Self-Supervised Disentanglement via Cluster-Dependent Subspace Rotational Equivariance

Junbing Liu, Zhiming Zhou
Shanghai University of Finance and Economics
heiju@stu.sufe.edu.cn, zhouzhiming@mail.shufe.edu.cn

Abstract

Conventional self-supervised methods learn well-clustered representations by enforcing invariance to data augmentations, but offer limited control over how factors of variation are organized in feature space. Recent methods seek to impose greater structure on the feature space by learning representations that are equivariant to data augmentations. However, existing approaches face two limitations: (i) the invariance objective can interfere with equivariance learning; and (ii) global rotational equivariance can constrain representations from different semantic clusters to a parallel configuration. To address these limitations, we introduce **C**luster-dependent subspace **R**otational **E**quivariance for **D**isentanglement (CRED), which decouples invariance and equivariance learning through two-stage training. The first stage uses invariance-based training for semantic clustering, and the second learns cluster-dependent rotational equivariance in mutually orthogonal two-dimensional subspaces. We establish conditions under which this construction yields exact disentanglement of augmentation-induced variations. Empirically, CRED achieves near-perfect scores (≥ 0.99) on all three DCI metrics after post-processing, using random forest evaluation across four synthetic datasets. On three real-world datasets, it outperforms all evaluated baselines on all four equivariance and disentanglement metrics.

1 Introduction

Learning disentangled representations, in which factors of variation in the data are encoded in distinct components of the feature space, is a fundamental goal of representation learning (Bengio et al., 2013; Hinton & Salakhutdinov, 2006; Higgins et al., 2018; Locatello et al., 2019; Weiler & Cesa, 2019). Such representations can improve interpretability (Chen et al., 2016), generalization (Lachapelle et al., 2023), and sample efficiency (van Steenkiste et al., 2019), while facilitating controllable generation (Ren et al., 2021) and transfer learning (Xiao et al., 2021).

Traditional approaches to disentangled representation learning include fully and weakly supervised methods, which rely on factor labels or auxiliary supervisory signals (Reed et al., 2014; Cheung et al., 2014; Shu et al., 2020; Locatello et al., 2020), and unsupervised latent-variable methods that encourage statistical independence among latent variables (Higgins et al., 2017; Kim & Mnih, 2018; Chen et al., 2018). The former depend on external supervision that can be costly to obtain at scale, whereas the latter favor statistically independent latent variables, a bias that may be unsuitable when the factors of interest are dependent.

Data augmentations provide a source of self-supervision by relating transformed views of the same observation, without requiring factor labels. Conventional self-supervised methods (Chen et al., 2020a; He et al., 2020; Tian et al., 2020; Wu et al., 2018) enforce invariance to these transformations, producing well-clustered representations but offering limited control over how factors of variation are organized in the feature space.

Recent self-supervised methods have sought to address this limitation along two main lines. The first seeks to capture distinct factors of variation by tailoring contrastive objectives to different data partitions (Wang et al., 2021) or by using augmentation-specific representation branches with different invariance constraints

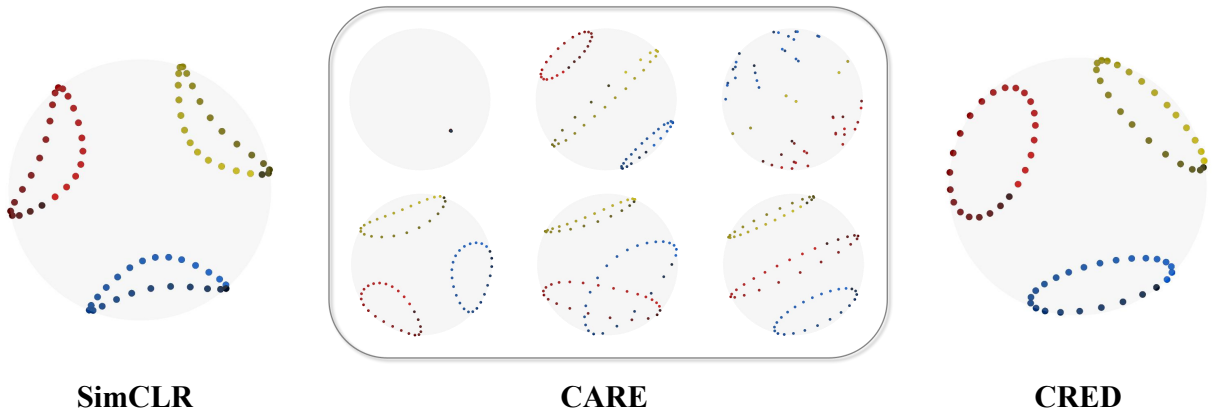


Figure 1: Feature trajectories traced out as each of three input images rotates. Each color corresponds to one input image. **SimCLR** (Chen et al., 2020a) forms uniformly distributed clusters with irregular trajectories. **CARE** (Gupta et al., 2024) is sensitive to the relative weights of its invariance, uniformity, and equivariance terms: the shown settings produce collapse, irregular trajectories, or circular trajectories in parallel planes with unequal radii. **CRED** (ours) forms uniformly distributed clusters with circular trajectories in planes whose orientations vary across clusters. Appendix B.1 gives the setup and results for a non-cyclic setting.

(Xiao et al., 2021; Eastwood et al., 2023). These methods encourage factor separation through invariance constraints, without explicitly modeling how features transform under augmentations. The second turns to equivariant representations, in which input transformations induce corresponding transformations in feature space (Dangovski et al., 2022; Devillers & Lefort, 2023).

However, some existing equivariance-based disentanglement approaches focus on single-cluster manifolds (Shakerinava et al., 2022), while others require cluster-related invariant information and variation-related equivariant information to be encoded in separate components of the learned representation (Garrido et al., 2023b; Marchetti et al., 2023). To address these limitations, we introduce **C**luster-dependent subspace **R**otational **E**quivariance for **D**isentanglement (CRED), a framework for learning multi-cluster equivariant representations on a single hypersphere. Within the broader literature on equivariant representation learning, CRED is most closely related to CARE (Gupta et al., 2024), which shares this representation setting but does not target disentanglement.

CRED departs from CARE in three substantive respects that collectively enable disentanglement. First, CARE jointly optimizes invariance and equivariance objectives for the same augmentations on a shared representation. The invariance objective pulls the embeddings of augmented views of each sample together on the hypersphere, competing with the nontrivial feature variation needed for equivariance-based disentanglement. Figure 1 illustrates the geometric challenge: SimCLR and some CARE loss-weight settings produce irregular, noncircular feature trajectories. CRED eliminates this interference through a two-stage training scheme that first uses an InfoNCE-based invariance objective (van den Oord et al., 2018) to learn representations for semantic clustering and then dedicates the second stage exclusively to equivariance learning.

Second, CARE (Gupta et al., 2024) requires each augmentation instance to induce the same orthogonal transformation across all samples. Under the conditions established in Appendix D.1, this shared action confines cluster representations to parallel affine subspaces, as illustrated in Figure 1. CRED uses cluster-dependent rotational coordinates, allowing subspace orientations to vary across clusters.

Third, CARE learns rotational equivariance implicitly through a similarity-preserving objective, without explicitly parameterizing a subspace decomposition for augmentation-induced variations. Discovering mutually orthogonal subspaces for distinct variation components is therefore left to optimization and can be challenging in practice. CRED instead models augmentation-induced feature transformations explicitly via the action of a direct product of SO(2) groups on mutually orthogonal two-dimensional subspaces. Each

group acts only within its corresponding subspace, allowing distinct variation components to be encoded through separate rotation angles.

Building on this structure, we establish conditions under which the proposed model achieves exact disentanglement with respect to orthogonal augmentations. We further establish strong disentanglement by preventing angular wrap-around: within each cluster, each non-cyclic variation component admits a single-valued, strictly monotonic angular readout over its full range (Proposition 3). We also identify sufficient conditions under which the distinct variation components induced by a single augmentation are disentangled across their corresponding representation subspaces. Unlike the group-theoretic formulation of Higgins et al. (2018), our theoretical analysis does not assume that the underlying data transformations form a group.

CRED achieves near-perfect scores (≥ 0.99) on all three DCI metrics (Eastwood & Williams, 2018) after post-processing, using random forest evaluation across four synthetic datasets with ground-truth factors of variation. For real-world datasets without such annotations, we combine equivariance measures with two new disentanglement metrics, Mean Overlap and Mean Domination, to assess representation structure beyond conventional linear probes. On three real-world datasets, CRED outperforms all evaluated baselines on all four equivariance and disentanglement metrics.

Our main contributions are as follows:

- We develop a two-stage training framework that decouples invariance and equivariance learning, eliminating interference from the invariance objective.
- We identify the geometric constraint behind parallel cluster configurations under global rotational equivariance and remove the requirement of a common orientation through cluster-dependent rotations.
- We introduce an explicit subspace rotational model that provides direct control over how augmentation-induced variations are organized in feature space.
- We prove that CRED achieves strong disentanglement of augmentation-induced variations under sufficient conditions, including cases where a single augmentation induces multiple variation components.
- We introduce two metrics, Mean Overlap and Mean Domination, to assess the disentanglement of augmentation-induced variations on real-world datasets beyond conventional linear probes.

2 Related Work

Group-equivariant architectures enforce equivariance through architectural constraints tied to a specified group action (Cohen & Welling, 2016; 2017; Weiler & Cesa, 2019). Among methods that learn equivariance through training objectives, EquiMod (Devillers & Lefort, 2023) uses an augmentation-conditioned predictor, SIE (Garrido et al., 2023b) uses a hypernetwork to generate augmentation-conditioned linear maps, and CARE (Gupta et al., 2024) encourages a shared orthogonal action through similarity preservation. CRED follows this objective-based approach, but constrains feature transformations to rotations in prescribed, mutually orthogonal two-dimensional subspaces, with a separate coordinate system for each cluster. This gives direct control over which subspaces an augmentation can affect while allowing their orientations to vary across clusters.

Disentanglement has also been studied in generative models, including VAEs (Higgins et al., 2017; Chen et al., 2018), GANs (Chen et al., 2016), and diffusion models (Yang et al., 2023; Wu & Zheng, 2024). Some generative approaches use grouped observations with shared content (Bouchacourt et al., 2018) or paired observations sharing some factors (Locatello et al., 2020) as weak supervision to guide factor separation. A more comprehensive discussion is provided in Appendix C.

3 Preliminaries

We consider observations $x = \varphi(z)$ generated from latent semantic factors $z = (z_1, \ldots, z_{d_z}) \in \mathcal{Z} \subseteq \mathbb{R}^{d_z}$ through an unknown invertible nonlinear map $\varphi : \mathcal{Z} \to \mathcal{X}$. A **feature extractor** $f : \mathcal{X} \to \mathcal{Y}$ maps each observation to the unit hypersphere $\mathcal{Y} = \mathbb{S}^{d-1} \triangleq \{y \in \mathbb{R}^d : \|y\|_2 = 1\}$. The resulting representation map is $\Phi = f \circ \varphi : \mathcal{Z} \to \mathcal{Y}$.

For self-supervised training, we consider a set of n augmentation families $\mathcal{A} = \{a_i\}_{i=1}^n$, each parameterized by t drawn from $\mathcal{T}_{a_i}$, applied individually or in composition. With a slight abuse of notation, $a \in \mathcal{A}$ also denotes an individual augmentation instance or a composition of such instances; we write $a(t, x)$ when its parameters are explicit and $a(x)$ otherwise. Given $x = \varphi(z)$, a positive sample $x^+ = a(x)$ is constructed, and we aim to relate $f(x)$ and $f(x^+)$ through appropriate objectives.

With respect to the chosen augmentation set $\mathcal{A}$, **content** refers to properties preserved by all augmentations, while **style** refers to properties varied by at least one augmentation. These properties may depend on combinations of the original generative factors. Invariance learning aims to retain content information while suppressing augmentation-induced style variation. Equivariance learning models the corresponding changes in the representation. Our disentanglement guarantees concern augmentation-induced variations; properties preserved by all chosen augmentations are not guaranteed to be separately encoded.

Augmentations in Invariance. Invariance learning seeks $f(a(x)) \approx f(x)$ for all $a \in \mathcal{A}$. In practice, this is encouraged by aligning two augmented views, $f(a_1(x)) \approx f(a_2(x))$, where $a_1, a_2 \in \mathcal{A}$. A widely used objective is InfoNCE (van den Oord et al., 2018; Tian et al., 2020; He et al., 2020; Wu et al., 2018; Chen et al., 2020a), written in SimCLR form for a batch of N observations with two augmented views each:

$$\mathcal{L}_{\text{InfoNCE}} = \tfrac{1}{2N} \sum_{i=1}^{2N} -\log \frac{\exp(f(x_i)^\top f(x_i^+)/\tau)}{\sum_{j=1, j\neq i}^{2N} \exp(f(x_i)^\top f(x_j)/\tau)}, \tag{1}$$

where $x_1, \ldots, x_{2N}$ are the augmented views, x_i^+ is the other view of the same observation as x_i, and $\tau > 0$ is the temperature. InfoNCE encourages positive-pair alignment and feature uniformity on the hypersphere (Wang & Isola, 2020). Under suitable assumptions on augmentation structure and the encoder class, optimal InfoNCE representations preserve the underlying data clusters (Wang et al., 2022; Parulekar et al., 2023).

Augmentations in Equivariance. Equivariance instead requires that input transformations induce predictable feature-space changes: $f(a(x)) \approx M_a(f(x))$, with the natural objective:

$$\mathcal{L}_{\text{equiv}_0} = \mathbb{E}_{x,a}\left[\|f(a(x)) - M_a(f(x))\|^2\right]. \tag{2}$$

EquiMod (Devillers & Lefort, 2023) and SIE (Garrido et al., 2023b) explicitly model the feature transformation M_a. Geometric structure can also be imposed through objectives based on group invariants (Shakerinava et al., 2022). CARE (Gupta et al., 2024), for example, encourages orthogonal equivariance by preserving pairwise inner products under a shared augmentation.

