}(3)$ representations.

2. Related Work

Post-hoc debiasing and concept erasure. Post-hoc debiasing and concept erasure have been extensively studied as practical approaches to removing protected or nuisance information from pretrained representations (Bolukbasi et al., 2016; Meade et al., 2022). INLP (Ravfogel et al., 2020) removes linearly decodable protected-attribute information by iteratively training linear probes and projecting representations onto the intersection of their nullspaces. RLACE (Ravfogel et al., 2022) and LEACE (Belrose et al., 2023) further formulate linear concept erasure through a constrained adversarial game and a closed-form oblique projection, respectively, establishing key baselines for linear concept erasure. SFID (Jung et al., 2024) mitigates bias through selective feature imputation over bias-relevant embedding coordinates, whereas SPD (Zhao et al., 2025) models bias as a linear subspace and removes it via subspace projection with neutral mean reinjection. While IsoAct adopts the probe-based nuisance-axis estimation pipeline from these works, it departs by executing the estimated axes as an admissible action primitive on the target manifold, rather than through Euclidean projection or coordinate replacement.

Geometry-aware debiasing for manifold-valued representations. Euclidean projection can cause information loss by erasing valid associations (Dev et al., 2021) and may leave residual bias in the embedding geometry beyond the targeted bias direction (Gonen & Goldberg, 2019), motivating geometry-aware debiasing methods. WRING (Gerych et al., 2026) replaces projection with rotations within a bias-relevant subspace of VLM embeddings, FAIRT2V (Zhong et al., 2026) uses anchor-based geodesic transformations on the hypersphere for training-free text-to-video debiasing, and PGD (Kumar et al., 2021) mitigates gender bias in Poincaré word embeddings via per-word Riemannian

Table 1. Target manifolds and validity constraints. For the Lorentz hyperboloid, $c > 0$ is the curvature parameter, h_0 is the time-like coordinate, and $\langle \cdot, \cdot \rangle_L$ denotes the Lorentz inner product.

Manifold	Representation	Validity Constraint	
Hypersphere	$q \in \mathbb{S}^{d-1}$	1) $\|q\|_2 = 1$	
Lorentz hyperboloid	$h \in \mathbb{H}^d_c$	1) $\langle h, h \rangle_L = -1/c$	2) $h_0 > 0$
SE(3)	$g = (R, t) \in \mathrm{SE}(3)$	1) $R^\top R = I$	2) $\det(R) = 1$

optimization. These methods are similarly motivated but remain tied to specific geometries, representations, or domains. IsoAct instead provides an axis-to-action framework for manifold-valued representations, including hyperspherical normalized contrastive embeddings (Radford et al., 2021; Wang & Isola, 2020), Lorentz hyperbolic embeddings (Desai et al., 2023; Nickel & Kiela, 2018), and SE(3) representations (Kooistra et al., 2016). In these settings, IsoAct removes nuisance information while preserving geometry-specific validity constraints.

Geometric fairness and training-time invariance. Fairness and debiasing have also been studied from the perspectives of representation geometry and invariance (He et al., 2020). CMF (Rathore, 2026) enforces preservation of local Riemannian geometry across factual and counterfactual representations induced by sensitive-attribute interventions, through training-time regularization of the decoder Jacobian and Hessian. Unlike such training-time approaches, IsoAct operates post hoc on fixed pretrained representations, complementing settings where retraining is costly or model internals are inaccessible.

3. Preliminaries

3.1. Problem Setting

We study post-hoc debiasing of pretrained representations on a known target manifold. Let $\mathcal{D} = \{(z_i, s_i, y_i)\}_{i=1}^n$, where n is the number of samples, $z_i \in \mathcal{M}$ is the i-th representation on the target manifold, s_i is the nuisance variable, and y_i is the utility variable. s_i denotes an attribute whose information should be reduced in the edited representation for the target use case. It includes protected attributes such as gender or race, as well as non-protected but undesired attributes such as data source, batch label, or domain label. In contrast, y_i denotes the task-relevant variable whose predictive information should be preserved after editing.

The goal of post-hoc debiasing is to reduce information about s_i in the edited representation while preserving information useful for predicting y_i. A post-hoc debiasing method constructs a map $T : \mathcal{M} \to \mathcal{M}$ that is applied after representation extraction, without retraining the upstream model. The transformation is expected to reduce information about s_i in $T(z_i)$, preserve information about y_i, ensure that the output remains valid on the target manifold, i.e., $T(z_i) \in \mathcal{M}$.

3.2. Target Manifolds

We consider three target geometries that commonly arise in structured representation learning. The hypersphere $\mathbb{S}^{d-1}$ represents normalized embeddings q with fixed Euclidean norm, as used in contrastive representation learning and vision-language models (Radford et al., 2021). The Lorentz hyperboloid $\mathbb{H}^d_c$ represents hyperbolic embeddings h with constant negative curvature, which are widely used for hierarchical or entailment-like structure and have been adopted in image-text representation learning, such as MERU (Desai et al., 2023). The group SE(3) represents rigid-body poses $g = (R, t)$, with rotation $R \in \mathrm{SO}(3)$ and translation $t \in \mathbb{R}^3$, and is standard in robotics, vision, structural modeling, and kinase–ligand pose representation (Murray et al., 2017; Hartley et al., 2013; Kooistra et al., 2016).

Table 1 summarizes the representation forms and validity constraints. Throughout the paper, manifold validity is measured by residuals derived from these standard constraints (Absil et al., 2008; Cannon et al., 1997; Murray et al., 2017; Hartley et al., 2013).

3.3. Euclidean Probe Features for Nuisance Estimation

For post-hoc editing of manifold-valued representations, preserving manifold validity is a natural requirement. This requirement is not guaranteed by generic coordinate-space interventions, since they may move a representation away from the target manifold. At the same time, nuisance information is often estimated in an auxiliary Euclidean probe space. For each representation $z_i \in \mathcal{M}$, we define

$$\phi : \mathcal{M} \to \mathbb{R}^D, \qquad v_i = \phi(z_i),$$

where D is the probe-feature dimension. The probe feature v_i is used to estimate nuisance-predictive axes and to compute the nuisance-axis scores used by IsoAct.

For ambient manifolds, such as the hypersphere and the Lorentz hyperboloid, ϕ uses the ambient coordinates of the representation. For structured representations, such as SE(3) poses, ϕ uses a coordinate representation. Since $R \in \mathrm{SO}(3) \subset \mathbb{R}^{3\times3}$ and $t \in \mathbb{R}^3$, we use $\phi(R, t) = [\mathrm{vec}(R), t] \in \mathbb{R}^{12}$. Thus, the nuisance axis is estimated in the probe-feature space, rather than assumed to be a tangent direction on the manifold. IsoAct uses this probe-space

estimate to compute a nuisance discrepancy, while the actual update is performed through a geometry-specific action on $\mathcal{M}$. Details of nuisance-axis estimation and neutral-reference construction are given in Section 4.

4. IsoAct: Isometric Action Debiasing

IsoAct is a post-hoc editing method for manifold-valued representations. Unlike coordinate- or subspace-based post-hoc methods such as SFID and SPD (Jung et al., 2024; Zhao et al., 2025), which edit representations through Euclidean coordinate replacement or projection, IsoAct replaces the final edit with geometry-specific action primitives. An action primitive is a constraint-preserving map $\mathcal{M} \to \mathcal{M}$, such as a hyperspherical rotation, a Lorentz transformation, or a group-valid $\mathrm{SE}(3)$ update. The full debiasing transformation is obtained by composing these primitives across the retained nuisance axes.

At a high level, IsoAct estimates two training-derived objects, nuisance axes in a Euclidean probe space and a neutral reference on the target manifold. Given representations $z_i \in \mathcal{M}$ and probe features $v_i = \phi(z_i) \in \mathbb{R}^D$, IsoAct obtains K nuisance axes $u_1, \ldots, u_K$ and a neutral reference $z_{\mathrm{ref}} \in \mathcal{M}$. The axes specify the probe-space directions along which nuisance information is measured, while z_{ref} provides a common neutral anchor for these measurements. For each sample and retained axis, IsoAct compares the current probe projection with the corresponding neutral-anchor projection, converts their discrepancy into an action score, and applies the action primitive on the target manifold.