Augmentations in Disentanglement. In the group-theoretic formulation of Higgins et al. (2018), let $\mathcal{G} = \mathcal{G}_1 \times \cdots \times \mathcal{G}_r$ act on $\mathcal{X}$. Viewing $\mathcal{Y} = \mathbb{S}^{d-1}$ as a subset of $\mathbb{R}^d$, f is *linearly disentangled* with respect to this group decomposition if there exist a decomposition $\mathbb{R}^d = V_1 \oplus \cdots \oplus V_r$ and group representations $\rho_i : \mathcal{G}_i \to \mathrm{GL}(V_i)$ satisfying, for each $i = 1, \ldots, r$,

$$f(\mathbf{g}_i \cdot x) = \Big(\mathrm{id}_{V_1} \oplus \cdots \oplus \rho_i(\mathbf{g}_i) \oplus \cdots \oplus \mathrm{id}_{V_r}\Big) f(x), \quad \forall\, x \in \mathcal{X},\ \mathbf{g}_i \in \mathcal{G}_i. \tag{3}$$

Here id_{V_j} is the identity map on V_j, so each $\mathcal{G}_i$ acts only on its corresponding feature subspace. For our augmentation-based analysis, we describe the separation of input variations directly through augmentation families, without requiring the input transformations to form a group. We call augmentation families **mutually orthogonal** when they independently control distinct sets of variation components in the data, with each family leaving the components controlled by the others unchanged. Each family may affect multiple components, which need not coincide with the original generative factors. This setting is related to the active decomposition of Shakerinava et al. (2022), which uses transformations acting on individual group factors.

4 Cluster-Dependent Subspace Rotational Equivariance

CRED decouples invariance and equivariance learning through two-stage training. Stage 1 uses InfoNCE to learn representations for semantic clustering; Stage 2 learns cluster-dependent rotational equivariance without the InfoNCE objective. Algorithm 1 summarizes training and optional post-processing.

Global rotational equivariance can constrain different clusters to affine subspaces with a common orientation. Specifically, when the shared rotation operators induce changes spanning a nonzero proper subspace V, exact equivariance and connectivity through within-cluster augmentation pairs place each cluster in an affine translate of V (Appendix D.1). CRED assigns each cluster its own rotational coordinate system, allowing these subspaces to have different orientations.

Within each cluster's coordinate system, CRED models augmentation-induced feature transformations as rotations in prescribed, mutually orthogonal 2D subspaces. Assigning each augmentation family to one or more subspaces provides an inductive bias for disentanglement; the feature extractor f learns the representation, and the angle networks predict the corresponding rotation angles.

4.1 Stage 1: Self-Supervised Invariance Learning for Clustering

In the first stage, we use InfoNCE to learn representations for semantic clustering. We then apply spherical k-means (Hornik et al., 2012) or agglomerative clustering (Müllner, 2011) to obtain cluster centroids $\{c_i\}$ and initial sample assignments $c(x)$.

4.2 Stage 2: Self-Supervised Equivariance Learning for Disentanglement

In the second stage, we learn cluster-dependent subspace rotational equivariance within each cluster.

4.2.1 Cluster-Dependent Subspace Rotation Systems

We construct each cluster's rotation system by aligning its anchor with a common reference direction, applying rotations in mutually orthogonal 2D subspaces in the aligned coordinates, and mapping the result back to the original feature space.

Rotation Center. For each cluster i, we define the rotation center r_i as the normalized cluster centroid from the first stage, i.e., $r_i = c_i/\|c_i\|$. The unit vector r_i serves as the cluster's rotation anchor, defining a reference direction left unchanged by the modeled rotations. By default, we keep these anchors fixed during the second stage, using the cluster configuration learned in Stage 1 as a reference. Optionally, we can make the anchors learnable and add the repulsion loss in Equation (23) to encourage their separation.

Rotation Group. To model augmentation-induced variations separately, we construct feature transformations from rotations in mutually orthogonal 2D subspaces. Each planar rotation belongs to SO(2) and is specified by a single angle. Two dimensions are the minimum needed for a nontrivial continuous rotation. Here, the special orthogonal group SO(2) is defined as

$$\mathrm{SO}(2) \triangleq \left\{ R \in \mathbb{R}^{2\times 2} \;\middle|\; R^\top R = I,\ \det(R) = 1 \right\}. \tag{4}$$

For any $\phi \in \mathbb{R}$, the corresponding rotation matrix is $R(\phi) = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}$.

In the basic construction, we assign one rotation plane in $\mathbb{R}^d$ to each mutually orthogonal augmentation family. In the cluster-aligned coordinate system introduced below, family i acts on coordinates $(2i-1, 2i)$, leaving the remaining coordinates unchanged. We fix this allocation in advance and train the feature extractor f so that each augmentation induces a rotation in its assigned feature plane. This avoids the need for an additional dimension-selection mechanism.

We combine these rotations in a block-diagonal matrix:

$$\mathcal{R}(\theta) = \mathrm{diag}(R(\theta^{(1)}), R(\theta^{(2)}), \ldots, R(\theta^{(m)}), I_{d-2m}) \in \mathbb{R}^{d\times d}, \tag{5}$$

where m is the number of rotation planes and $\theta^{(i)}$ is the rotation angle in the i-th plane. We choose $d \geq 2m+1$, with the identity block I_{d-2m} leaving the remaining $d-2m$ coordinates unchanged. For sequentially composed augmentations, angles accumulate within each plane.

The same block structure allows a complex augmentation to act on multiple 2D subspaces, with a separate rotation angle for each affected subspace (see Section 5.3).

In the basic construction, each angle network g_a predicts a rotation angle for each assigned subspace from the augmentation parameter t alone. The predicted angles are shared across samples in the aligned coordinates. Like parameter-conditioned methods such as EquiMod (Devillers & Lefort, 2023), CRED supports stochastic augmentations when their sampled parameters are recorded. Since angles differing by integer multiples of 2π represent the same rotation, we use $\pi \tanh(\cdot)$ as the base angular parameterization, with output range $(-\pi, \pi)$. The symmetric range allows rotations in either direction.

Coordinate Alignment. To use the same block-diagonal rotation structure in different clusters, we align each cluster's rotation anchor r_i with the reference direction $e_d = (0, \ldots, 1)$. For $r_i \neq e_d$, we use the *Householder transformation* (Householder, 1958):

$$u_i = r_i - e_d, \qquad H_i = I - 2\frac{u_i u_i^\top}{u_i^\top u_i}. \tag{6}$$

When $r_i = e_d$, we set $H_i = I$. In either case, $H_i r_i = e_d$. Since H_i is orthogonal, the corresponding rotation in the original feature space is $H_i^\top \mathcal{R}(\theta) H_i$, allowing plane orientations to vary across clusters.

4.2.2 Cluster-Dependent Subspace Rotational Equivariance

In the aligned coordinates, cluster-dependent subspace rotational equivariance is expressed as

$$H_{c(x)} f(a(t, x)) = \mathcal{R}(\theta_{a(t)})\, H_{c(x)} f(x), \quad \forall x \in \mathcal{X},\ a(t) \in \mathcal{A}. \tag{7}$$

Here, the cluster assignment $c(x) := \arg\max_i\ f(x)^\top r_i$ is recomputed at every forward pass using the current representation and rotation anchors. This allows the model to revise the initial cluster assignments during training. The original sample and its augmented view are aligned using the same map $H_{c(x)}$, determined by the original sample's assignment.

The rotation matrix $\mathcal{R}(\theta_{a(t)})$ has the block-diagonal form in Equation (5), where $\theta^{(i)}_{a(t)}$ denotes the predicted angle in subspace i. For sequential compositions, $\theta^{(i)}_{a(t)}$ denotes the accumulated angle in subspace i.

We decompose the coordinate-aligned representation as

$$H_{c(x)} f(x) = ([H_{c(x)} f(x)]_1, \ldots, [H_{c(x)} f(x)]_m, [H_{c(x)} f(x)]_{\text{extra}}), \tag{8}$$

where each equivariant block $[\cdot]_i \in \mathbb{R}^2$ and the remaining coordinates $[\cdot]_{\text{extra}} \in \mathbb{R}^{d-2m}$.

Equivariance alone admits invariant solutions: a feature block remains unchanged if the predicted rotation is the identity or the block is zero. We therefore regularize both the rotation angles and the subspace norms.

- **Non-Identity Rotation**. For augmentation instances sufficiently far from the identity transformation, we penalize a total angular magnitude below a small margin $\epsilon_\theta > 0$ across the assigned subspaces. This allows individual angle components to be zero while discouraging an entirely invariant response.

$$\mathcal{L}_{\text{theta}} = \mathbb{E}_{x,\, a(t) \in \mathcal{A}_{\text{nontrivial}}} \left[\max\left(0, \epsilon_\theta - \sum_{i \in \mathcal{S}_a} |\theta^{(i)}_{a(t)}| \right) \right]. \tag{9}$$

  Here, $\mathcal{A}_{\text{nontrivial}} \subseteq \mathcal{A}$ contains these augmentation instances, and $\mathcal{S}_a$ indexes the subspaces assigned to a. The penalty does not require the exact latent displacement induced by each augmentation. For sequential compositions, the angular penalty is applied separately to the constituent calls in $\mathcal{A}_{\text{nontrivial}}$.

- **Non-Zero Subspace Norm**. We encourage the norm of each aligned 2D feature block to be close to a target radius $\omega > 0$, for both the original and augmented features:

$$\mathcal{L}_{\text{radius}} = \mathbb{E}_{x, a(t)} \left[\sum_{i=1}^{m} \left((\|[H_{c(x)} f(x)]_i / \omega\| - 1)^2 + (\|[H_{c(x)} f(a(t, x))]_i / \omega\| - 1)^2 \right) \right]. \tag{10}$$

Since the features have unit norm, m blocks of radius ω leave a residual norm of $\sqrt{1-m\omega^2}$. We choose $m\omega^2 < 1$ to keep the residual nonzero. When the radius targets and the residual alignment below are satisfied, a smaller ω places features closer to their cluster anchors, supporting cluster separation.

To confine within-cluster variation to the rotation blocks, we constrain the residual components of both aligned views to a common fixed target. Since $H_{c(x)} r_{c(x)} = e_d$, we choose this target along the positive direction of the last coordinate axis, with the residual norm derived above:

$$e_{\text{extra}} := \big(0, \ldots, 0, \sqrt{1-m\omega^2}\big) \in \mathbb{R}^{d-2m}. \tag{11}$$

We combine the rotated blocks of the original aligned representation with this fixed residual to form the target for the augmented view:

$$\bar{f}(x, a(t)) := \big([\mathcal{R}(\theta_{a(t)})\, H_{c(x)} f(x)]_{1:m},\ e_{\text{extra}}\big) \in \mathbb{R}^d. \tag{12}$$

We align the original and augmented features with their respective targets using

$$\mathcal{L}_{\text{equiv}} = \mathbb{E}_{x,a(t)}\Big[\big\|(H_{c(x)} f(a(t,x)) - \bar{f}(x, a(t)))/\omega\big\|^2 + \big\|(H_{c(x)} f(x) - \bar{f}(x,0))/\omega\big\|^2\Big]. \tag{13}$$

Here, $\bar{f}(x,0)$ is the target with zero rotation angles. The first term matches the aligned augmented feature to the rotated blocks and the fixed residual target. Since $\bar{f}(x,0)$ retains the original 2D blocks, the second term penalizes only the original feature's residual error. Dividing the feature errors by ω measures them relative to the target block radius.

Combining this loss with the radius and angle penalties gives the Stage 2 objective:

$$\mathcal{L}(f, \{g_a\}) = \mathcal{L}_{\text{equiv}} + \lambda_{\text{radius}}\, \mathcal{L}_{\text{radius}} + \lambda_{\text{theta}}\, \mathcal{L}_{\text{theta}}. \tag{14}$$

5 Theoretical and Empirical Results

Here, we present theoretical results and empirical validations of CRED. The definitions and assumptions used in the analysis are given in Appendix D.2, and the full statements and proofs in Appendices D.3–D.8.

5.1 Exact Equivariance

Proposition 1 (Perfect Equivariance)**.** *Under the assumptions of Proposition 7 in Appendix D.3, the three losses $\mathcal{L}_{\text{equiv}}, \mathcal{L}_{\text{theta}}, \mathcal{L}_{\text{radius}}$ admit a common zero, which is a global minimum of the Stage 2 objective for positive loss weights. At any such solution, the evaluated augmentation pairs satisfy exact cluster-dependent equivariance. For each augmentation call to which $\mathcal{L}_{\text{theta}}$ is applied, the block rotation is non-identity and the feature representation changes.*

This is because the three losses constrain different aspects of the representation and are mutually compatible by construction. Empirically, all three losses decrease to near-zero values, as shown in Figure 13.

5.2 Disentanglement under Orthogonal Augmentations

We now show that, under the conditions below, perfect equivariance yields a basic form of disentanglement for orthogonal augmentations.

Proposition 2 (Weak Disentanglement)**.** *Suppose the augmentation families $\{a_i\}_{i=1}^n$ are mutually orthogonal and each acts on a single one-dimensional variation component. Each family a_i is assigned exclusively to rotation plane i. Under the assumptions in Appendix D.4, perfect equivariance implies that a_i acts only on the i-th feature block:*

$$\big[H_{c(x)}\, f(a_i(t,x))\big]_i = R\big(\theta^{(i)}_{a_i(t)}\big)\big[H_{c(x)}\, f(x)\big]_i, \quad \big[H_{c(x)}\, f(a_i(t,x))\big]_j = \big[H_{c(x)}\, f(x)\big]_j,\ \forall\, j \neq i. \tag{15}$$

Moreover, there exists a strictly monotonic function $\bar{g}_{a_i}$ on the considered parameter range such that

$$R\big(\theta^{(i)}_{a_i(t)}\big) = R\big(\bar{g}_{a_i}(t)\big). \tag{16}$$

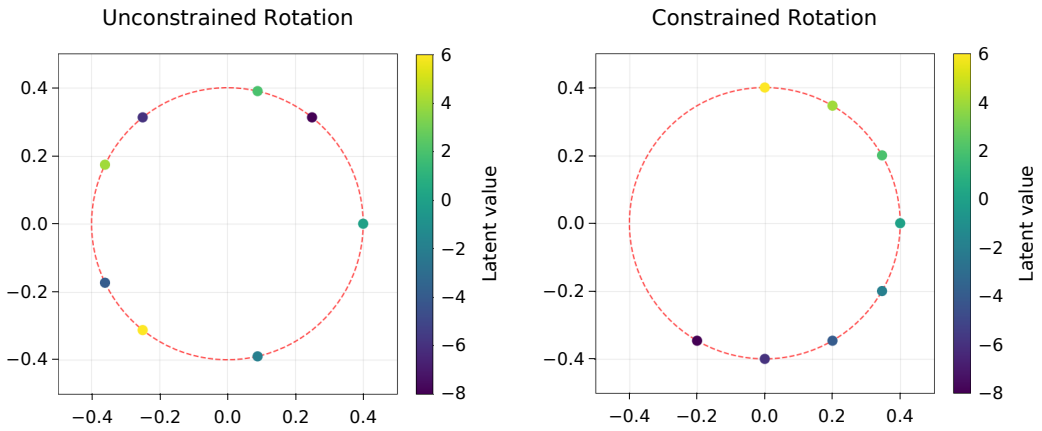


Figure 2: Angular readout for a non-cyclic component. Colors indicate the corresponding latent values. **Left:** With wrap-around, angles read from the features do not vary monotonically with the component values. **Right:** Keeping the unwrapped angular span below 2π permits a monotonic angular readout.

Here t denotes the natural scalar augmentation parameter, such as translation displacement or scaling factor. The analytical function $\bar{g}_{a_i}$ is distinct from the learned network g_{a_i}. The two angles represent the same rotation and can differ only by an integer multiple of 2π.

That is, orthogonal augmentations induce rotations in separate subspaces. Each rotation can be represented by an angle that varies monotonically with the corresponding latent displacement.

Rotation Angle Constraints. Even with a monotonic angle representation, distinct component values can share the same feature direction if a non-cyclic component wraps around its rotation circle or a cyclic component completes multiple turns over one period. To obtain a one-to-one correspondence between component values and feature directions, we require the full-support angular span to remain below 2π for non-cyclic components (Figure 2) and exactly one turn per period for full-period cyclic components. To control these rotations, we bound the predicted angle for each augmentation–subspace pair:

$$\theta^{(i)}_{a(t)} \in \eta^{(i)}_a(-\pi, \pi), \tag{17}$$

where $\pi\eta^{(i)}_a$ is the upper bound on the magnitude of a single predicted rotation. We use $0 < \eta^{(i)}_a \leq 1$ by default. Smaller values impose tighter angle bounds. The bound needed to control the full-support angular span depends on the augmentation displacement relative to the component's full variation range. Remark 8 in Appendix D.5 derives sufficient bounds for non-cyclic and full-period cyclic components and describes their practical selection.

Proposition 3 (Strong Disentanglement)**.** *Under the assumptions of Proposition 2, consider a non-cyclic augmentation-induced variation component whose unwrapped angular span over a cluster's full support is less than 2π. Its values then correspond one-to-one to feature directions in the assigned rotation plane. Relative to a suitable fixed reference direction, the features admit a single-valued, strictly monotonic angular readout of the component within that cluster.*

For full-period cyclic components, Corollary 3 establishes the corresponding one-to-one relation under its stated assumptions, with exactly one turn per period.

A variation isolated to a rotation plane can still affect multiple raw feature coordinates. We express this learned geometry in interpretable coordinates through unsupervised post-processing. For non-cyclic components, **relative angles (RAs)** are measured from the estimated midpoint of each cluster's covered arc; cyclic components retain their two-dimensional circular representation. We concatenate these coordinates with a one-hot **cluster indicator (CI)** encoding the learned cluster assignment. Appendix D.6 gives the construction and the conditions for monotonic readout and alignment across clusters.

Table 1: DCI scores using random forest regression on four datasets. D, C, and I denote disentanglement, completeness, and informativeness. For CRED, raw denotes the original learned features, and RAs + CI denotes the post-processed representation.

Method	MPI3D			Shape3D			3DIdent			3DIEBench		
	D	C	I	D	C	I	D	C	I	D	C	I
SimCLR	0.09	0.09	0.88	0.07	0.07	0.89	0.13	0.10	0.93	0.10	0.10	0.90
EquiMod	0.14	0.13	0.96	0.13	0.12	0.97	0.07	0.02	0.93	0.09	0.09	0.91
CARE	0.08	0.09	0.89	0.08	0.08	0.93	0.15	0.12	0.93	0.12	0.11	0.94
IP-IRM	0.22	0.20	0.96	0.14	0.11	0.95	0.15	0.09	0.93	0.11	0.11	0.93
Eastwood et al. (2023)	0.95	0.83	**1.00**	0.94	0.77	**1.00**	0.94	0.77	**1.00**	0.36	0.30	0.95
CRED (raw)	0.63	0.51	**1.00**	0.63	0.49	**1.00**	**0.99**	0.94	**1.00**	0.60	0.48	**1.00**
CRED (RAs + CI)	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**0.99**	**0.99**	**1.00**	**1.00**	**1.00**	**1.00**

5.2.1 Empirical Results

We evaluate CRED in two complementary settings. **(i) Datasets with ground-truth latents.** We consider MPI3D (Gondal et al., 2019) and Shape3D (Kim & Mnih, 2018) (primarily with discrete latents), as well as 3DIdent (Zimmermann et al., 2021) and 3DIEBench (Garrido et al., 2023b) (with continuous latents). Latent transformations are used as augmentations. We use DCI (Eastwood & Williams, 2018) as the evaluation metric, which quantifies disentanglement, completeness, and informativeness via supervised regressors mapping features to ground-truth latents, with higher scores indicating better performance.

Table 1 compares CRED with the invariance baseline SimCLR (Chen et al., 2020a), the equivariance baselines EquiMod (Devillers & Lefort, 2023) and CARE (Gupta et al., 2024), and the disentanglement baselines IP-IRM (Wang et al., 2021) and the method of Eastwood et al. (2023). Under random forest evaluation, CRED's raw features already achieve an informativeness score of 1.00 on all four datasets and outperform all compared baselines in disentanglement and completeness on 3DIdent and 3DIEBench. With RAs + CI post-processing, CRED achieves at least 0.99 on all three DCI metrics across all four datasets, surpassing all compared baselines in disentanglement and completeness while matching or exceeding their informativeness. Complementary DCI results using linear regression are reported in Table 6.

Ablations on Shape3D show that removing either the radius or angle regularizer leads to collapse in the rotation subspaces, while omitting SimCLR pretraining substantially reduces informativeness (Appendix B.3).

(ii) Datasets without ground-truth latents. We further evaluate CRED and the same five baselines on MNIST, CIFAR-10 and CIFAR-100 with standard data augmentations. Since ground-truth latents are unavailable, DCI is not applicable. Instead, we assess augmentation-induced representation structure using four metrics: rotational equivariance (ρ_f), relative equivariance (γ_f), and two newly proposed disentanglement metrics, Mean Overlap and Mean Domination. Linear probing and clustering accuracies provide complementary assessments of class discriminability.

For structural evaluation, we use the feature coordinates in which each method's equivariance objective is defined, or the original feature coordinates for methods without such an objective. In the metric definitions below, $f(x)$ and $f(a(x))$ denote the original and augmented representations in these evaluation coordinates. For CRED, both views are aligned by the same matrix $H_{c(x)}$, determined by the original sample's cluster assignment; the baselines use their original feature coordinates. For a second sample x', its original and augmented views similarly share $H_{c(x')}$.

Rotational Equivariance (ρ_f, lower is better) measures how well a single rotation explains the feature changes induced by the same augmentation instance across samples. We define ρ_f as the ratio of the equivariance residual to the invariance residual:

$$\rho_f := \frac{\mathbb{E}_x \|f(a(x)) - R_a f(x)\|^2}{\mathbb{E}_x \|f(a(x)) - f(x)\|^2}. \tag{18}$$

For our method (CRED), R_a corresponds to the learned rotation $\mathcal{R}(\theta_{a(t)})$. For methods without explicit rotation parameterization, we recover the optimal rotation as $R_a = \arg\min_{R \in \mathrm{SO}(d)} \mathbb{E}_x \|f(a(x)) - Rf(x)\|^2$. Equivariance alone admits degenerate invariant solutions; near such solutions, even the identity rotation yields a small absolute residual. We therefore normalize by the augmentation-induced feature change. For a nonzero denominator, the identity rotation gives $\rho_f = 1$, whereas a small ρ_f indicates that the learned or fitted rotation explains most of this change.

Relative Equivariance (γ_f, lower is better) (Gupta et al., 2024) measures whether applying the same augmentation preserves pairwise distances in feature space:

$$\gamma_f := \mathbb{E}_{a,x,x'} \left[\frac{(\|f(a(x')) - f(a(x))\|^2 - \|f(x') - f(x)\|^2)^2}{(\|f(a(x')) - f(x')\|^2 + \|f(a(x)) - f(x)\|^2)^2} \right]. \tag{19}$$

Since shared rotations preserve these distances, the metric assesses this consequence of rotational equivariance without estimating a rotation matrix.

To probe disentanglement without ground-truth latents, we propose two metrics that assess whether different augmentations affect disjoint subsets of feature dimensions, inspired by the subspace view of disentanglement in Higgins et al. (2018). For an augmentation set $\mathcal{A}$, we first define the per-dimension sensitivity to augmentation a as

$$s_a^{(k)} := \frac{\mathbb{E}_x \left[\left(f_k(a(x)) - f_k(x) \right)^2 \right]}{\sum_{k'=1}^{d} \mathbb{E}_x \left[\left(f_{k'}(a(x)) - f_{k'}(x) \right)^2 \right]}, \quad k = 1, \ldots, d. \tag{20}$$

When the denominator is nonzero, $s_a^{(k)}$ is coordinate k's share of the total expected squared feature change induced by a, and $\sum_k s_a^{(k)} = 1$.

Mean Overlap (lower is better) measures the extent to which different augmentations activate overlapping subsets of feature dimensions:

$$\text{Mean Overlap} := \frac{1}{|\mathcal{A}|(|\mathcal{A}| - 1)} \sum_{\substack{a,a' \in \mathcal{A} \\ a \neq a'}} \sum_{k=1}^{d} \min\left(s_a^{(k)}, s_{a'}^{(k)} \right). \tag{21}$$

The inner sum measures overlap between two normalized sensitivity distributions. The metric is 0 when augmentations affect disjoint feature dimensions, and 1 when all sensitivity vectors are identical.

Mean Domination (higher is better) measures the extent to which each feature dimension is primarily associated with a single augmentation. We weight each dimension by its total sensitivity, $\text{mass}(k) := \sum_a s_a^{(k)}$, so dimensions with little total sensitivity receive less weight. The weighted average excludes dimensions with zero mass and simplifies to:

$$\text{Mean Domination} := \frac{\sum_{\substack{1 \leq k \leq d \\ \text{mass}(k) > 0}} \text{mass}(k) \cdot \frac{\max_{a \in \mathcal{A}} s_a^{(k)}}{\text{mass}(k)}}{\sum_{k=1}^{d} \text{mass}(k)} = \frac{1}{|\mathcal{A}|} \sum_{k=1}^{d} \max_{a \in \mathcal{A}} s_a^{(k)}, \tag{22}$$

The mass weights cancel in the numerator, zero-mass dimensions contribute zero, and $\sum_k \text{mass}(k) = |\mathcal{A}|$ because each sensitivity vector sums to 1. The metric lies in $[1/|\mathcal{A}|, 1]$. It equals $1/|\mathcal{A}|$ when all augmentations have identical sensitivity vectors, and 1 when each active dimension responds to only one augmentation.

We estimate the expectations in the four structural metrics using 2,000 samples and 50 randomly sampled augmentation instances. The same samples and augmentation parameters are used for all methods.

Table 2: Comparison on MNIST, CIFAR-10, and CIFAR-100. Linear probing accuracy (top-1) and clustering accuracy are reported as percentages on the training and test splits. Arrows indicate the preferred direction.

	Method	Accuracy				Equivariance		Disentanglement	
		Linear eval. ↑		Clustering ↑		Rotational	Relative	Mean ↓	Mean ↑
		Train	Test	Train	Test	$\rho_f \downarrow$	$\gamma_f \downarrow$	Overlap	Domination
MNIST	SimCLR	99.82	99.70	98.90	96.71	0.556	9.967	0.908	0.179
	EquiMod	**99.88**	99.65	97.41	97.17	0.646	5.612	0.894	0.186
	CARE	99.87	99.66	97.17	90.01	0.461	0.516	0.839	0.211
	IP-IRM	99.65	99.75	45.61	47.18	0.656	0.396	0.816	0.221
	Eastwood et al. (2023)	99.82	**99.78**	93.22	91.58	0.328	0.036	0.384	0.405
	CRED	99.65	99.72	**99.25**	**99.49**	**0.015**	**0.012**	**0.043**	**0.942**
CIFAR-10	SimCLR	**92.68**	92.03	87.57	83.19	0.930	2.660	0.919	0.541
	EquiMod	92.48	**92.24**	84.62	80.65	0.362	0.164	0.772	0.614
	CARE	92.12	92.09	86.09	84.32	0.896	1.041	0.911	0.544
	IP-IRM	92.62	92.04	40.75	42.94	0.970	2.381	0.915	0.543
	Eastwood et al. (2023)	92.65	92.18	84.43	83.32	0.318	0.103	0.166	0.917
	CRED	92.07	92.06	**88.60**	**86.81**	**0.022**	**0.089**	**0.136**	**0.932**
CIFAR-100	SimCLR	83.65	72.05	49.74	**45.81**	0.983	8.368	0.877	0.562
	EquiMod	**84.96**	**72.41**	48.22	43.80	0.968	2.687	0.924	0.538
	CARE	83.37	71.84	48.67	45.72	0.973	3.622	0.880	0.560
	IP-IRM	83.41	72.17	48.01	45.24	0.969	6.207	0.895	0.553
	Eastwood et al. (2023)	74.94	72.19	38.54	36.08	0.222	0.181	0.136	0.932
	CRED	82.55	71.96	**50.55**	42.49	**0.076**	**0.090**	**0.120**	**0.940**

Table 2 shows that CRED outperforms all five baselines on all four structural metrics across MNIST, CIFAR-10, and CIFAR-100. It also achieves competitive linear-probe test accuracy and the highest clustering accuracy on MNIST and CIFAR-10. Experimental details are provided in Appendix A.

5.3 Disentanglement under Complex Augmentations

Real-world augmentations can couple multiple types of variation. We aim to separate these components into distinct feature subspaces while using the same subspace for a given component across augmentations. We first analyze augmentations that combine variations induced by existing orthogonal augmentations, then consider those introducing an additional variation that can be isolated at an anchor configuration.

Proposition 4 (Disentanglement of Simple Composite Augmentations)**.** *Suppose n mutually orthogonal augmentation families are each assigned exclusively to a distinct rotation plane. Under the assumptions of Proposition 9 in Appendix D.7, perfect equivariance implies weak disentanglement under a composite augmentation $\tilde{a}$ affecting $s \leq n$ of these variations: the rotation in each affected plane admits the same strictly monotonic angle representation of displacement as the corresponding orthogonal augmentation, while unaffected planes remain unchanged.*

Under the chain condition, a sample and its composite-augmented view are connected through the orthogonal augmentation constraints. At perfect equivariance, the directly predicted composite rotation and the rotation inferred along this chain must produce the same augmented feature. Since each 2D feature block is nonzero, the two rotations must agree in every plane.