4.1. Nuisance Axes and Neutral Reference

We estimate nuisance axes in the Euclidean probe space by finding directions along which the nuisance variable is linearly decodable. These axes provide probe-space directions for measuring sample-wise nuisance discrepancies relative to a neutral reference. We use Iterative Nullspace Projection (INLP) (Ravfogel et al., 2020) to estimate the axes. This yields K ordered, normalized axes $u_1, \ldots, u_K \in \mathbb{R}^D$, with $\|u_k\|_2 = 1$. The axes are fitted using the training split only. For multiclass nuisance variables, we use one-vs-rest linear nuisance probes and orthonormalize the extracted directions. Implementation details are provided in Appendix B.1.

The neutral reference specifies the nuisance-axis score toward which edited representations are moved. Following SFID and SPD (Jung et al., 2024; Zhao et al., 2025), we identify low-confidence training examples using an auxiliary random-forest nuisance classifier, average them to form z_{ref}, and project or retract the result to the target manifold when necessary. If the low-confidence set is unstable, we use the training mean instead. Oracle references are used only as controlled-benchmark diagnostics.

After the nuisance axes and neutral reference are fixed, IsoAct computes an action score for each sequential update. For sample i and axis k, we first compare the current probe score with the neutral-reference score

$$\delta_{ik}^{\mathrm{score}} = \langle \phi(z_{\mathrm{ref}}), u_k \rangle - \langle \phi(z_i^{(k-1)}), u_k \rangle. \tag{1}$$

This signed discrepancy measures how the current representation differs from the neutral reference along u_k. A nonnegative scalar α, selected on validation data, scales this discrepancy, giving the action score $\alpha\delta_{ik}^{\mathrm{score}}$. For the hypersphere, Lorentz hyperboloid, and the translation component of $\mathrm{SE}(3)$, this action score directly determines the update strength. For the rotation component of $\mathrm{SE}(3)$, IsoAct instead uses an $\mathrm{SO}(3)$-native discrepancy, since displacement in vectorized rotation coordinates does not directly define a valid motion on $\mathrm{SO}(3)$.

4.2. Geometry-specific Action Primitives

We now instantiate how the action score is converted into a manifold-valid action for each target geometry. For the hypersphere and Lorentz hyperboloid, the action score, together with the direction induced by the retained nuisance axis, specifies an isometric transformation. For $\mathrm{SE}(3)$, IsoAct separates rotation and translation actions so that each component remains valid after editing.

Hyperspherical and Lorentz actions. For hyperspherical and Lorentz representations, the action score $\alpha\delta_{ik}^{\mathrm{score}}$ is used to parameterize a geometry-preserving linear map. Let $\eta_{ik} = \alpha\delta_{ik}^{\mathrm{score}}$. For the sphere, the update can be written as $O_{ik} = \exp(\eta_{ik} G_k)$, where $G_k^\top = -G_k$ is a skew-symmetric generator for the selected rotation plane. For the Lorentz hyperboloid, the analogous form is $L_{ik} = \exp(\eta_{ik} B_k)$, where $B_k^\top J + J B_k = 0$, so L_{ik} preserves the Lorentz inner product. On the hypersphere, IsoAct uses an orthogonal map to update the representation of the sample index i, $q_i^{(k)}$ as

$$q_i^{(k)} = O_{ik} q_i^{(k-1)}, \qquad O_{ik}^\top O_{ik} = I.$$

Since O_{ik} is orthogonal, the unit-norm constraint is preserved. On the Lorentz hyperboloid, to update the representation $h_i^{(k)}$ IsoAct uses

$$h_i^{(k)} = L_{ik} h_i^{(k-1)},$$

where L_{ik} is a proper orthochronous Lorentz transformation satisfying

$$L_{ik}^\top J L_{ik} = J, \qquad J = \mathrm{diag}(-1, 1, \ldots, 1).$$

This preserves the Lorentz quadratic constraint and the upper-sheet condition. Additional construction details are provided in Appendix B.2.

SE(3) action and fixed modes. For $g_i = (R_i, t_i) \in \mathrm{SE}(3)$, with $R_i \in \mathrm{SO}(3)$ and $t_i \in \mathbb{R}^3$, the nuisance axis u_k is split into rotation-generator direction $a_k \in \mathbb{R}^3$ and translation direction $b_k \in \mathbb{R}^3$. Appendix B.3 describes this extraction, including zero-norm cases and vectorization. The rotation component lies in $\mathrm{SO}(3)$, so IsoAct measures the reference-directed correction in the rotation tangent space instead of using the probe-score discrepancy. The rotation action and action factor are

$$R_i^{(k)} = R_i^{(k-1)} Q_{ik}, \qquad Q_{ik} = \mathrm{Exp}_{\mathrm{SO}(3)}\left(\alpha \delta_{ik}^R \widehat{a}_k\right), \quad (2)$$

where the native rotation discrepancy is $\delta_{ik}^R = \langle (\mathrm{Log}_{\mathrm{SO}(3)}((R_i^{(k-1)})^\top R_{\mathrm{ref}}))^\vee, a_k \rangle$. Here R_{ref} is the rotation component of g_{ref}, $\widehat{a}_k$ is the skew-symmetric matrix of a_k, and $(\cdot)^\vee$ maps a skew-symmetric matrix to its vector form. Since $Q_{ik} \in \mathrm{SO}(3)$, the rotation remains valid after the action.

The translation action uses the probe-score discrepancy from Eq. (1)

$$t_i^{(k)} = t_i^{(k-1)} + \alpha \delta_{ik}^{\mathrm{score}} b_k. \quad (3)$$

For $\mathrm{SE}(3)$, the final action mode is fixed for a run and selected on validation data. Rotation-only IsoAct applies Eq. (2) and leaves t fixed. Translation-only IsoAct applies Eq. (3) and leaves R fixed. Joint rotation–translation IsoAct applies both updates sequentially for each retained axis. We use component-level nuisance probes on the rotation block, translation block, and their concatenation to diagnose where nuisance information is decodable. Probe accuracy above the majority-class baseline indicates measurable nuisance information in the corresponding component.

The following theorem formalizes the manifold-validity guarantee for the $\mathrm{SE}(3)$ updates. The proof is given in Appendix B.5.

Theorem 4.1. *IsoAct on* $\mathrm{SE}(3)$ *maps into* $\mathrm{SE}(3)$.

Scope of the action maps. IsoAct guarantees per-sample manifold validity, not global distance preservation. For fixed action scores, the hyperspherical and Lorentz primitives are isometries, and the $\mathrm{SE}(3)$ update is group-valid by construction. Because debiasing uses sample-dependent nuisance discrepancies scaled by α, the resulting dataset-level map is generally not a single global isometry. Pairwise distances are preserved only when the same action is applied globally to all samples.

4.3. Selection protocol

Model selection uses the validation split. We consider a grid over the number of nuisance axes K, the global action strength α, and, when applicable, the action mode. Selection follows a utility-preserving debiasing criterion. Candidates that collapse the designated utility signal relative to the original representation are not treated as successful debiasing, since removing nuisance information by destroying task-relevant structure is representation corruption rather than a useful edit. Among utility-preserving candidates, we choose the configuration with the strongest nuisance removal under validation probes, with experiment-specific metrics specified in the corresponding experimental section. The held-out test split is used only for final reporting and uncertainty estimation.

5. Experiments

We evaluate post-hoc debiasing along three axes: *nuisance removal*, measured by how much nuisance information remains decodable after editing; *utility preservation*, measured by whether the task-relevant signal is retained; and *manifold validity*, measured by constraint residuals of the edited representation. This joint evaluation is essential because an ambient-space edit may reduce nuisance-probe performance while perturbing task-relevant structure or leaving the target manifold. For probe-based metrics, predictors are trained on edited training representations and evaluated on the corresponding edited validation or test representations. Method-specific hyperparameters are selected on validation data, as described in Section 4.3, and final numbers are reported on held-out test data.

Our experiments have two roles. The controlled synthetic benchmarks isolate the mechanism of manifold-native editing under known nuisance structure and known geometry. The KLIFS benchmark then tests the same principle in a real-world ligand-pose setting on $\mathrm{SE}(3)$.