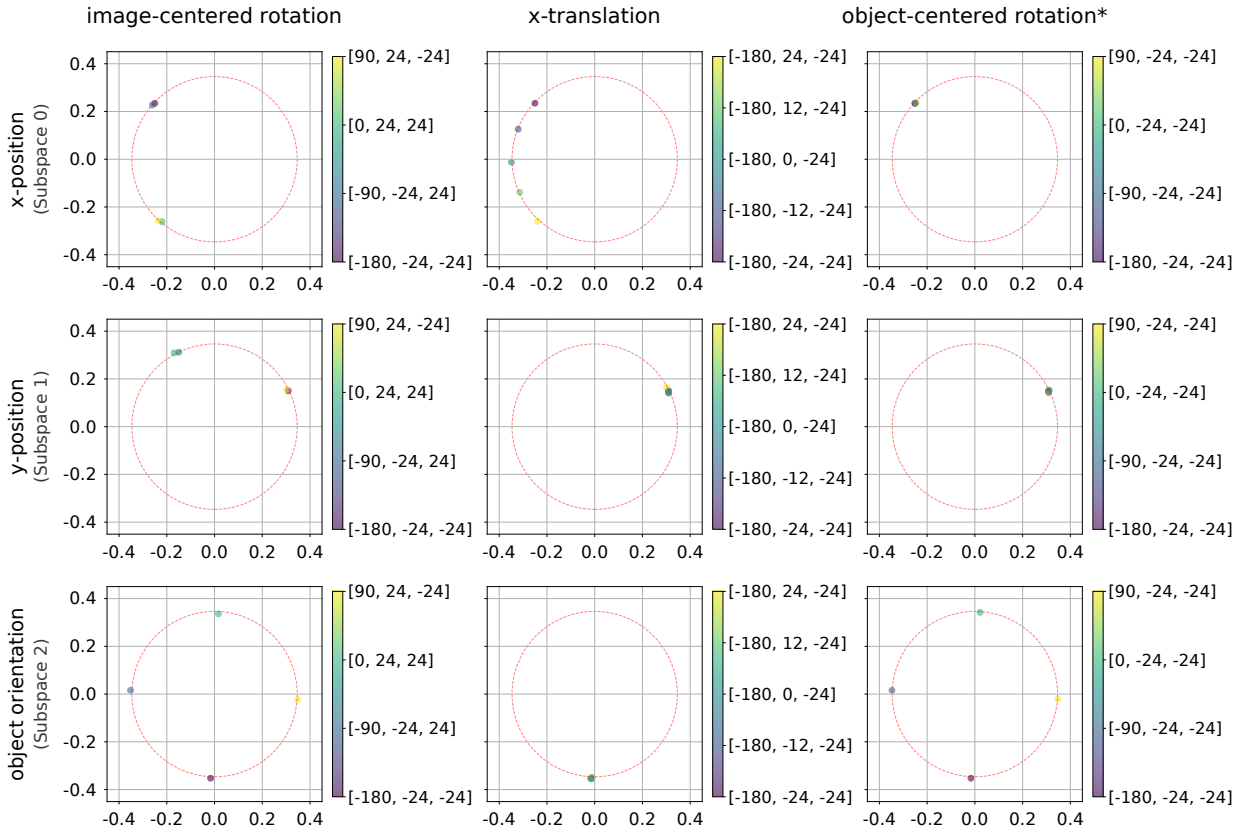


Figure 3: CRED feature projections from one cluster. The model is trained using x-translation, y-translation, and image-centered rotation. Columns show feature projections under image-centered rotation, x-translation, and object-centered rotation*, from left to right; rows show the feature planes for x-position, y-position, and object orientation, from top to bottom. Here, $*$ indicates that object-centered rotation is virtual. This rotation acts about the object's own center and is not supplied as a separate training augmentation. CRED disentangles the changes in x-position, y-position, and object orientation induced by image-centered rotation into separate feature planes. Colorbar labels give (object orientation, x-position, y-position).

The chain argument applies separately to each sample, so the result also covers sample-dependent displacements of the existing variation components. We predict the corresponding rotation angles using $g_{\tilde{a}}(t, x)$, retaining $\theta_{\tilde{a}(t)}$ as shorthand for the predicted angle vector. The angles may vary across samples, while each plane follows the same relation between component displacement and feature rotation. The subspace allocation and losses remain unchanged.

We next consider augmentations that introduce a new variation component. We use an anchor configuration to establish the relation between its displacement and rotations in an additional plane.

Proposition 5 (Disentanglement of Anchored Complex Augmentation)**.** *Suppose $n-1$ mutually orthogonal augmentation families are each assigned exclusively to a distinct rotation plane. Let $\tilde{a}$ affect a subset of these variations together with one new variation. Suppose there is an anchor configuration at which it changes only the new variation, with a displacement determined by t alone. At this anchor, we use the basic parameter-only construction, restricting the predicted rotation to an additional n-th plane. Under the assumptions of Proposition 10 in Appendix D.8, perfect equivariance implies weak disentanglement on the covered augmentation pairs: the rotation in each affected plane admits a common strictly monotonic angle representation of the corresponding displacement.*

At the anchor, the single-axis analysis establishes the relation between displacement and rotation for the new component. The same chain argument then combines this relation with those of the existing orthogonal augmentations to determine the composite rotation away from the anchor.

Both results extend to strong disentanglement when each non-cyclic component has a full-support angular span below 2π and each cyclic component completes exactly one turn per period (Corollary 4).

Empirical Results. In the anchored setting, image-centered rotation changes both the position and orientation of an off-center object, with the position change depending on its initial position. An object at the image center provides the anchor configuration, where only orientation changes. Figure 3 shows CRED disentangling the position and orientation changes induced by image-centered rotation into separate feature planes. In the simple composite setting, CRED disentangles the changes in x-position and y-position induced by xy-translation into the planes assigned to x-translation and y-translation, respectively (Figure 14). Additional visualizations are provided in Appendix B.6, and experimental details in Appendix A.

6 Limitations and Future Directions

Like other augmentation-based self-supervised methods, CRED depends on the choice and coverage of augmentations (Wang et al., 2022; Parulekar et al., 2023). Our guarantees assume within-cluster connectivity and perfect equivariance, which may not hold exactly on natural images. Nevertheless, CRED achieves better equivariance and subspace separation than the evaluated baselines, including on CIFAR-100 with imperfect clustering (Table 2). Designing augmentation families that better connect samples within clusters and reflect underlying style variations could help narrow this gap between the theoretical setting and practical data.

The current formulation requires manual configuration of the rotation subspaces, including their number and their assignment to the basic orthogonal augmentation families. The anchored extension additionally requires prior knowledge of a configuration at which a composite augmentation induces only the new variation component, together with a way to identify or construct samples at that configuration. Automating subspace configuration and reducing reliance on such anchor configurations could broaden applicability to less controlled settings. More general models of feature transformations could further extend the framework beyond the current product of SO(2) rotations.

7 Conclusion

CRED decouples invariance and equivariance learning to avoid interference between the two objectives. Its cluster-dependent SO(2) rotations in mutually orthogonal two-dimensional subspaces support multi-cluster representations on a single hypersphere. We establish conditions for exact disentanglement of variations induced by orthogonal and composite augmentations. After post-processing, all three DCI scores reach at least 0.99 across four controlled datasets under random forest evaluation. On three real-world datasets, CRED outperforms all evaluated baselines on all four equivariance and disentanglement metrics.

References

Brandon Anderson, Truong Son Hy, and Risi Kondor. Cormorant: Covariant molecular neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022a.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICRegL: Self-supervised learning of local visual features. *Advances in Neural Information Processing Systems*, 35:8799–8810, 2022b.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Diane Bouchacourt, Ryota Tomioka, and Sebastian Nowozin. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 2095–2102, 2018.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision*, pp. 132–149, 2018.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020.

Junxiang Chen and Kayhan Batmanghelich. Weakly supervised disentanglement by pairwise similarities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(4):3495–3502, 2020.

Ricky T. Q. Chen, Xuechen Li, Roger B. Grosse, and David K. Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in Neural Information Processing Systems*, 31, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607. PMLR, 2020a.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in Neural Information Processing Systems*, 29, 2016.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Brian Cheung, Jesse A. Livezey, Arjun K. Bansal, and Bruno A. Olshausen. Discovering hidden factors of variation in deep networks. *arXiv preprint arXiv:1412.6583*, 2014.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning*, pp. 2990–2999. PMLR, 2016.

Taco S. Cohen and Max Welling. Steerable CNNs. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=rJQKYt5ll.

Taco S. Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical CNNs. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=Hkbd5xZRb.

Rumen Dangovski, Li Jing, Charlotte Loh, Seungwook Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljačić. Equivariant self-supervised learning: Encouraging equivariance in representations. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gKLAAfiytI.

Alexandre Devillers and Mathieu Lefort. EquiMod: An equivariance module to improve visual instance discrimination. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=eDLwjKmtYFt.

Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

Cian Eastwood, Julius von Kügelgen, Linus Ericsson, Diane Bouchacourt, Pascal Vincent, Bernhard Schölkopf, and Mark Ibrahim. Self-supervised disentanglement by leveraging structure in data augmentations. *arXiv preprint arXiv:2311.08815*, 2023.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *International Conference on Machine Learning*, pp. 3015–3024. PMLR, 2021.

Babak Esmaeili, Hao Wu, Sarthak Jain, Alican Bozkurt, N. Siddharth, Brooks Paige, Dana H. Brooks, Jennifer Dy, and Jan-Willem van de Meent. Structured disentangled representations. In *International Conference on Artificial Intelligence and Statistics*, pp. 2525–2534. PMLR, 2019.

Marc Finzi, Samuel Stanton, Pavel Izmailov, and Andrew Gordon Wilson. Generalizing convolutional neural networks for equivariance to Lie groups on arbitrary continuous data. In *International Conference on Machine Learning*, pp. 3165–3176. PMLR, 2020.

Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. SE(3)-Transformers: 3D roto-translation equivariant attention networks. *Advances in Neural Information Processing Systems*, 33:1970–1981, 2020.

Quentin Garrido, Yubei Chen, Adrien Bardes, Laurent Najman, and Yann LeCun. On the duality between contrastive and non-contrastive self-supervised learning. In *International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=kDEL91Dufpa.

Quentin Garrido, Laurent Najman, and Yann LeCun. Self-supervised learning of split invariant equivariant representations. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10975–10996. PMLR, 2023b. URL https://proceedings.mlr.press/v202/garrido23b.html.

Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. *Advances in Neural Information Processing Systems*, 32:15714–15725, 2019.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: a new approach to self-supervised learning. *Advances in Neural Information Processing Systems*, 33:21271–21284, 2020.

Sharut Gupta, Joshua Robinson, Derek Lim, Soledad Villar, and Stefanie Jegelka. Structuring representation geometry with rotationally equivariant contrastive learning. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/d607a260b3bc0b9a704a1a04dd64040a-Abstract-Conference.html.

Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *Advances in Neural Information Processing Systems*, 34:5000–5011, 2021.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738, 2020.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

Geoffrey E. Hinton and Ruslan R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

Kurt Hornik, Ingo Feinerer, Martin Kober, and Christian Buchta. Spherical k-means clustering. *Journal of Statistical Software*, 50(10):1–22, 2012.

Haruo Hosoya. Group-based learning of disentangled representations with generalizability for novel contents. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 2506–2513, 2019. doi: 10.24963/ijcai.2019/348. URL https://ijcai.org/proceedings/2019/348.

Alston S. Householder. Unitary triangularization of a nonsymmetric matrix. *Journal of the ACM (JACM)*, 5(4):339–342, 1958.

Vincent Tao Hu, Wei Zhang, Meng Tang, Pascal Mettes, Deli Zhao, and Cees Snoek. Latent space editing in transformer-based flow matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 2247–2255, 2024.

Insu Jeon, Wonkwang Lee, Myeongjang Pyeon, and Gunhee Kim. IB-GAN: Disentangled representation learning with information bottleneck generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7926–7934, 2021.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pp. 2649–2658. PMLR, 2018.

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

Sebastien Lachapelle, Tristan Deleu, Divyat Mahajan, Ioannis Mitliagkas, Yoshua Bengio, Simon Lacoste-Julien, and Quentin Bertrand. Synergies between disentanglement and sparsity: Generalization and identifiability in multi-task learning. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 18171–18206. PMLR, 2023. URL https://proceedings.mlr.press/v202/lachapelle23a.html.

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

Attila Lengyel, Ombretta Strafforello, Robert-Jan Bruintjes, Alexander Gielisse, and Jan van Gemert. Color equivariant convolutional networks. *Advances in Neural Information Processing Systems*, 36:29831–29850, 2023.

Junnan Li, Pan Zhou, Caiming Xiong, and Steven C. H. Hoi. Prototypical contrastive learning of unsupervised representations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=KmykpuSrjcq.

Yang Li, Quan Pan, Suhang Wang, Haiyun Peng, Tao Yang, and Erik Cambria. Disentangled variational auto-encoder for semi-supervised learning. *Information Sciences*, 482:73–85, 2019.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pp. 4114–4124. PMLR, 2019.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pp. 6348–6359. PMLR, 2020.

Giovanni Luca Marchetti, Gustaf Tegnér, Anastasiia Varava, and Danica Kragic. Equivariant representation learning via class-pose decomposition. In *International Conference on Artificial Intelligence and Statistics*, pp. 4745–4756. PMLR, 2023.

Emile Mathieu, Tom Rainforth, N. Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders. In *International Conference on Machine Learning*, pp. 4402–4412. PMLR, 2019.

Daniel Müllner. Modern hierarchical, agglomerative clustering algorithms. *arXiv preprint arXiv:1109.2378*, 2011.

Utkarsh Ojha, Krishna Kumar Singh, Cho-Jui Hsieh, and Yong Jae Lee. Elastic-InfoGAN: Unsupervised disentangled representation learning in class-imbalanced data. *Advances in Neural Information Processing Systems*, 33:18063–18075, 2020.

Advait Parulekar, Liam Collins, Karthikeyan Shanmugam, Aryan Mokhtari, and Sanjay Shakkottai. InfoNCE loss provably learns cluster-preserving representations. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 1914–1961. PMLR, 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

Ashwini Pokle, Jinjin Tian, Yuchen Li, and Andrej Risteski. Contrasting the landscape of contrastive and non-contrastive learning. In *International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 8592–8618. PMLR, 2022. URL `https://proceedings.mlr.press/v151/pokle22a.html`.

Scott Reed, Kihyuk Sohn, Yuting Zhang, and Honglak Lee. Learning to disentangle factors of variation with manifold interaction. In *International Conference on Machine Learning*, pp. 1431–1439. PMLR, 2014.

Xuanchi Ren, Tao Yang, Yuwang Wang, and Wenjun Zeng. Rethinking content and style: exploring bias for unsupervised disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1823–1832, 2021.

Víctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E(n) equivariant graph neural networks. In *International Conference on Machine Learning*, pp. 9323–9332. PMLR, 2021.

Mehran Shakerinava, Arnab Kumar Mondal, and Siamak Ravanbakhsh. Structuring representations using group invariants. *Advances in Neural Information Processing Systems*, 35:34162–34174, 2022.

Rui Shu, Yining Chen, Abhishek Kumar, Stefano Ermon, and Ben Poole. Weakly supervised disentanglement with guarantees. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=HJgSwyBKvr`.

Yue Song, Andy Keller, Nicu Sebe, and Max Welling. Flow factorized representation learning. *Advances in Neural Information Processing Systems*, 36:49761–49782, 2023.

Kai Sheng Tai, Peter Bailis, and Gregory Valiant. Equivariant transformer networks. In *International Conference on Machine Learning*, pp. 6086–6095. PMLR, 2019.

Pieter Merkus Lambertus Tammes. On the origin of number and arrangement of the places of exit on the surface of pollen-grains. *Recueil des travaux botaniques néerlandais*, 27(1):1–84, 1930.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision*, pp. 776–794. Springer, 2020.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Sjoerd van Steenkiste, Francesco Locatello, Jürgen Schmidhuber, and Olivier Bachem. Are disentangled representations helpful for abstract visual reasoning? In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL `https://proceedings.neurips.cc/paper_files/paper/2019/hash/bc3c4a6331a8a9950945a1aa8c95ab8a-Abstract.html`.

Tan Wang, Zhongqi Yue, Jianqiang Huang, Qianru Sun, and Hanwang Zhang. Self-supervised learning disentangled group representation as feature. *Advances in Neural Information Processing Systems*, 34: 18225–18240, 2021.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Yifei Wang, Qi Zhang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=ECvgmYVyeUz.

Maurice Weiler and Gabriele Cesa. General E(2)-equivariant steerable CNNs. *Advances in Neural Information Processing Systems*, 32, 2019.

Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco S. Cohen. 3D steerable CNNs: Learning rotationally equivariant features in volumetric data. *Advances in Neural Information Processing Systems*, 31, 2018.

Ancong Wu and Wei-Shi Zheng. Factorized diffusion autoencoder for unsupervised disentangled representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5930–5939, 2024.

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3733–3742, 2018.

Tete Xiao, Xiaolong Wang, Alexei A. Efros, and Trevor Darrell. What should not be contrastive in contrastive learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=CZ8Y3NzuVzO.

Tao Yang, Yuwang Wang, Yan Lu, and Nanning Zheng. DisDiff: Unsupervised disentanglement of diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 69130–69156, 2023. doi: 10.52202/075280-3026. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/da47bfaf3f3a8d5bbab0d60c5195dc18-Abstract-Conference.html.

Yulong Yang, Felix O'Mahony, and Christine Allen-Blanchette. Learning color equivariant representations. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/2e098001625a8b8f22bf6c1aef09c4e5-Abstract-Conference.html.

Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. In *European Conference on Computer Vision*, pp. 668–684. Springer, 2022.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320. PMLR, 2021.

Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In *International Conference on Machine Learning*, pp. 12979–12990. PMLR, 2021.

Appendix guide.

- Appendix A: Algorithm and Experimental Details p. 19
- Appendix B: Additional Results and Visualizations p. 24
- Appendix C: Extended Related Work p. 36
- Appendix D: Theoretical Analysis and Proofs p. 36

A Algorithm and Experimental Details

Algorithm 1 summarizes training and optional post-processing. For sequential compositions, we sum the predicted angles within each plane and apply $\mathcal{L}_{\text{theta}}$ separately to each constituent call in $\mathcal{A}_{\text{nontrivial}}$.

A.1 Dataset

Shape3D (Kim & Mnih, 2018) This dataset provides 5 style latents (floor hue, wall hue, object hue, object scale, azimuth) and 4 object shapes as the class latent. For each style latent, we keep the middle half of its range and include all 4 classes, yielding a total of 16,000 images.

MPI3D (Gondal et al., 2019) This dataset contains 6 style latents (object color, object size, camera height, background color, horizontal rotation, vertical rotation) and 6 object shapes as the class latent. We keep the first four style latents in full. For the two rotations, we subsample the angles to $\{0, 4, 8, 12\}$ (uniformly spaced indices), resulting in a total of 10,368 images.

3DIdent (Zimmermann et al., 2021) This dataset has a single class and 10 continuous style factors: position (x, y, z); rotation (ϕ, θ, ψ); spotlight position; object hue; ground hue; and spotlight hue. We randomly sample 50,000 images for training.

3DIEBench (Garrido et al., 2023b) 3DIEBench is derived from a ShapeNetCore subset (sourced from 3D Warehouse). To generate the non-symmetric images, we use 10 chair models as classes and render images in Blender, varying 5 parameters uniformly over their ranges. We generate 20,000 images at a resolution of 224×224. Parameter ranges are listed in Table 3, and example renderings are shown in Figure 4.

Synthetic letter dataset The synthetic letter dataset used in Section 5.3 provides controlled variation in object position and orientation for evaluating disentanglement under composite augmentations. It consists of 96×96 grayscale images of three distinct uppercase letters, "R," "E," and "D." For each letter, we apply three controlled augmentations: horizontal translation (x-shift), vertical translation (y-shift), and in-plane rotation about the image center. In the image-centered rotation experiment, the anchor configuration is obtained by placing the object at the image center, where rotation changes only its orientation.

MNIST (LeCun et al., 1998) The MNIST dataset contains $70,000$ grayscale handwritten digit images of size 28×28, with $60,000$ images for training and $10,000$ for testing across 10 classes. Unlike the synthetic and benchmark datasets above, MNIST does not provide ground-truth style latents. We instead apply standard data augmentations as the source of variation.

CIFAR-10 and CIFAR-100 (Krizhevsky, 2009) CIFAR-10 and CIFAR-100 contain 10 and 100 classes, respectively. Each contains $60,000$ color images of size 32×32, split into $50,000$ training and $10,000$ test images. Neither provides ground-truth style latents, so we use standard augmentations to induce variation.

A.2 Hardware

All experiments were performed on 2 NVIDIA Tesla V100 GPUs with 32GB accelerator RAM for a single training run. All experiments use the PyTorch deep learning framework (Paszke et al., 2019).

Algorithm 1 Cluster-Dependent Subspace Rotational Equivariance (with post-processing)

Input Dataset $\{x\}$, augmentations $\mathcal{A} = \{a_i\}_{i=1}^n$; feature extractor f, angle networks $G = \{g_i\}_{i=1}^n$; number of clusters k, number of rotation planes m, assigned rotation planes and angle-bound parameters $\{\eta_a^{(i)}\}$ for each augmentation a, subspace radius ω; loss weights $\lambda_{\text{radius}}, \lambda_{\text{theta}}$, epochs T_1, T_2; optional phase-alignment matrices $\{A_{j,i}\}$
Output Trained feature extractor f and angle networks $G = \{g_i\}_{i=1}^n$; optionally, post-processed representations $\{\widehat{r}(x)\}$.
Stage 1: Self-Supervised Invariance Learning for Clustering
Optional; skipped when $k = 1$.
for $e = 1$ to T_1 **do**
  for each minibatch $\{x_j\}$ **do**
    Generate two augmented views of each image.
    Compute $\mathcal{L}_{\text{InfoNCE}}$
    Update f by $\nabla\mathcal{L}_{\text{InfoNCE}}$
  end for
end for
Stage 2: Self-Supervised Equivariance Learning for Disentanglement
Each angle network g_a uses $\eta_a^{(i)}\pi\tanh(\cdot)$ as its output parameterization on each assigned plane i.
Initialize cluster centroids $\{c_\ell\}$ by clustering $f(x)/\|f(x)\|_2$ with the chosen method.
Set $r_\ell \leftarrow \text{normalize}(c_\ell)$.
for $e = 1$ to T_2 **do**
  for each minibatch $\{x_j\}$ **do**
    Generate $x_j^+ = a(t, x_j)$ with $a(t) \in \mathcal{A}$
    $u_j \leftarrow f(x_j)/\|f(x_j)\|_2, \quad u_j^+ \leftarrow f(x_j^+)/\|f(x_j^+)\|_2$
    Assign $c(x_j) = \arg\max_i u_j^\top r_i$
    Construct $H_{c(x_j)}$ with $H_{c(x_j)} r_{c(x_j)} = e_d$, using it for both u_j and u_j^+
    Predict $\theta_j \in \mathbb{R}^m$ using $g_a(t)$ (or $g_a(t, x_j)$ if sample-dependent); zero unassigned entries
    Build $\mathcal{R}(\theta_j)$ according to Equation (5)
    Compute $\mathcal{L}_{\text{equiv}}$, $\mathcal{L}_{\text{radius}}$, $\mathcal{L}_{\text{theta}}$
    Update $f, \{g_i\}$ by $\nabla(\mathcal{L}_{\text{equiv}} + \lambda_{\text{radius}}\mathcal{L}_{\text{radius}} + \lambda_{\text{theta}}\mathcal{L}_{\text{theta}})$
  end for
end for
Optional: post-processing and final representation
Let S_i denote the reference samples assigned to cluster i.
For each non-cyclic plane j, define $\theta^{(j)}(x) := \text{atan2}\big([H_{c(x)}f(x)]_{j,y}, [H_{c(x)}f(x)]_{j,x}\big)$.
for each cluster i and non-cyclic plane j **do**
  Sort the distinct values in $\{\theta^{(j)}(x) : x \in S_i\}$ as $\theta_1 < \cdots < \theta_n$ $(n \geq 2)$
  $\theta_{n+1} \leftarrow \theta_1 + 2\pi, \quad \delta_\ell \leftarrow \theta_{\ell+1} - \theta_\ell$ for $\ell = 1, \ldots, n$
  Choose $\ell^* \in \arg\max_{1 \leq \ell \leq n} \delta_\ell$
  $\theta_R \leftarrow \theta_{\ell^*+1}, \quad L \leftarrow 2\pi - \delta_{\ell^*}$
  $\alpha_{i,j} \leftarrow \text{atan2}\big(\sin(\theta_R + L/2), \cos(\theta_R + L/2)\big)$
end for
for each x with cluster $i = c(x)$ **do**
  for each plane $j = 1, \ldots, m$ **do**
    if plane j is non-cyclic **then**
      $\widehat{f}^{(j)}(x) \leftarrow \text{atan2}\big(\sin(\theta^{(j)}(x) - \alpha_{i,j}),\ \cos(\theta^{(j)}(x) - \alpha_{i,j})\big)$ ▷ scalar
    else
      $\widehat{f}^{(j)}(x) \leftarrow [H_i f(x)]_j \in \mathbb{R}^2$ ▷ retain the two-dimensional cyclic block
      if phase-alignment matrices are supplied **then**
        $\widehat{f}^{(j)}(x) \leftarrow A_{j,i}^{-1}\widehat{f}^{(j)}(x)$
      end if
    end if
  end for
  $\widehat{r}(x) \leftarrow \text{concat}\big(\widehat{f}^{(1)}(x), \ldots, \widehat{f}^{(m)}(x), e_i\big)$ ▷ e_i: one-hot cluster indicator
end for

Table 3: Parameter ranges used in our 3DIEBench setup.

Parameter	**Minimum value**	**Maximum value**
Object position X	-0.3	0.3
Object position Y	-0.3	0.3
Object position Z	-0.3	0.3
Object rotation ψ (radians)	$-\pi/2$	$\pi/2$
Floor hue	0.3	0.7

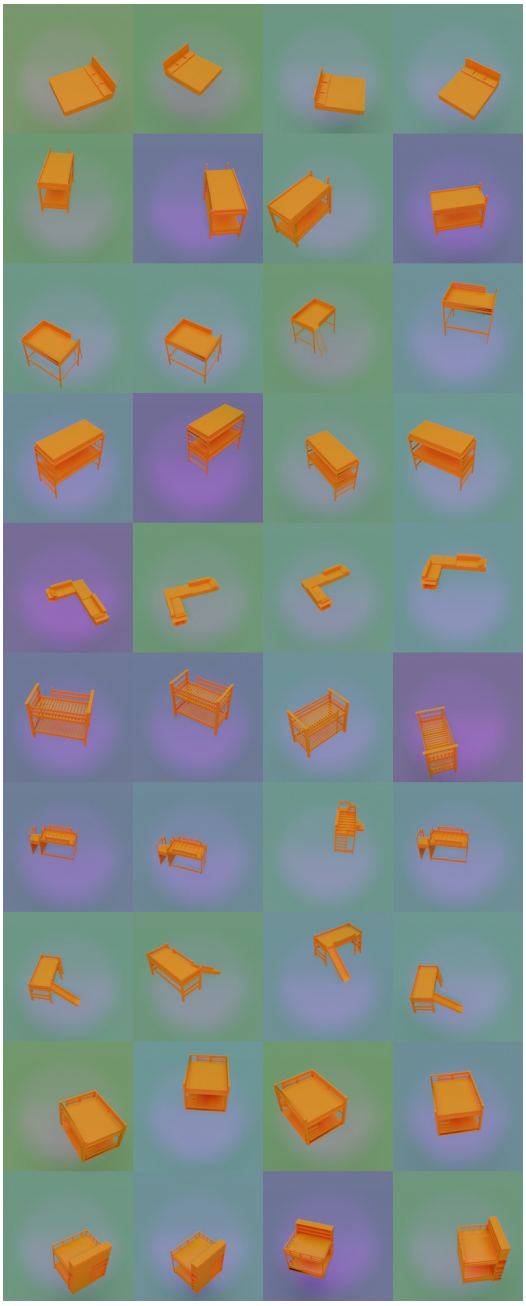


Figure 4: Example images generated for our 3DIEBench experiments. Each row shows four renderings of the same object model with different positions, orientations, and floor hues.

A.3 Training

A.3.1 Datasets with ground-truth latents

On Shape3D, MPI3D, 3DIdent, 3DIEBench, and the synthetic letter dataset, all methods are trained for 500 epochs with a ResNet-18 backbone and an MLP projection head with three hidden layers of width 2048, PReLU activations, and output dimension d. For each experimental setting, all methods use the same output dimension $d = 2m+1$, where m is the number of rotation planes in CRED. CRED uses the additional coordinate as its residual component. We use a batch size of 500 and Adam ($\text{lr} = 10^{-4}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$). Method-specific hyperparameters are detailed below.

SimCLR. We set the InfoNCE temperature to $\tau = 0.5$ for all experiments.

CARE. We use $\tau = 0.5$ in the InfoNCE loss and $\lambda_{\text{equiv}} = 10$. The number of equivariance chunks is set to half the batch size.

EquiMod. We use $\tau = 0.5$ for InfoNCE loss and $\lambda_{\text{equimod}} = 1$. The predictor head mirrors the projection head architecture.

IP-IRM. We use $\tau = 0.5$ for the InfoNCE loss and set $\lambda_1 = \lambda_2 = 0.2$. We update the data partition every 30 epochs of representation learning, running each partition-optimization phase for 50 epochs.