5.1. Controlled synthetic benchmarks

5.1.1. Setup

We construct controlled benchmarks on three target geometries: the hypersphere $\mathbb{S}^{d-1}$, the Lorentz hyperboloid $\mathbb{H}_c^d$, and $\mathrm{SE}(3)$. In each benchmark, the nuisance variable is injected into a known geometric degree of freedom, while a separate coordinate is treated as the task-relevant variable. The goal is not to evaluate a semantic downstream task, but to test whether an edit removes nuisance separability while preserving the designated task-relevant variable and manifold validity.

The injection patterns expose geometry-specific failure modes: on the hypersphere, nuisance and task signals occupy separate angular degrees of freedom because the norm is fixed; on the Lorentz hyperboloid, nuisance is injected into the angular coordinate θ and the task variable is the radial coordinate r; on $\mathrm{SE}(3)$, nuisance is injected into the rotation angle φ, and the task variable is the translation magnitude $\|t\|$.

Table 2. Controlled mechanism checks on three target manifolds. It compares reference, learning-based, and post-hoc methods on debiasing capability, utility preservation, and unit-norm violation. The Original and Oracle rows report raw nuisance-probe metrics as references. For edited methods, the debiasing columns report oracle discrepancies, with $|\Delta m|$ denoting the absolute deviation from the oracle value for metric m. IsoAct matches oracle-level debiasing while preserving near-zero constraint violation. † denotes a learning-based method, and reference results are shown in gray. For the constraint columns, C1 denotes unit-norm error on the hypersphere; C1/C2 denote Lorentz quadratic error and upper-sheet violation on the Lorentz hyperboloid; and C1/C2 denote $\mathrm{SO}(3)$ orthogonality and determinant errors on $\mathrm{SE}(3)$.

		Debiasing Capability				Utility Preservation		Manifold Constraint Violation	
	Method	$\lvert\Delta\mathrm{Acc}\rvert\downarrow$	$\lvert\Delta\mathrm{Recall}\rvert\downarrow$	$\lvert\Delta\mathrm{F1}\rvert\downarrow$	$\lvert\Delta\mathrm{AUROC}\rvert\downarrow$	Task MAE↓	Task In-Band↑	Constraint 1↓	Constraint 2↓
Hypersphere	Original	1.0000	1.0000	1.0000	1.0000	0.0000	1.0000	0.0000	—
	Oracle	0.5083	0.5083	0.5080	0.4956	0.0000	1.0000	0.0000	—
	CMF†	0.0083	0.0083	0.0086	0.0023	3.2653	0.1050	0.0211	—
	SFID	0.0083	0.0083	0.0102	**0.0000**	13.1276	0.0525	0.0813	—
	SPD	0.0083	0.0083	0.0081	0.0109	2.9891	0.3050	0.0242	—
	IsoAct	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**1.0000**	**0.0000**	—
Lorentz	Original	1.0000	1.0000	1.0000	1.0000	0.0000	1.0000	0.0000	0.0000
	Oracle	0.5000	0.5000	0.4999	0.5008	0.0000	1.0000	0.0000	0.0000
	CMF†	0.5000	0.5000	0.5001	0.4992	0.0161	0.9775	3.6672	**0.0000**
	SFID	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**1.0000**	15.2018	**0.0000**
	SPD	0.0167	0.0167	0.0202	0.0125	0.0009	**1.0000**	10.9841	**0.0000**
	IsoAct	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**1.0000**	**0.0000**	**0.0000**
SE(3)	Original	1.0000	1.0000	1.0000	1.0000	0.0000	1.0000	0.0000	0.0000
	Oracle	0.4833	0.4833	0.4828	0.5133	0.0000	1.0000	0.0000	0.0000
	CMF†	0.0584	0.0584	0.0586	0.0067	0.4223	0.6300	1.2249	0.8678
	SFID	0.0250	0.0250	0.0245	0.0377	0.4636	0.5825	0.1797	0.0086
	SPD	**0.0000**	**0.0000**	**0.0000**	0.0009	0.0104	0.9925	1.0358	0.7299
	IsoAct	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**0.0000**	**1.0000**	**0.0000**	**0.0000**

We compare the unedited representation, an oracle edit available only in the controlled construction, CMF as a learning-based geometric-fairness baseline, and Euclidean post-hoc feature-intervention and subspace-projection baselines. Because the oracle is known, the nuisance columns in Table 2 report oracle discrepancies: for each nuisance metric m, $|\Delta m| = |m_{\text{method}} - m_{\text{oracle}}|$. This measures whether a method matches the intended neutralized reference, rather than rewarding any raw decrease in probe performance. The task columns report the absolute error of the task-relevant variable and the fraction of edited points that remain inside the original band of the task-relevant variable.

5.1.2. Results

Table 2 shows that nuisance removal alone is insufficient for manifold-valued debiasing. Euclidean baselines reduce the injected nuisance signal, confirming that the nuisance direction is identifiable in the probe space, but they often perturb the task-relevant variable or produce nonzero constraint residuals. This is most visible in the Lorentz and $\mathrm{SE}(3)$ rows, where ambient feature replacement or projection achieve low nuisance discrepancy while violating the target geometry. IsoAct matches the oracle discrepancy in these controlled settings while keeping the task-relevant variable and validity residuals at their intended values.

Figure 2 visualizes this mechanism for the $\mathrm{SE}(3)$ benchmark, with analogous hyperspherical and Lorentz hyperbolic visualizations provided in Appendix C.1. The horizontal axis is the injected nuisance coordinate and the vertical axis is the task-relevant variable. The desired edit removes horizontal separability while keeping points in the shaded task-relevant band. IsoAct moves the nuisance coordinate toward the reference value while preserving the vertical structure, whereas some baselines either leave residual nuisance separation or distort the task-relevant variable.

These benchmarks establish a mechanism-level result in which manifold-native actions provide a favorable nuisance–task–validity trade-off when nuisance aligns with a known geometric degree of freedom. We next test whether this principle remains useful in real-world $\mathrm{SE}(3)$ representations.

5.2. Real-world SE(3) ligand-pose experiment on KLIFS

We evaluate IsoAct on KLIFS kinase–ligand structural data (Kooistra et al., 2016). Each example is a ligand pose $z = (R, t) \in \mathrm{SE}(3)$, where R denotes ligand orientation and t denotes its position in the binding site (Shen et al., 2021; Zhang et al., 2022). This setting naturally requires manifold-aware editing, since a non-orthogonal rotation block or a determinant different from one is not a valid rigid pose, even if its vectorized coordinates yield a favorable probe score (Mueller, 2019).

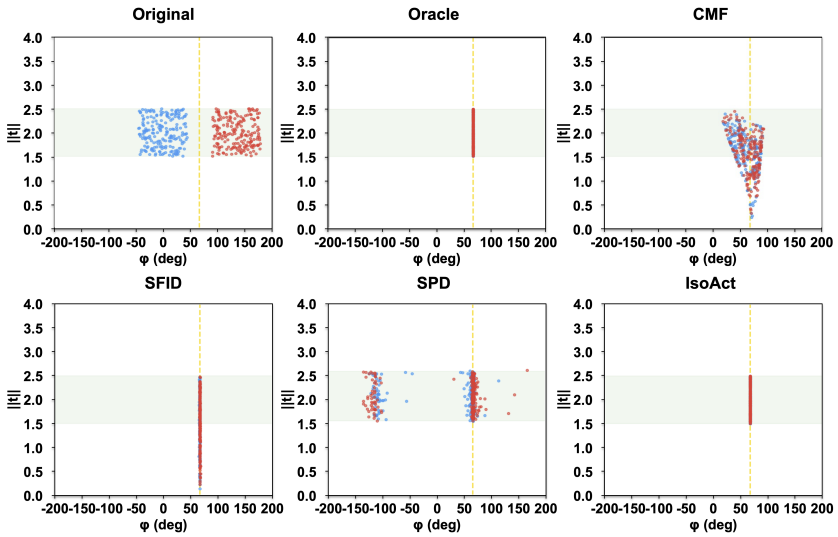

Figure 2. SE(3) controlled synthetic experiment scatter plot. The horizontal axis φ is the injected nuisance coordinate, and the vertical axis $\|t\|$ is the task-relevant variable. Blue/red points indicate the two nuisance groups, the yellow dashed line marks the neutral reference, and the light-green band marks the target task-relevant range. An ideal edit removes horizontal group separation while preserving the vertical structure.