Eastwood et al. (2023). We use $\tau = 0.5$ and set each embedding subspace to two dimensions.

Ours. We adopt a two-stage schedule. Stage 1: SimCLR pretraining for 100 epochs with $\tau = 0.5$ (skipped on the single-class dataset **3DIdent**). Stage 2: we jointly train the feature extractor f and angle networks $G = \{g_i\}_i$ for the remaining 400 epochs (or 500 epochs on **3DIdent** when Stage 1 is skipped). Each g_i is an MLP with three hidden layers of width 2048, PReLU activations, LayerNorm, and one output per assigned rotation plane. We set the subspace radius to $\omega = 0.05$, with a one-dimensional residual component. We set $\lambda_{\text{radius}} = \lambda_{\text{theta}} = 1$ on Shape3D, MPI3D, 3DIdent, and 3DIEBench.

A.3.2 Datasets without ground-truth latents

We train all models for 800 epochs on MNIST and CIFAR-10 using a ResNet-18 backbone, and for 1500 epochs on CIFAR-100 using a ResNet-50 backbone. All models use an MLP projection head with three hidden layers of width 2048, ReLU activations, and output dimension d. We use a batch size of 500. The baselines and CRED's pretraining stage use LARS with a 10-epoch linear warm-up followed by cosine learning-rate decay and weight decay 10^{-6}. The LARS learning rate is 2.0 on MNIST and 4.0 on CIFAR-10 and CIFAR-100. CRED uses Adam in Stage 2, as specified below. The feature dimension is 256 on MNIST and 64 on CIFAR-10 and CIFAR-100.

Augmentations and subspace assignments in CRED. *MNIST.* We assign each of the seven augmentations in Table 4 to its own 2D subspace, giving 14 equivariant dimensions and 242 residual dimensions.

CIFAR-10 and CIFAR-100. We assign spatial and appearance augmentations to separate sets of five 2D subspaces (Table 5). Augmentations within each group share the same set, giving 20 equivariant dimensions and 44 residual dimensions.

SimCLR. We set the InfoNCE temperature to $\tau = 0.5$ on all three datasets.

CARE. We use $\tau = 0.5$ in the InfoNCE loss and $\lambda_{\text{equiv}} = 0.01$ on all three datasets. The number of equivariance chunks is set to be 4.

EquiMod. We use $\tau = 0.5$ for the InfoNCE loss and $\tau' = 0.2$ for EquiMod's contrastive equivariance loss, with $\lambda_{\text{equimod}} = 1$. The predictor head mirrors the projection head architecture.

Table 4: Augmentations and subspace assignments for CRED on MNIST.

Augmentation	Probability	Parameters	Subspaces
Scaling	0.80	Scale range $[0.90, 1.10]$	1 exclusive
Rotation	0.80	Angle range $[-12°, 12°]$	1 exclusive
Shear	0.80	Angle range $[-8°, 8°]$	1 exclusive
Dilation	0.50	Kernel size 3	1 exclusive
Erosion	0.50	Kernel size 3	1 exclusive
x-translation	0.90	Range $[-3, 3]$	1 exclusive
y-translation	0.90	Range $[-3, 3]$	1 exclusive

Table 5: Augmentations and subspace assignments for CRED on CIFAR-10 and CIFAR-100. p denotes the application probability.

Group	Augmentation	p	Parameters	Subspaces
Spatial	Random resized crop	1.00	Size 32×32; scale $[0.2, 1.0]$; aspect ratio $[3/4, 4/3]$	5 shared
	Horizontal flip	0.50	—	
Appearance	Color jitter	0.80	Brightness/contrast/saturation = 0.4; hue = 0.1	5 shared
	Random grayscale	0.20	—	

Ours. On MNIST and CIFAR-10, we use 500 epochs of LARS pretraining followed by 300 epochs with Adam. On CIFAR-100, we use 1000 epochs of LARS pretraining followed by 500 epochs with Adam. These schedules match the baseline epoch budget on each dataset. In the second stage, Adam uses a 10-epoch linear warm-up, $\beta_1 = 0.9$, and $\beta_2 = 0.999$. The learning rate is 10^{-4} on MNIST and CIFAR-10 and 2×10^{-4} on CIFAR-100. The subspace radius is $\omega = 0.1$ on all three datasets. We set $\lambda_{\text{radius}} = \lambda_{\text{theta}} = 0.1$ on MNIST and CIFAR-10 and $\lambda_{\text{radius}} = \lambda_{\text{theta}} = 0.15$ on CIFAR-100.

A.4 Evaluation

A.4.1 Clustering and linear evaluation protocols

For clustering, we use spherical k-means on CIFAR-10 and CIFAR-100 and agglomerative clustering on MNIST. We set the number of clusters to the known number of classes in each dataset. For agglomerative clustering, merging stops when this number of clusters remains.

Following the standard linear evaluation protocol (Chen et al., 2020a; He et al., 2020), we discard the projection head and train a linear classifier on the frozen backbone features. We train the classifier for 100 epochs with a batch size of 250, using SGD with Nesterov momentum 0.9 and no weight decay. The learning rate starts at 0.2 and follows a cosine decay schedule. We use the same protocol for all methods.

A.4.2 DCI Evaluation

For datasets with ground-truth latents, we evaluate with **DCI** (Eastwood & Williams, 2018), which summarizes three properties of a representation: *disentanglement*, *completeness*, and *informativeness*. Following Eastwood & Williams (2018), we train a supervised predictor from the evaluated representation to the ground-truth factors and extract a nonnegative *importance matrix* R, where R_{ij} measures the contribution of input coordinate i to predicting factor j. In our implementation, we use either a linear regressor or a random forest regressor. We take $R_{ij} = |w_{ij}|$ (absolute regression weights) for linear regression and use the mean decrease in impurity for random forest regression (Breiman, 2001).

For CRED's RAs + CI representation, we treat the one-hot cluster indicator as a single representation unit. Before normalization, we sum the corresponding rows of R into one row and compute disentanglement and completeness from the resulting matrix.

For disentanglement, we normalize each row of the resulting importance matrix over ground-truth factors; for completeness, we normalize each column over representation units. The entropies are normalized by the logarithms of the numbers of ground-truth factors and representation units, respectively. Disentanglement is computed for each representation unit and completeness for each ground-truth factor, in both cases as one minus the corresponding normalized entropy. The scores are then aggregated using the standard DCI weights. Informativeness is measured by the mean test-set R^2 across ground-truth factors, using either linear regression or random forest regression.

B Additional Results and Visualizations

B.1 Additional Visualizations of Figure 1

Settings. We construct a toy dataset of three uppercase letters ("R", "E", "D") rendered as 96×96 grayscale images. For each letter, we generate images at angles $\{-165°, -150°, \ldots, 180°\}$, sampling a full rotation cycle in 15° increments. The only training augmentation is a 15° rotation implemented with bilinear interpolation. The InfoNCE temperature is set to $\tau = 0.5$.

Following the alignment and uniformity view of contrastive learning (Wang & Isola, 2020), we vary the relative weights of the invariance, uniformity, and equivariance terms in CARE (Gupta et al., 2024) to obtain the configurations shown in Figure 1.

Extension: non-cyclic setting. We repeat the visualization using the same three input images, with rotation angles restricted to $[-120°, 120°]$ in 15° increments. Figure 5 shows the resulting feature trajectories.

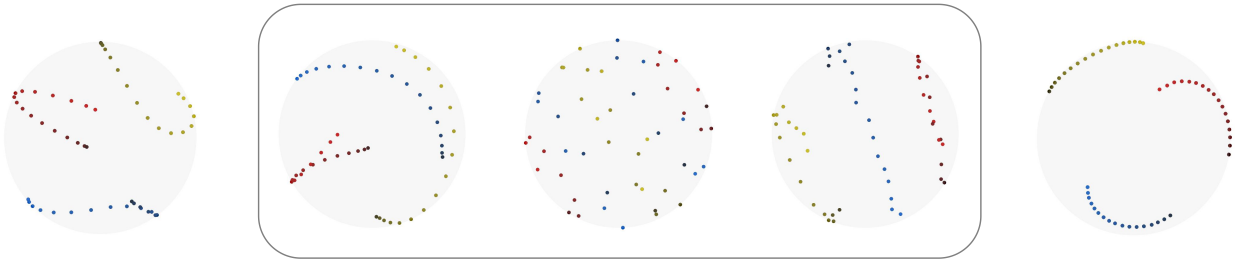


Figure 5: Feature trajectories traced out as each of three input images rotates over $[-120°, 120°]$. Each color corresponds to one input image. SimCLR produces irregular within-cluster trajectories. Across the shown loss-weight settings, CARE produces irregular trajectories or circular arcs in approximately parallel planes. CRED forms circular arcs in planes whose orientations vary across the three clusters.

B.2 Additional Results under Orthogonal Augmentations

Table 6 complements the random forest evaluation in the main text with linear regression. We focus on disentanglement and completeness, which assess how factors are organized across representation units; informativeness provides a complementary measure of factor predictability under the chosen regressor. With RAs + CI, CRED achieves D and C scores of at least 0.95 and an I score of 1.00 across all four datasets. On Shape3D, SimCLR achieves a D score of 1.00, but its C and I scores are only 0.10 and 0.17, respectively, illustrating why a high D score alone is insufficient to assess representation quality.

Figures 6 and 7 compare the feature distributions learned by SimCLR, CARE, and CRED. CRED's feature projections in Figures 8 and 9 show that varying one ground-truth style factor mainly changes feature angles in its corresponding rotation subspace.

Table 6: DCI scores using linear regression on four datasets. D, C, and I denote disentanglement, completeness, and informativeness. For CRED, raw denotes the original learned features, and RAs + CI denotes the post-processed representation.

Method	**MPI3D**			**Shape3D**			**3DIdent**			**3DIEBench**		
	D	C	I	D	C	I	D	C	I	D	C	I
SimCLR	0.23	0.08	0.14	**1.00**	0.10	0.17	0.29	0.08	0.31	0.23	0.10	0.17
EquiMod	0.55	0.12	0.20	0.85	0.17	0.48	0.23	0.08	0.31	0.74	0.12	0.29
CARE	0.24	0.11	0.17	0.84	0.07	0.18	0.24	0.08	0.33	0.17	0.10	0.31
IP-IRM	0.65	0.15	0.35	0.36	0.11	0.13	0.20	0.07	0.43	0.53	0.12	0.25
Eastwood et al. (2023)	**0.95**	0.70	0.41	0.96	0.82	0.69	0.88	0.69	0.71	0.95	0.50	0.28
CRED (raw)	0.63	0.51	**1.00**	0.63	0.49	**1.00**	0.60	0.47	**1.00**	0.11	0.07	0.31
CRED (RAs + CI)	**0.95**	**0.95**	**1.00**	0.99	**0.98**	**1.00**	**0.98**	**0.98**	**1.00**	**0.99**	**0.99**	**1.00**

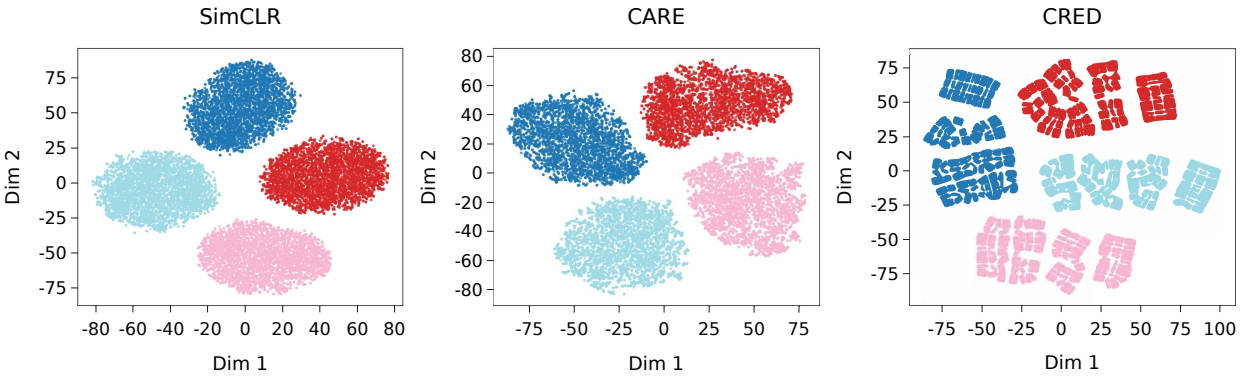


Figure 6: t-SNE visualization of learned features on **Shape3D**.

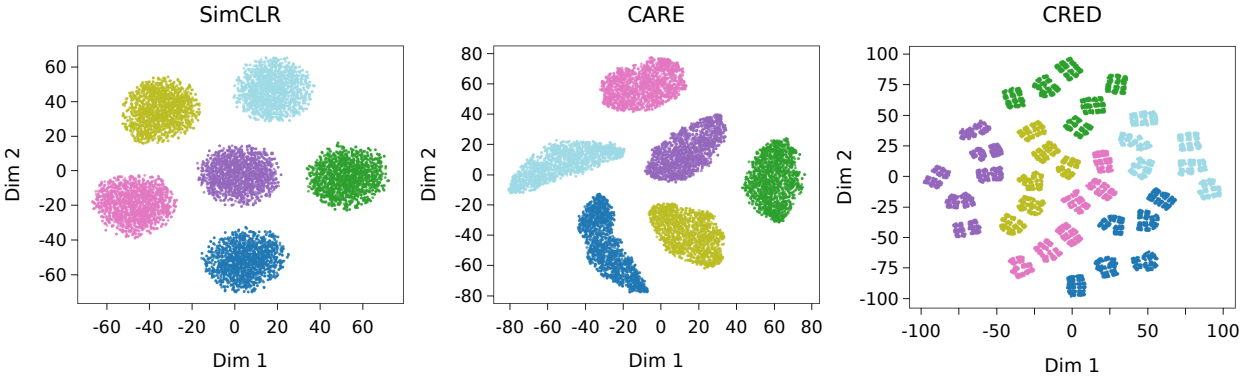


Figure 7: t-SNE visualization of learned features on **MPI3D**.

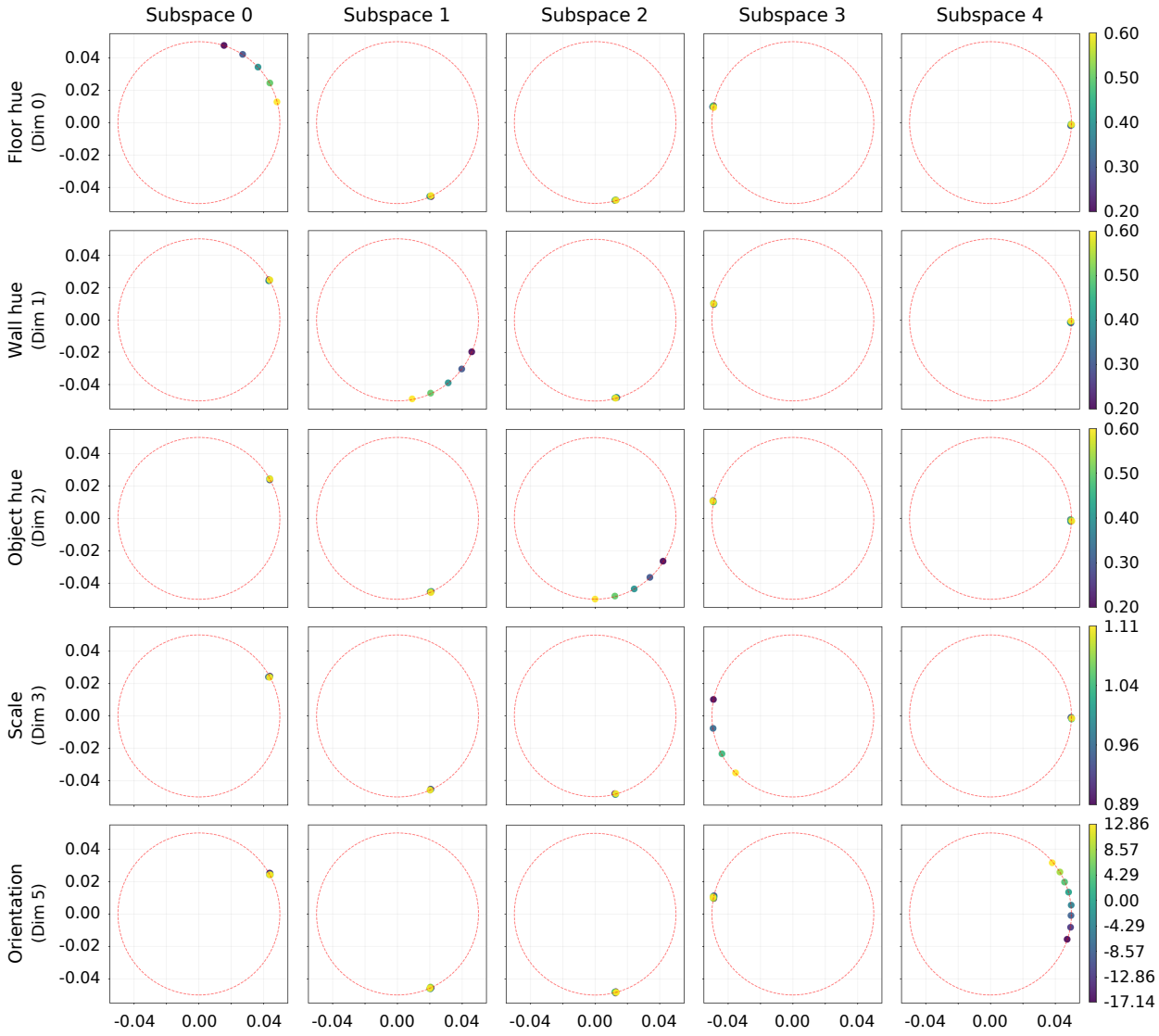


Figure 8: CRED feature projections on **Shape3D**. Starting from a reference sample in one cluster, each row varies one ground-truth style factor: floor hue (Dim 0), wall hue (Dim 1), object hue (Dim 2), scale (Dim 3), or orientation (Dim 5). Columns show the aligned rotation subspaces. Colors indicate the value of the factor varied in each row. Each factor mainly changes the feature angle in its corresponding subspace.

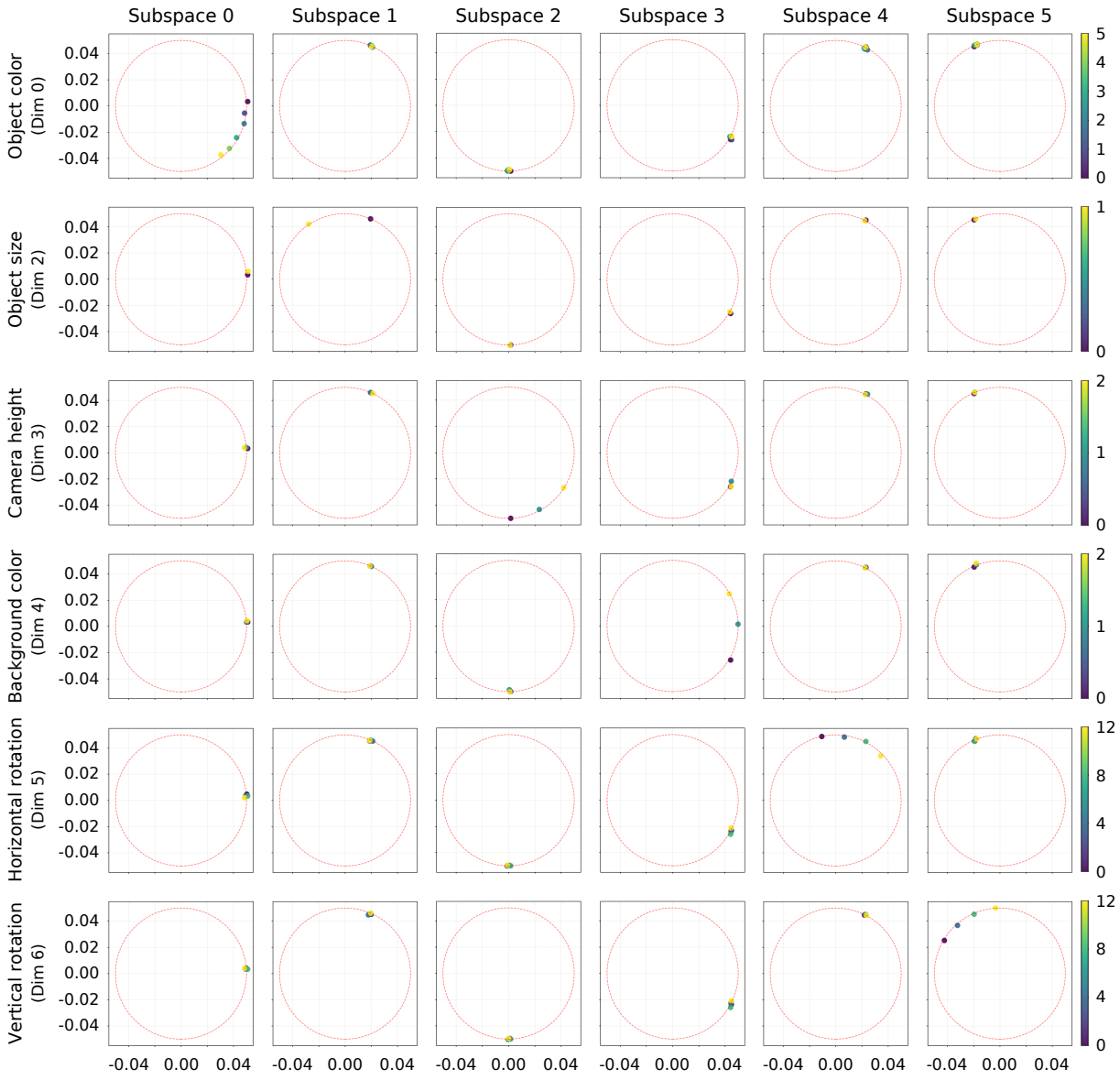


Figure 9: CRED feature projections on **MPI3D**. Starting from a reference sample in one cluster, each row varies one ground-truth style factor: object color (Dim 0), object size (Dim 2), camera height (Dim 3), background color (Dim 4), horizontal rotation (Dim 5), or vertical rotation (Dim 6). Columns show the aligned rotation subspaces. Colors indicate the value of the factor varied in each row. Each factor mainly changes the feature angle in its corresponding subspace.

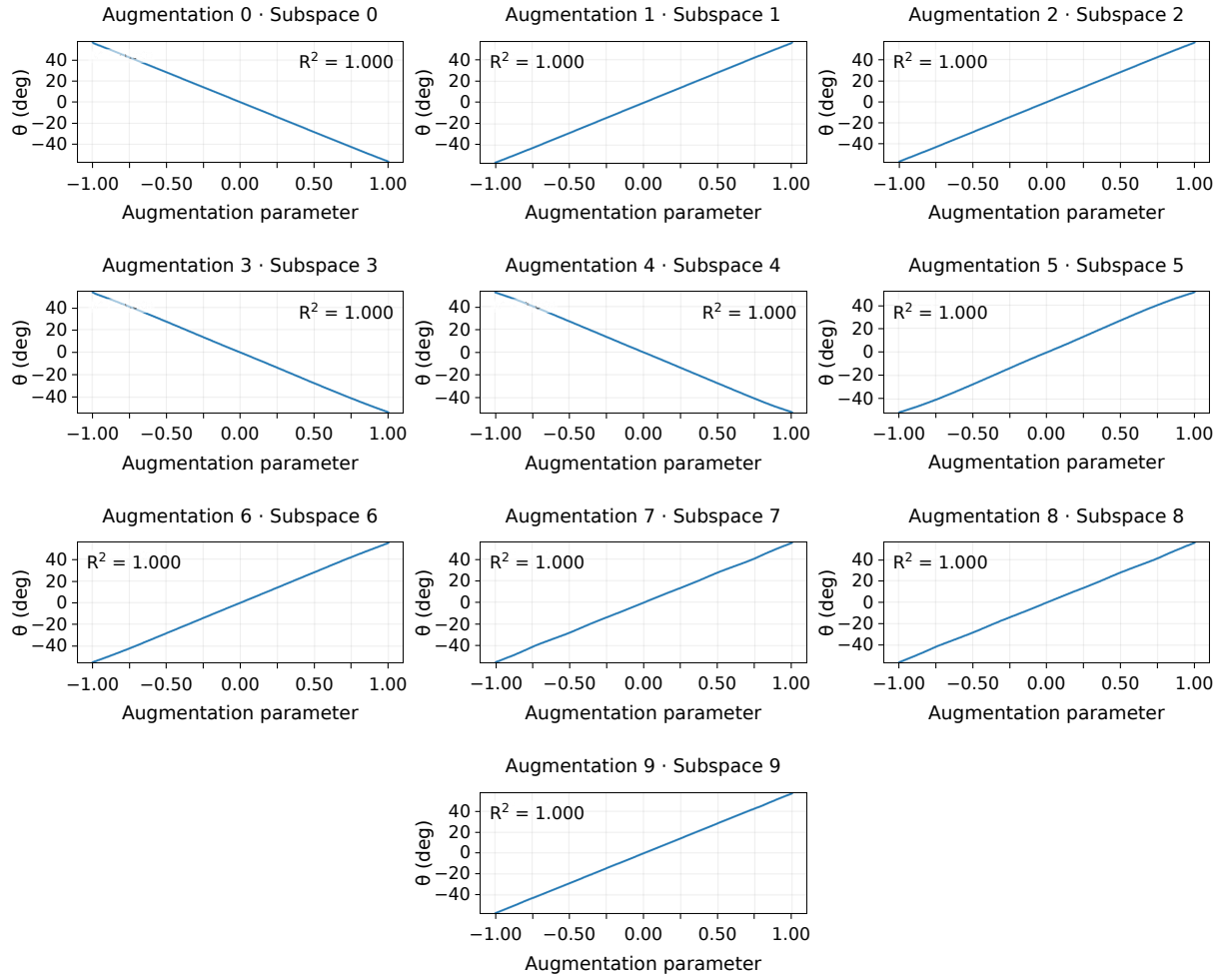


Figure 10: CRED's predicted rotation angles versus augmentation parameters on **3DIdent**, evaluated at 1,000 randomly sampled parameter values. Each panel corresponds to an augmentation and its assigned rotation subspace. The predicted angles vary approximately linearly over the plotted parameter ranges.

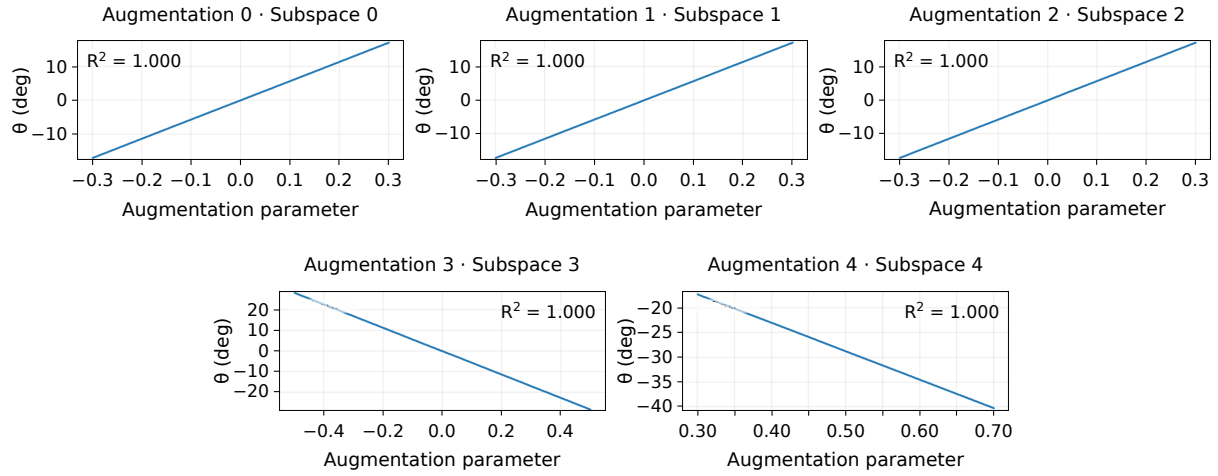


Figure 11: CRED's predicted rotation angles versus augmentation parameters on **3DIEBench**, evaluated at 1,000 randomly sampled parameter values. Each panel corresponds to an augmentation and its assigned rotation subspace. The predicted angles vary approximately linearly over the plotted parameter ranges.

B.3 Ablation Studies

We ablate SimCLR pretraining and the two regularizers on Shape3D. Without SimCLR pretraining, we initialize the rotation centers using k-means on randomly initialized features. In this setting, informativeness drops substantially across all evaluated factors. With pretraining retained, removing either $\mathcal{L}_{\text{radius}}$ or $\mathcal{L}_{\text{theta}}$ leads to collapse in the rotation subspaces (Figure 12). Informativeness under linear regression drops to 0.0000 for all five style factors, while the class score remains 1.0000 in both variants (Table 7).

Table 7: Ablation study on **Shape3D**: per-factor informativeness under linear regression. All variants use RAs + CI post-processing.