Table 3. Real-world KLIFS SE(3) ligand-pose debiasing results. Lower debiasing and constraint values are better, while higher utility values are better; IsoAct gives a favorable nuisance–utility–validity operating point, with the strongest reduction on most nuisance metrics while preserving competitive structural utility and near-zero SO(3) residuals.

		Debiasing Capability				Utility Preservation								Manifold Constraint Violation	
						Task 1: αC_helix state Classification				Task 2: pocket occupancy Classification				Constraint 1	Constraint 2
	Method	Acc↓	Recall↓	F1↓	AUROC↓	Acc↑	Recall↑	F1↑	AUROC↑	Acc↑	Recall↑	F1↑	AUROC↑	Mean↓	Mean↓
KLIFS	Null bias	0.0827	0.1172	0.0679	0.4996	—	—	—	—	—	—	—	—	—	—
	Original	0.1395	0.2427	0.1404	0.6158	0.8532	0.3374	0.3846	0.7169	0.9248	0.7357	0.7921	0.9708	0.0000	0.0000
	CMF†	0.1438	0.2802	0.1432	**0.5780**	0.8393	0.1266	0.1765	0.5623	0.8985	0.6724	0.7365	0.9347	0.0072	0.0048
	SFID	0.1395	0.2427	0.1404	0.6158	0.8509	0.3263	0.3730	0.7184	0.9268	0.7388	0.7860	0.9712	**0.0000**	**0.0000**
	SPD	0.1383	0.2408	0.1392	0.6151	0.8577	0.3490	0.4002	0.7162	**0.9271**	0.7534	0.7820	**0.9724**	0.0037	0.0012
	IsoAct	**0.1156**	**0.1853**	**0.1037**	0.5853	**0.8609**	**0.3752**	**0.4231**	**0.7564**	0.9228	**0.8052**	**0.7998**	0.9649	**0.0000**	**0.0000**

The nuisance variable is kinase group, a coarse kinase-domain grouping derived from standard kinase classifications (Modi & Dunbrack Jr, 2019). Prior work has shown that kinase sequence similarity and phylogenetic relationships can induce shortcut learning or information leakage in kinase-related prediction tasks. In our KLIFS experiments, we therefore use kinase group as a coarse, biologically meaningful proxy for kinase-family identity, and ask whether post-hoc debiasing can reduce this group-level signal while preserving ligand-pose and kinase-structure utility (Kanakala et al., 2023; Ong et al., 2023). The utility tasks are αC-helix state classification and pocket-occupancy classification, which capture kinase conformation and ligand placement within the binding site (Kooistra et al., 2016).

Before applying IsoAct, we diagnose where kinase-group information is decodable. Unlike the controlled synthetic setting, where nuisance is injected into a single chosen component, we find that kinase-group information is not clearly decodable from the rotation or translation block alone, but becomes measurable from the concatenated pose (R, t). We therefore use an SE(3) IsoAct configuration that allows both rotation and translation updates for KLIFS, testing whether IsoAct remains effective when nuisance information is distributed across the pose representation rather than isolated in a single component.

Table 3 reports the main compact comparison. Nuisance removal is measured by linear probes, utility preservation by the two structural tasks, and manifold validity by the mean SO(3) orthogonality and determinant residuals. Since the structural utility labels are imbalanced, we report Recall and F1 alongside accuracy and AUROC. The null bias row is obtained by evaluating the same nuisance probes under shuffled kinase-group labels and serves as the ideally debiased point. For readability, Table 3 reports central values only; the full table with bootstrap uncertainty estimates is provided in Appendix C.2. Uncertainty is estimated using fixed-test bootstrap resampling with 1,000 resamples, with deterministic debiasing transforms held fixed.

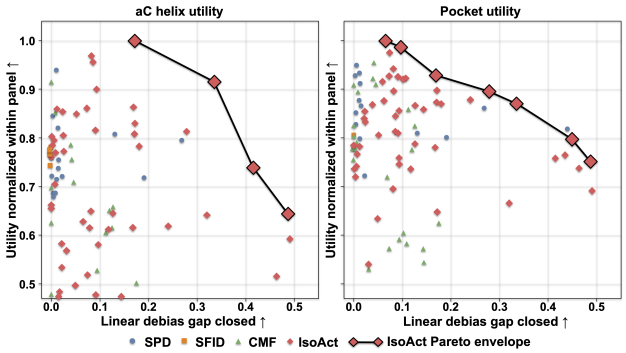

Figure 3. IsoAct trade-off envelope on KLIFS. The horizontal axis measures the fraction of the linear debiasing gap closed, and the vertical axis measures utility normalized within each panel. The envelope shows the nuisance–utility trade-offs within the IsoAct configuration pool.

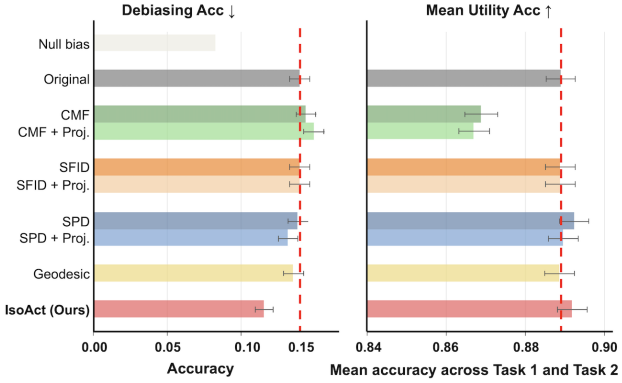

Figure 4. KLIFS accuracy comparison with projected and geodesic diagnostic variants. The geodesic row is not an external baseline but an on-manifold diagnostic testing whether generic manifold movement is sufficient once raw projection is questioned. The comparison suggests that projection repair or generic on-manifold movement alone is insufficient, and that the action must be tied to the relevant geometry and nuisance direction.

The table shows that IsoAct reduces kinase-group recoverability relative to the original representation while keeping the $\mathrm{SE}(3)$ constraint residuals at numerical zero. It also preserves the selected structural utilities: IsoAct improves the αC-helix metrics in this comparison and remains competitive on pocket occupancy, with stronger recall and F1 but not the best value on every pocket metric. Thus, the relevant claim is not uniform dominance on every score, but a favorable nuisance–utility–validity operating point in a structural pose setting.

Table 3 reports one validation-selected operating point. Figure 3 shows a complementary Pareto view over a representative matched pool of candidate configurations with comparable hyperparameter choices. IsoAct spans multiple nuisance–utility regimes: some configurations close more of the linear debiasing gap at a utility cost, while others preserve more utility. We use this figure to interpret the trade-off and selection behavior, not to claim that every IsoAct instance dominates every baseline.

Figure 4 complements the table by including projected variants and a geodesic diagnostic, with the corresponding full table in Appendix C.2. The projected variants test whether repairing Euclidean edits by projecting back to the manifold is sufficient. The geodesic row is an on-manifold diagnostic asking whether a generic manifold move is enough once raw projection is challenged. The comparison indicates that validity repair and generic on-manifold movement alone are insufficient, and that the edit must be tied to the nuisance direction and representation geometry.

5.3. Empirical summary and scope

Together, the experiments support the central claim that post-hoc debiasing benefits from respecting representation geometry. The controlled benchmarks show the mechanism under known nuisance locations: manifold-native actions remove injected nuisance information while preserving the designated task-relevant variable and maintaining validity. The KLIFS experiment shows that the same principle can yield a useful trade-off in a real-world $\mathrm{SE}(3)$ ligand-pose setting, including the harder case where nuisance information is decodable only from the joint rotation–translation representation. The empirical claim is intentionally limited. We do not claim that IsoAct is uniformly best across all metrics, baselines, manifolds, or domains. The selected KLIFS configuration should be interpreted through the joint nuisance–utility–validity trade-off rather than through a single metric.

6. Conclusion

We introduced IsoAct, a post-hoc debiasing framework for representations constrained to known manifolds. IsoAct preserves the practicality of probe-based concept erasure while replacing Euclidean projection or coordinate repair with manifold-native action primitives. Across controlled synthetic benchmarks and a real-world KLIFS $\mathrm{SE}(3)$ ligand-pose setting, IsoAct improves the nuisance–utility–validity trade-off by reducing decodable nuisance information while preserving task-relevant structure and manifold validity. Additional projection and geodesic diagnostics suggest that validity repair or generic on-manifold movement alone is insufficient; effective debiasing should be tied to both nuisance directions and target geometry. IsoAct does not guarantee complete concept removal or fairness, and its sample-dependent actions are not a single global isometry. Future work can extend action-based debiasing to learned geometries, richer nuisance definitions, and broader manifold-valued representations.

Impact Statement

This paper studies post-hoc debiasing for representations constrained to known manifolds. Its potential positive impact is to make representation editing more reliable in settings where nuisance information should be reduced while geometric validity must be preserved, such as normalized embeddings, hyperbolic representations, and rigid-pose representations. By evaluating nuisance removal jointly with utility preservation and manifold-constraint residuals, the work encourages more complete reporting of debiasing behavior than nuisance-probe performance alone.

At the same time, IsoAct should not be interpreted as a guarantee of fairness, safety, or complete concept removal. The method depends on the availability and quality of nuisance labels, the chosen probe features, the neutral reference, and validation-selected hyperparameters; nonlinear or unmeasured biases may remain. In sensitive or biomedical applications, an edit that suppresses a proxy attribute may also remove task-relevant structure or create a misleading appearance of fairness if evaluated only through a narrow set of probes. IsoAct should therefore be used only as one component of a broader evaluation pipeline that includes subgroup analysis, domain-specific auditing, downstream utility and safety checks, and clear disclosure of the debiasing objective and its limitations.

References

Absil, P.-A., Mahony, R., and Sepulchre, R. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

Belrose, N., Schneider-Joseph, D., Ravfogel, S., Cotterell, R., Raff, E., and Biderman, S. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063, 2023.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Cannon, J. W., Floyd, W. J., Kenyon, R., Parry, W. R., et al. Hyperbolic geometry. *Flavors of geometry*, 31(59-115): 2, 1997.

Desai, K., Nickel, M., Rajpurohit, T., Johnson, J., and Vedantam, S. R. Hyperbolic image-text representations. In *International Conference on Machine Learning*, pp. 7694–7731. PMLR, 2023.

Dev, S., Li, T., Phillips, J. M., and Srikumar, V. Oscar: Orthogonal subspace correction and rectification of biases in word embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5034–5050, 2021.

Gerych, W., Parent, C., Perian, Q., Javed, R., Solomon, J., and Ghassemi, M. Wring out the bias: A rotation-based alternative to projection debiasing. In *The Fourteenth International Conference on Learning Representations*, 2026.

Gonen, H. and Goldberg, Y. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *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. 609–614, 2019.

Hartley, R., Trumpf, J., Dai, Y., and Li, H. Rotation averaging. *International journal of computer vision*, 103(3): 267–305, 2013.

He, Y., Burghardt, K., and Lerman, K. A geometric solution to fair representations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 279–285, 2020.

Jung, H., Jang, T., and Wang, X. A unified debiasing approach for vision-language models across modalities and tasks. *Advances in Neural Information Processing Systems*, 37:21034–21058, 2024.

Kanakala, G. C., Aggarwal, R., Nayar, D., and Priyakumar, U. D. Latent biases in machine learning models for predicting binding affinities using popular data sets. *ACS omega*, 8(2):2389–2397, 2023.

Kooistra, A. J., Kanev, G. K., van Linden, O. P., Leurs, R., de Esch, I. J., and de Graaf, C. Klifs: a structural kinase-ligand interaction database. *Nucleic acids research*, 44 (D1):D365–D371, 2016.

Kumar, V., Bhotia, T. S., and Chakraborty, T. Identifying and mitigating gender bias in hyperbolic word embeddings. *arXiv preprint arXiv:2109.13767*, 2021.

Meade, N., Poole-Dayan, E., and Reddy, S. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 1878–1898, 2022.

Modi, V. and Dunbrack Jr, R. L. A structurally-validated multiple sequence alignment of 497 human protein kinase domains. *Scientific reports*, 9(1):19790, 2019.

Mueller, A. Modern robotics: Mechanics, planning, and control [bookshelf]. *IEEE Control Systems Magazine*, 39 (6):100–102, 2019.

Murray, R. M., Li, Z., and Sastry, S. S. *A mathematical introduction to robotic manipulation*. CRC press, 2017.

Nickel, M. and Kiela, D. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International conference on machine learning*, pp. 3779–3788. PMLR, 2018.

Ong, W. J. G., Kirubakaran, P., and Karanicolas, J. Poor generalization by current deep learning models for predicting binding affinities of kinase inhibitors. *bioRxiv*, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Rathore, V. Causal manifold fairness: Enforcing geometric invariance in representation learning. *arXiv preprint arXiv:2601.03032*, 2026.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., and Goldberg, Y. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 7237–7256, 2020.

Ravfogel, S., Twiton, M., Goldberg, Y., and Cotterell, R. D. Linear adversarial concept erasure. In *International Conference on Machine Learning*, pp. 18400–18421. PMLR, 2022.

Shen, C., Hu, X., Gao, J., Zhang, X., Zhong, H., Wang, Z., Xu, L., Kang, Y., Cao, D., and Hou, T. The impact of cross-docked poses on performance of machine learning classifier for protein–ligand binding pose prediction. *Journal of cheminformatics*, 13(1):81, 2021.

Wang, T. and Isola, P. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pp. 9929–9939. PMLR, 2020.

Zhang, Y., Cai, H., Shi, C., Zhong, B., and Tang, J. E3bind: An end-to-end equivariant network for protein-ligand docking. *arXiv preprint arXiv:2210.06069*, 2022.

Zhao, D., Li, W., Shen, Z., Qiu, Y., Xu, B., Chen, H., and Chen, Y. Bias is a subspace, not a coordinate: A geometric rethinking of post-hoc debiasing in vision-language models. *arXiv preprint arXiv:2511.18123*, 2025.

Zhong, H., Song, W., Han, T., Pagnucco, M., Xue, J., and Song, Y. Fairt2v: Training-free debiasing for text-to-video diffusion models. *arXiv preprint arXiv:2601.20791*, 2026.

A. Limitations

This work is an initial study of validity-aware post-hoc debiasing for manifold-valued representations. IsoAct assumes that the target geometry is known and that suitable action primitives can be specified for that geometry. The experiments cover three representative cases—the hypersphere, the Lorentz hyperboloid, and $SE(3)$—but they do not establish that the same behavior will hold for all manifolds, learned latent geometries, product manifolds, approximate constraints, or application domains. Extending the framework to settings where the manifold structure is learned, only partially known, or not exactly satisfied by the representation remains future work.

IsoAct also depends on how nuisance information is estimated. The nuisance axes are obtained from Euclidean probe features using labeled training data, so nonlinear, interaction-based, weakly labeled, or distribution- shifted nuisance information may not be fully captured. Multiple nuisance attributes may interact with each other or with the utility variable in ways that are not represented by a small set of linear probe axes. The neutral reference, the number of retained axes, the action strength, and the fixed action mode are selected using validation data, and different choices can produce different nuisance–utility trade-offs. The selected configuration should therefore be interpreted as one operating point, not as a universally optimal edit.

The validity guarantee is also limited in scope. IsoAct preserves the manifold validity of each edited representation by construction, but the sample-dependent transformation is generally not a single global isometry and does not preserve all pairwise manifold distances. It also does not guarantee complete concept removal, fairness, causal invariance, or semantic correctness. A downstream model may still recover nuisance information through nonlinear probes, unmeasured features, or correlations not targeted by the selected action family.

The empirical evaluation is necessarily limited. The controlled experiments are mechanism checks under known nuisance structure, not evidence of universal behavior in semantic tasks. In the KLIFS experiment, kinase group is used as a coarse nuisance proxy and the utility tasks measure selected structural properties. Preserving $SE(3)$ validity of ligand poses does not by itself guarantee biochemical plausibility, physical feasibility, binding-affinity preservation, or safety in downstream drug-discovery workflows. Future work should study richer nuisance definitions, nonlinear nuisance estimators, adaptive action policies, broader manifolds and datasets, and more comprehensive downstream audits.

B. Additional method details

This appendix expands the compact method description in Section 4. It includes the algorithmic steps, nuisance-axis estimation, neutral-reference construction, and the geometry-specific updates used by IsoAct. The final method uses a fixed action mode for a given run.

B.1. Nuisance-axis estimation

In the main experiments, IsoAct uses Iterative Nullspace Projection (INLP) to estimate nuisance axes (Ravfogel et al., 2020). The estimator operates on Euclidean probe features $v_i = \phi(z_i)$. Binary nuisance variables are treated as the two-class case. For multiclass nuisance variables, we use one-vs-rest linear nuisance probes and remove the subspace spanned by the corresponding classifier weight vectors after orthonormalization.

Concretely, INLP repeatedly fits a linear nuisance probe on the current residual features, extracts the nuisance directions associated with the probe weights, orthonormalizes the newly extracted directions against the previously retained axes, and projects the features onto the orthogonal complement of the retained nuisance subspace. Axes are retained in discovery order until K normalized directions have been collected. The discovery order is also the order used for the sequential IsoAct updates.

B.2. Geometry-specific action constructions

IsoAct uses manifold-native action primitives rather than Euclidean coordinate edits. For a fixed action strength, the primitive is an isometry of the corresponding geometry. Since the strength can depend on the sample through Eq. (1) or Eq. (4), the overall map need not be a single global isometry, but each update remains valid on the target manifold.

Spherical actions. For the spherical synthetic benchmark, IsoAct applies an orthogonal rotation in the selected action plane. If O_{ik} denotes the resulting orthogonal update, then

$$z_i^{(k)} = O_{ik} z_i^{(k-1)}, \qquad O_{ik}^\top O_{ik} = I.$$

The rotation angle is controlled by $\alpha \delta_{ik}^{\mathrm{score}}$. Orthogonality gives $\|z_i^{(k)}\|_2 = \|z_i^{(k-1)}\|_2$, so the unit-norm constraint is preserved.

Lorentz hyperbolic actions. For Lorentz hyperbolic representations, IsoAct applies a Lorentz transformation with strength controlled by $\alpha \delta_{ik}^{\mathrm{score}}$. Let $J = \mathrm{diag}(-1, 1, \ldots, 1)$. The update has the form

$$h_i^{(k)} = L_{ik} h_i^{(k-1)}, \qquad L_{ik}^\top J L_{ik} = J,$$

with L_{ik} chosen in the proper orthochronous Lorentz group. Thus the Lorentz quadratic constraint and the upper-sheet condition are preserved. In implementation, spatial Lorentz rotations are used for angular components, and Lorentz boosts are used for radial-changing components.

B.3. Detailed SE(3) construction

For $\mathrm{SE}(3)$, the probe feature exposes the pose blocks, $\phi(R, t) = [\mathrm{vec}(R), t] \in \mathbb{R}^{12}$. We decompose a nuisance axis as $u_k = (u_{R,k}, u_{t,k})$, where $u_{R,k} \in \mathbb{R}^9$ corresponds to the vectorized rotation block and $u_{t,k} \in \mathbb{R}^3$ corresponds to translation. The rotation block is reshaped into $U_{R,k} \in \mathbb{R}^{3\times 3}$ using the same convention as $\mathrm{vec}(R)$, and its skew-symmetric part is

$$\Omega_k = \frac{1}{2}(U_{R,k} - U_{R,k}^\top).$$

The corresponding unit rotation axis is

$$a_k = \frac{\Omega_k^\vee}{\|\Omega_k^\vee\|_2} \quad \text{when } \|\Omega_k^\vee\|_2 > 0.$$

Here $(\cdot)^\vee$ maps a skew-symmetric matrix in $\mathfrak{so}(3)$ to its vector form. The matrix Ω_k is not an edited rotation matrix; it is the rotation generator extracted from the nuisance axis. If $\|\Omega_k^\vee\|_2 = 0$, the rotation branch is the identity for that axis. Similarly, the translation direction is

$$b_k = \frac{u_{t,k}}{\|u_{t,k}\|_2} \quad \text{when } \|u_{t,k}\|_2 > 0,$$

with the translation branch set to the identity when $\|u_{t,k}\|_2 = 0$.

Rotation branch. The rotation branch uses the native rotation discrepancy

$$\delta_{ik}^R = \left\langle \left(\mathrm{Log}_{\mathrm{SO}(3)} \left((R_i^{(k-1)})^\top R_{\mathrm{ref}} \right) \right)^\vee, a_k \right\rangle. \tag{4}$$

The relative rotation $(R_i^{(k-1)})^\top R_{\mathrm{ref}}$ moves the current orientation toward the reference orientation; the logarithm maps this relative rotation to a tangent vector, and the inner product extracts its component along the nuisance-derived axis a_k. The update is

$$Q_{ik} = \mathrm{Exp}_{\mathrm{SO}(3)} \left(\alpha \delta_{ik}^R \widehat{a}_k \right), \qquad R_i^{(k)} = R_i^{(k-1)} Q_{ik}. \tag{5}$$

Since $Q_{ik} \in \mathrm{SO}(3)$, $R_i^{(k)} \in \mathrm{SO}(3)$ whenever $R_i^{(k-1)} \in \mathrm{SO}(3)$.

Translation branch. The translation branch uses the probe-space discrepancy in Eq. (1) and updates

$$t_i^{(k)} = t_i^{(k-1)} + \alpha \delta_{ik}^{\mathrm{score}} b_k. \tag{6}$$

This keeps the translation component in $\mathbb{R}^3$.

Table 4. Fixed $\mathrm{SE}(3)$ action modes used by IsoAct.

Mode	Rotation branch	Translation branch
Rotation-only IsoAct	Eq. (5)	identity
Translation-only IsoAct	identity	Eq. (6)
Joint rotation–translation IsoAct	Eq. (5)	Eq. (6)

Fixed $\mathrm{SE}(3)$ action modes. IsoAct uses three fixed $\mathrm{SE}(3)$ modes. Rotation-only updates only R, translation-only updates only t, and joint rotation–translation updates both. Equivalently, the joint update can be written in homogeneous form as

$$\begin{bmatrix} R_i^{(k-1)} & t_i^{(k-1)} \\ 0 & 1 \end{bmatrix} \mapsto \begin{bmatrix} R_i^{(k-1)} Q_{ik} & t_i^{(k-1)} + \alpha \delta_{ik}^{\mathrm{score}} b_k \\ 0 & 1 \end{bmatrix}.$$

Because rotations are non-commutative, changing the order of rotation updates can change the final pose. We therefore apply axes in INLP discovery order.

B.4. Scope and implementation conventions

The final IsoAct implementation uses one fixed action mode from Table 4 for a given run. The action strength is controlled by the discrepancy and the validation-selected global scale α. The notation in the main paper therefore uses K, α, $\delta_{ik}^{\mathrm{score}}$, and δ_{ik}^{R} without introducing an additional generic action parameter.

The method should be interpreted as a post-hoc framework for manifold-constrained representations using manifold-native action primitives. Because the action strength can vary by sample, the composed map is generally not a single global isometry and should not be interpreted as preserving all pairwise manifold distances. The claim used by IsoAct is that each individual update is constructed to remain valid on the target manifold.

B.5. Proof of Theorem

Let Φ be the mapping of $\mathrm{SE}(3)$ action, that is, IsoAct on $\mathrm{SE}(3)$. Then Φ maps a matrix in $\mathrm{SE}(3)$ to a matrix with a rotation-only IsoAct, to a matrix with a translation-only IsoAct, or to a matrix with a joint rotation–translation IsoAct. Also Φ represents any composition of the 3 mappings (rotation-only IsoAct, translation-only IsoAct, joint rotation–translation IsoAct).

For $z_i \in \mathrm{SE}(3)$, $z_i = (R_i, t_i)$ with $R_i \in \mathrm{SO}(3)$ and $t_i \in \mathbb{R}^3$. Let Φ_R, Φ_T be the mappings of a rotation-only IsoAct with direction $a_k \in \mathbb{R}^3$ and of a translation-only IsoAct with direction $b_k \in \mathbb{R}^3$, respectively. Then

$$\begin{aligned} \Phi_R(z_i) &= \Phi_R(R_i, t_i) = (R_i Q_{ik}, t_i) \\ \Phi_T(z_i) &= \Phi_T(R_i, t_i) = (R_i, t_i + \alpha b_k), \end{aligned} \tag{7}$$

where $Q_{ik} = \mathrm{Exp}_{\mathrm{SO}(3)}(\alpha \widehat{a}_k)$, for some scalar α, and $\widehat{a}_k$ is the skew-symmetric matrix associated with a_k, that is, for $a = (a_x, a_y, a_z) \in \mathbb{R}^3$,

$$\widehat{a} = \begin{bmatrix} 0 & -a_z & a_y \\ a_z & 0 & -a_x \\ -a_y & a_x & 0 \end{bmatrix}.$$

First, we will prove that Φ_R and Φ_T map $\mathrm{SE}(3)$ into $\mathrm{SE}(3)$.

Since for $\omega \in \mathbb{R}^3$,

$$\mathrm{Exp}_{\mathrm{SO}(3)}(\widehat{\omega}) = I + \left(\frac{\sin|\omega|}{|\omega|}\right)\widehat{\omega} + \left(\frac{1-\cos|\omega|}{|\omega|^2}\right)(\widehat{\omega})^2 := \mathrm{R} \in \mathrm{SO}(3),$$

we have, in Eq. (7), $Q_{ik} = \mathrm{Exp}_{\mathrm{SO}(3)}(\alpha \widehat{a}_k) \in \mathrm{SO}(3)$.

And since the composition of two rotation matrices in $\mathrm{SO}(3)$ is in $\mathrm{SO}(3)$, we have $R_i Q_{ik} \in \mathrm{SO}(3)$. Hence, $\Phi_R(z_i) \in \mathrm{SE}(3)$.

Since $t_i \in \mathbb{R}^3$ and $t_i + \alpha b_k \in \mathbb{R}^3$, we have $\Phi_T(z_i) \in \mathrm{SE}(3)$.

Now we will prove that Φ maps $\mathrm{SE}(3)$ into $\mathrm{SE}(3)$.

Since Φ_R and Φ_T are a rotation-only IsoAct and a translation-only IsoAct mappings, respectively, for joint rotation–translation IsoAct mapping Φ, it is the composition mapping of rotation-only and translation-only IsoActs. For one rotation-only and one translation-only IsoActs, Φ has the following form

$$\Phi(z) = \Phi_R \cdot \Phi_T(z) \;\text{ or }\; \Phi(z) = \Phi_T \cdot \Phi_R(z).$$

Since for $z \in \mathrm{SE}(3)$, $\Phi_T(z) \in \mathrm{SE}(3)$ and $\Phi_R(z) \in \mathrm{SE}(3)$, the composition of $\mathrm{SE}(3)$ matrices $\Phi_R(z)$ and $\Phi_T(z)$ are also in $\mathrm{SE}(3)$. Hence Φ maps $\mathrm{SE}(3)$ into $\mathrm{SE}(3)$ for a joint rotation–translation IsoAct mapping.

This also holds for any composition of the 3 maps (rotation-only IsoAct, translation-only IsoAct, joint rotation–translation IsoAct), for the compositions of any matrices in $\mathrm{SE}(3)$ are also in $\mathrm{SE}(3)$.

Therefore IsoAct on $\mathrm{SE}(3)$ maps into $\mathrm{SE}(3)$.

C. Full Experimental Tables

C.1. Controlled Experiments

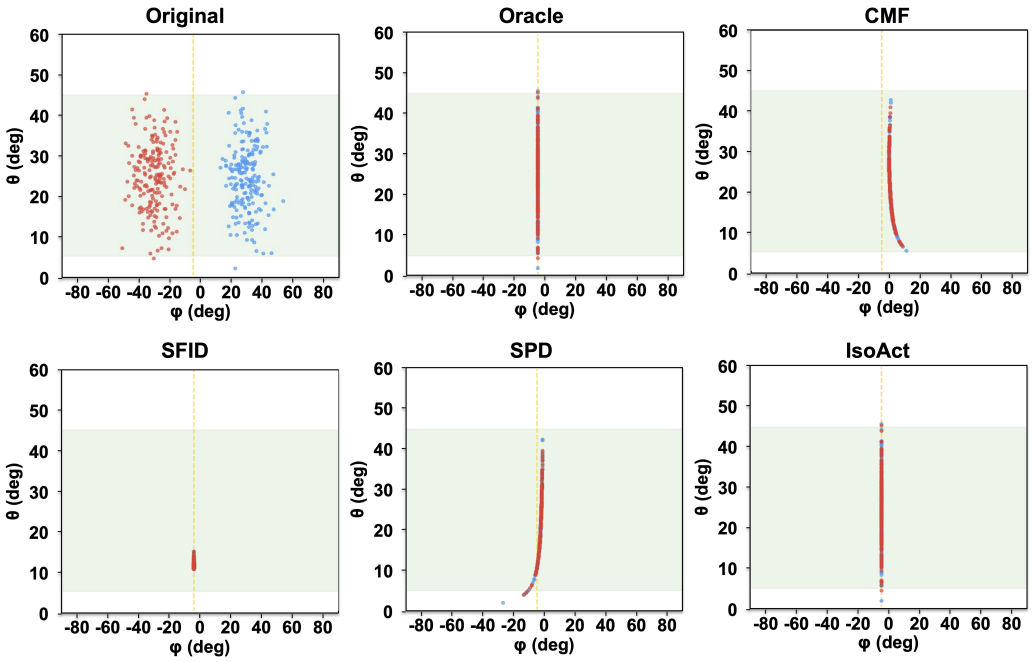

Figure 5. Hyperspherical controlled experiment scatter plot. The horizontal axis φ is the injected nuisance angular coordinate, and the vertical axis θ is the task-relevant angular coordinate. The dashed vertical line marks the neutral nuisance reference, and the shaded band marks the target range for the task-relevant variable. An ideal edit removes horizontal separability while preserving the vertical structure.

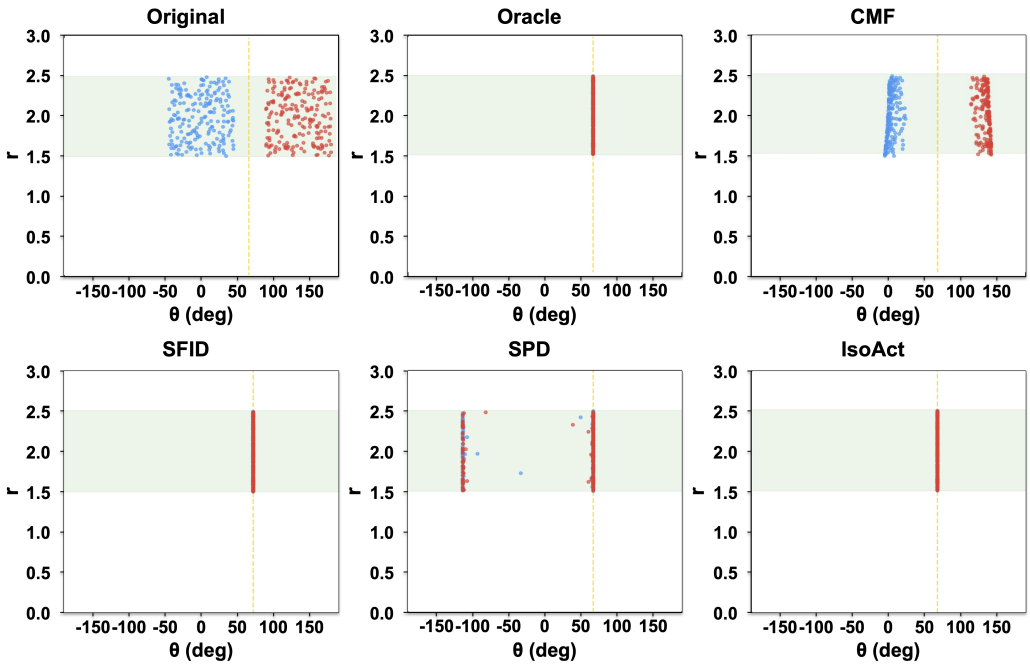

Figure 6. Lorentz hyperbolic controlled experiment scatter plot. The horizontal axis θ is the injected nuisance angular coordinate, and the vertical axis r is the task-relevant radial coordinate. The dashed vertical line marks the neutral nuisance reference, and the shaded band marks the target range for the task-relevant variable. An ideal edit removes horizontal separability while preserving the vertical structure.

C.2. Real-world SE(3) ligand-pose experiment on KLIFS

Table 5. KLIFS SE(3) ligand-pose debiasing results with uncertainty estimates. Uncertainty is estimated using fixed-test bootstrap resampling with 1,000 resamples of the frozen test split, with deterministic debiasing transforms held fixed.

		Debiasing Capability				Utility Preservation								Manifold Constraint Violation	
						Task 1: αC_helix state Classification				Task 2: pocket occupancy Classification					
	Method	Acc↓	Recall↓	F1↓	AUROC↓	Acc↑	Recall↑	F1↑	AUROC↑	Acc↑	Recall↑	F1↑	AUROC↑	Constraint 1 Mean↓	Constraint 2 Mean↓
KLIFS	Null bias	0.0827 ±0.0000	0.1172 ±0.0000	0.0679 ±0.0000	0.4996 ±0.0000	—	—	—	—	—	—	—	—	—	—
	Original	0.1395 ±0.0068	0.2427 ±0.0113	0.1404 ±0.0069	0.6158 ±0.0069	0.8532 ±0.0069	0.3374 ±0.0259	0.3846 ±0.0256	0.7169 ±0.0151	0.9248 ±0.0029	0.7357 ±0.0214	0.7921 ±0.0169	0.9708 ±0.0019	0.0000	0.0000
	CMF†	0.1438 ±0.0069	0.2802 ±0.0111	0.1432 ±0.0067	**0.5780 ±0.0082**	0.8393 ±0.0074	0.1266 ±0.0178	0.1765 ±0.0232	0.5623 ±0.0165	0.8985 ±0.0034	0.6724 ±0.0224	0.7365 ±0.0174	0.9347 ±0.0032	0.0072 ±0.0001	0.0048 ±0.0000
	SFID	0.1395 ±0.0068	0.2427 ±0.0113	0.1404 ±0.0069	0.6158 ±0.0069	0.8509 ±0.0070	0.3263 ±0.0255	0.3730 ±0.0255	0.7184 ±0.0150	0.9268 ±0.0029	0.7388 ±0.0225	0.7860 ±0.0176	0.9712 ±0.0019	**0.0000**	**0.0000**
	SPD	0.1383 ±0.0068	0.2408 ±0.0114	0.1392 ±0.0069	0.6151 ±0.0070	0.8577 ±0.0068	0.3490 ±0.0258	0.4002 ±0.0254	0.7162 ±0.0148	**0.9271 ±0.0029**	0.7534 ±0.0208	0.7820 ±0.0172	**0.9724 ±0.0019**	0.0037 ±0.0000	0.0012 ±0.0000
	IsoAct	**0.1156 ±0.0060**	**0.1853 ±0.0100**	**0.1037 ±0.0057**	0.5853 ±0.0071	**0.8609 ±0.0069**	**0.3752 ±0.0265**	**0.4231 ±0.0258**	**0.7564 ±0.0150**	0.9228 ±0.0030	**0.8052 ±0.0127**	**0.7998 ±0.0145**	0.9649 ±0.0022	**0.0000**	**0.0000**

Table 6. Real-world experiments with full metrics. The table reports debiasing capability, two downstream utility tasks, and manifold constraint violations on SE(3) representations.

		Debiasing Capability				Utility Preservation								Manifold Constraint Violation			
						Task 1: αC_helix state Classification				Task 2: pocket occupancy classification				SO(3) Orthogonality Error		SO(3) Determinant Error	
	Method	Acc↓	Recall↓	F1↓	AUROC↓	Acc↑	Recall↑	F1↑	AUROC↑	Acc↑	Recall↑	F1↑	AUROC↑	Mean↓	Max↓	Mean↓	Max↓
Reference	Null bias	0.0827 ±0.0000	0.1172 ±0.0000	0.0679 ±0.0000	0.4996 ±0.0000	—	—	—	—	—	—	—	—	—	—	—	—
	Original	0.1395 ±0.0068	0.2427 ±0.0113	0.1404 ±0.0069	0.6158 ±0.0069	0.8532 ±0.0069	0.3374 ±0.0259	0.3846 ±0.0256	0.7169 ±0.0151	0.9248 ±0.0029	0.7357 ±0.0214	0.7921 ±0.0169	0.9708 ±0.0019	0.0000	0.0000	0.0000	0.0000
Learning-based	CMF	0.1438 ±0.0069	0.2802 ±0.0111	0.1432 ±0.0067	**0.5780 ±0.0082**	0.8393 ±0.0074	0.1266 ±0.0178	0.1765 ±0.0232	0.5623 ±0.0165	0.8985 ±0.0034	0.6724 ±0.0224	0.7365 ±0.0174	0.9347 ±0.0032	0.0072 ±0.0001	0.0147 ±0.0000	0.0048 ±0.0000	0.0100 ±0.0000
	CMF + Proj	0.1491 ±0.0072	0.2061 ±0.0101	0.1160 ±0.0064	0.6491 ±0.0066	0.8519 ±0.0069	0.1328 ±0.0175	0.1960 ±0.0236	0.6535 ±0.0156	0.8823 ±0.0035	0.6878 ±0.0142	0.7070 ±0.0149	0.9271 ±0.0035	**0.0000**	**0.0000**	**0.0000**	**0.0000**
Post-hoc	SFID	0.1395 ±0.0068	0.2427 ±0.0113	0.1404 ±0.0069	0.6158 ±0.0069	0.8509 ±0.0070	0.3263 ±0.0255	0.3730 ±0.0255	0.7184 ±0.0150	0.9268 ±0.0029	0.7388 ±0.0225	0.7860 ±0.0176	0.9712 ±0.0019	**0.0000**	**0.0000**	**0.0000**	**0.0000**
	SFID + Proj	0.1395 ±0.0068	0.2427 ±0.0113	0.1404 ±0.0069	0.6158 ±0.0069	0.8509 ±0.0070	0.3263 ±0.0255	0.3730 ±0.0255	0.7184 ±0.0150	0.9268 ±0.0029	0.7388 ±0.0225	0.7860 ±0.0176	0.9712 ±0.0019	**0.0000**	**0.0000**	**0.0000**	**0.0000**
	SPD	0.1383 ±0.0068	0.2408 ±0.0114	0.1392 ±0.0069	0.6151 ±0.0070	0.8577 ±0.0068	0.3490 ±0.0258	0.4002 ±0.0254	0.7162 ±0.0148	0.9271 ±0.0029	0.7534 ±0.0208	0.7820 ±0.0172	**0.9724 ±0.0019**	0.0037 ±0.0000	0.0143 ±0.0004	0.0012 ±0.0000	0.0065 ±0.0005
	SPD + Proj	0.1315 ±0.0065	0.2350 ±0.0111	0.1348 ±0.0066	0.6123 ±0.0074	0.8506 ±0.0070	0.3320 ±0.0260	0.3766 ±0.0255	0.7250 ±0.0142	**0.9285 ±0.0028**	0.7664 ±0.0202	**0.8142 ±0.0156**	0.9703 ±0.0023	**0.0000**	**0.0000**	**0.0000**	**0.0000**
	Geodesic	0.1352 ±0.0067	0.2308 ±0.0112	0.1338 ±0.0067	0.6232 ±0.0064	0.8525 ±0.0069	0.3401 ±0.0258	0.3853 ±0.0255	0.7377 ±0.0146	0.9248 ±0.0029	0.7857 ±0.0174	0.8034 ±0.0151	0.9704 ±0.0020	**0.0000**	**0.0000**	**0.0000**	**0.0000**
	IsoAct	**0.1156 ±0.0060**	**0.1853 ±0.0100**	**0.1037 ±0.0057**	0.5853 ±0.0071	**0.8609 ±0.0069**	**0.3752 ±0.0265**	**0.4231 ±0.0258**	**0.7564 ±0.0150**	0.9228 ±0.0030	**0.8052 ±0.0127**	0.7998 ±0.0145	0.9649 ±0.0022	**0.0000**	**0.0000**	**0.0000**	**0.0000**

D. Compute Resources

The reported experiments do not require GPU computation, and the finalized result-regeneration scripts were run on an Ubuntu CPU node with an Intel Xeon Gold 6326 CPU at 2.90GHz, 64 CPU threads, and 251 GiB RAM. For the synthetic experiments, regenerating the finalized $\mathrm{SE}(3)$, spherical, and hyperbolic tables and figures took 22.1s, 15.0s, and 14.7s, respectively, with peak memory below 1 GB, excluding earlier exploratory development runs.