Method	Floor Hue	Wall Hue	Object Hue	Scale	Orientation	Class
w/o pretraining	0.0648	0.2712	0.0177	0.0172	0.0010	0.0034
w/o $\mathcal{L}_{\text{radius}}$	0.0000	0.0000	0.0000	0.0000	0.0000	**1.0000**
w/o $\mathcal{L}_{\text{theta}}$	0.0000	0.0000	0.0000	0.0000	0.0000	**1.0000**
CRED (RAs + CI)	**0.9995**	**0.9995**	**0.9994**	**0.9984**	**0.9994**	**1.0000**

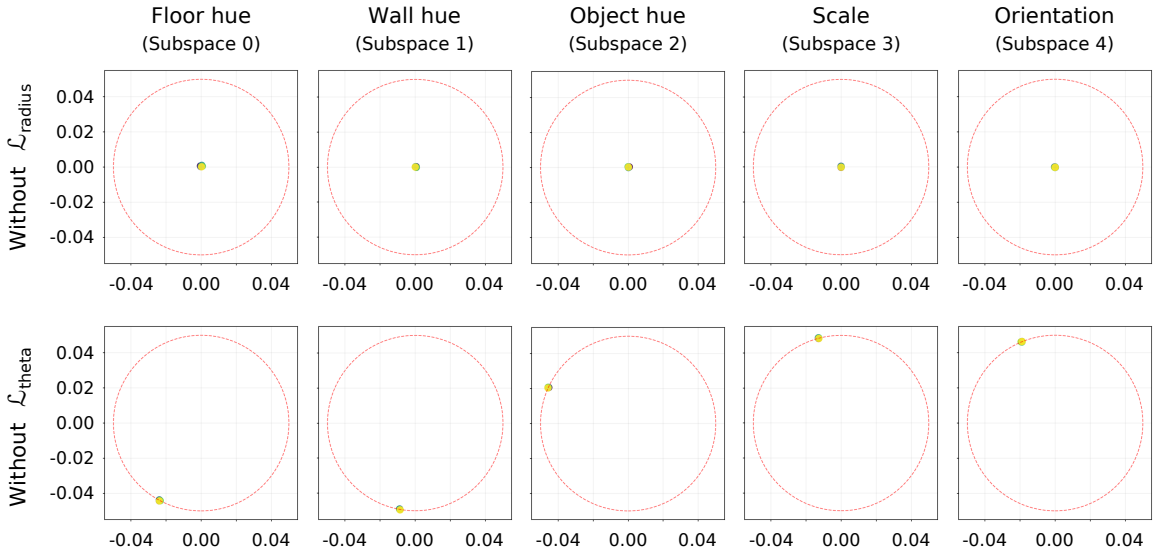


Figure 12: Shape3D feature projections from one cluster. Each column shows projections in the aligned rotation subspace corresponding to one ground-truth style factor. That factor varies while the others remain fixed. Colors indicate factor values. Top: without $\mathcal{L}_{\text{radius}}$, projections collapse near the origin. Bottom: without $\mathcal{L}_{\text{theta}}$, projections retain nonzero radii but collapse to a single point as the factor varies.

Uniformity loss for rotation centers (optional). To encourage a more uniform distribution of rotation centers on the unit sphere, we maximize their average pairwise Euclidean distance using the repulsion loss

$$\mathcal{L}_{\text{uniform}} = -\mathbb{E}_{i \neq j} \left\| r_i - r_j \right\|_2, \tag{23}$$

where the expectation is taken uniformly over distinct center pairs. We initialize the learnable centers from the normalized Stage 1 cluster centroids and add this loss to the Stage 2 objective.

For three unit-norm centers, every global minimizer of this loss forms an equilateral triangle with pairwise distance $\sqrt{3} \approx 1.732$, the largest achievable minimum separation (Tammes, 1930). Under the same settings as Figure 5, the learnable centers approach this optimal spacing, while the fixed centers have a minimum pairwise distance of 1.719.

B.4 Training Loss Convergence

Figure 13 reports the training behavior of the three losses on 3DIEBench, complementing the compatibility analysis in Appendix D.3.

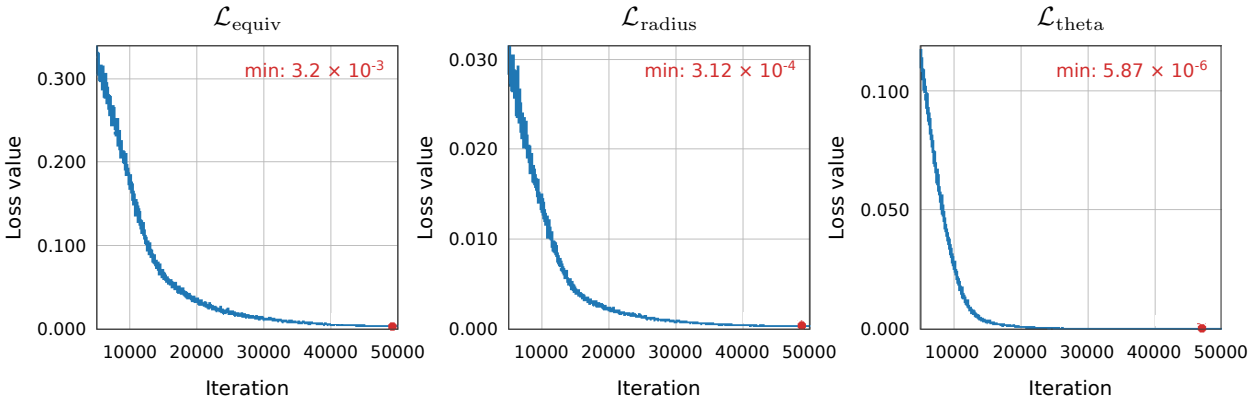


Figure 13: Training curves of the three CRED losses on 3DIEBench (Garrido et al., 2023b) over 50,000 steps. All three losses reach near-zero values in the reported run.

B.5 Compositional OOD Generalization

On Shape3D and MPI3D, we hold out selected joint factor combinations before training CRED, while retaining every individual factor value in the in-distribution training set. After training, we freeze the feature extractor f and use only in-distribution data to determine the reference angles for CRED's standard post-processing and fit one linear regressor per factor from its corresponding post-processed coordinate. The resulting mappings are applied unchanged to the held-out combinations for out-of-distribution (OOD) evaluation. Table 8 summarizes performance on these held-out combinations.

Table 8: Compositional OOD generalization of CRED on held-out factor combinations. Scores are mean R^2 across factors, evaluated by linear regression on the post-processed representation.

Dataset	**Shape3D**	**MPI3D**
Mean OOD R^2	0.99543	0.98653

B.6 Additional Results under Complex Augmentations

Figures 14 and 15 illustrate the simple composite setting, showing how xy-translation is separated into x-position and y-position changes. Figure 16 complements Figure 3 with projections across three clusters, showing the separation of position and orientation changes in the anchored complex setting.

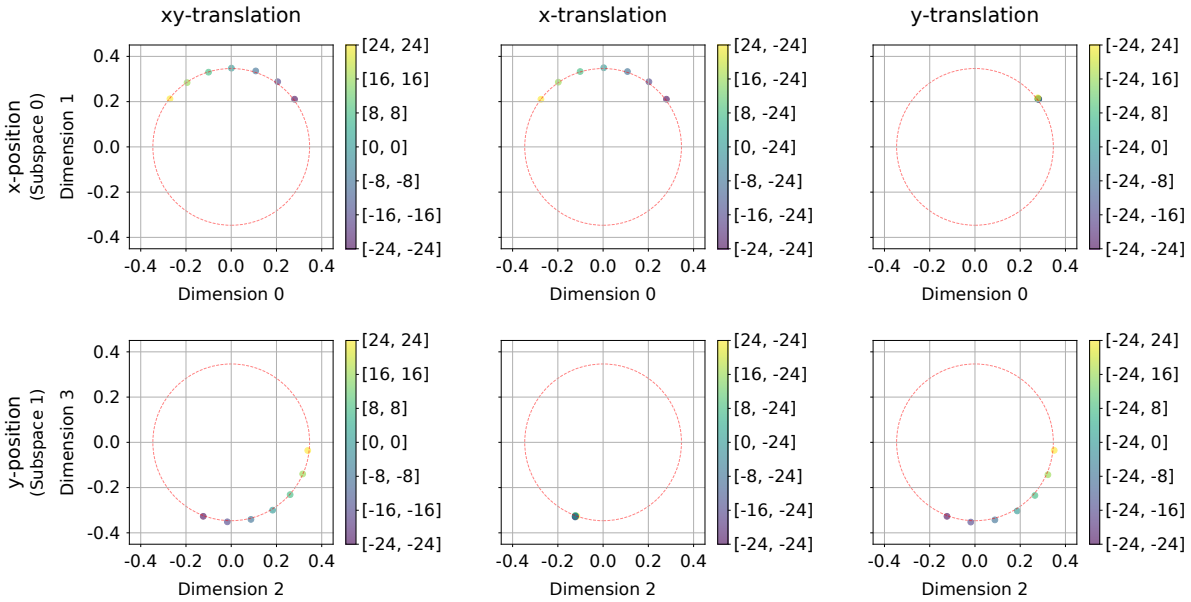


Figure 14: CRED feature projections from one cluster. Columns show xy-translation, x-translation, and y-translation, from left to right; rows show the feature planes for x-position and y-position, from top to bottom. Colorbar labels give (x-position, y-position). CRED disentangles the changes in x-position and y-position induced by xy-translation into separate feature planes. The feature changes in these planes align with those induced by x-translation and y-translation, respectively.

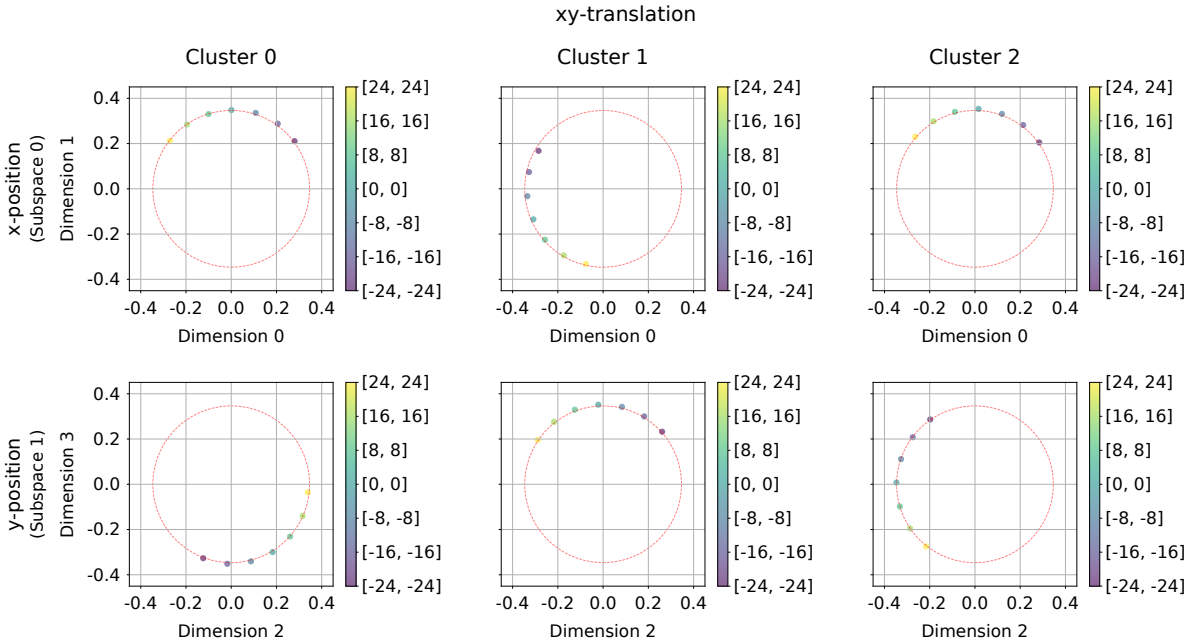


Figure 15: Full results for the experiment in Figure 14, showing feature projections across three clusters. This page shows xy-translation. Columns correspond to clusters; rows show the feature planes for x-position and y-position, from top to bottom. Colorbar labels give (x-position, y-position). CRED disentangles the changes in x-position and y-position into the planes assigned to x-translation and y-translation, respectively. The following pages show x-translation and y-translation.

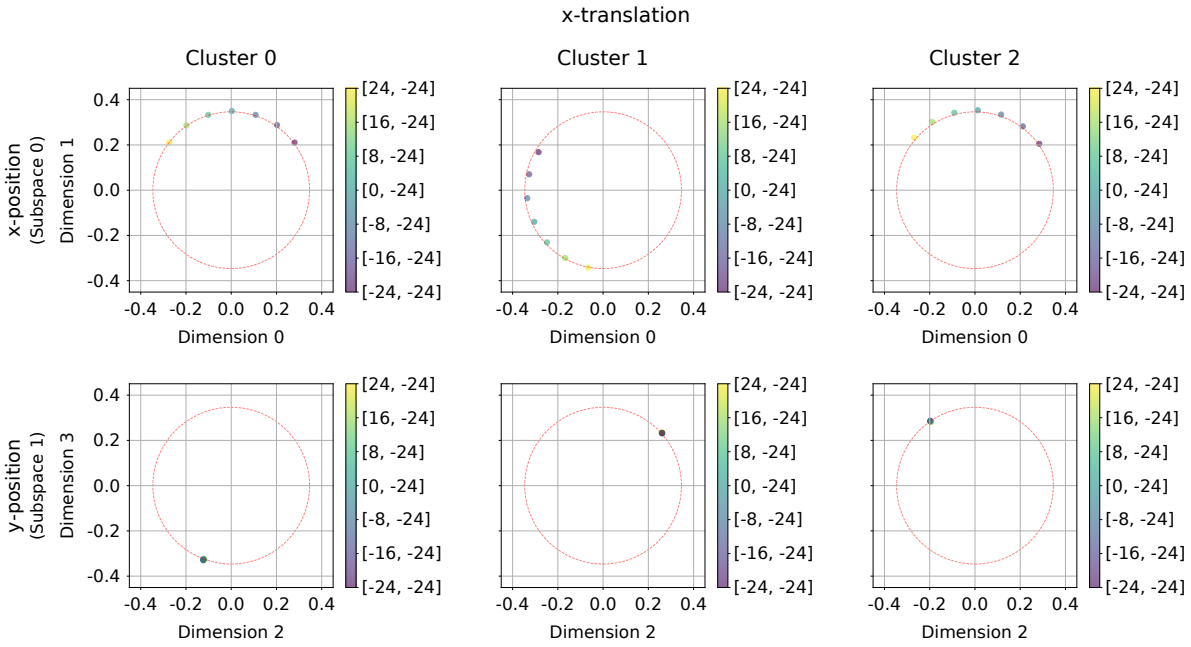


Figure 15: Continued: projections under x-translation, with the same cluster and rotation-plane layout.

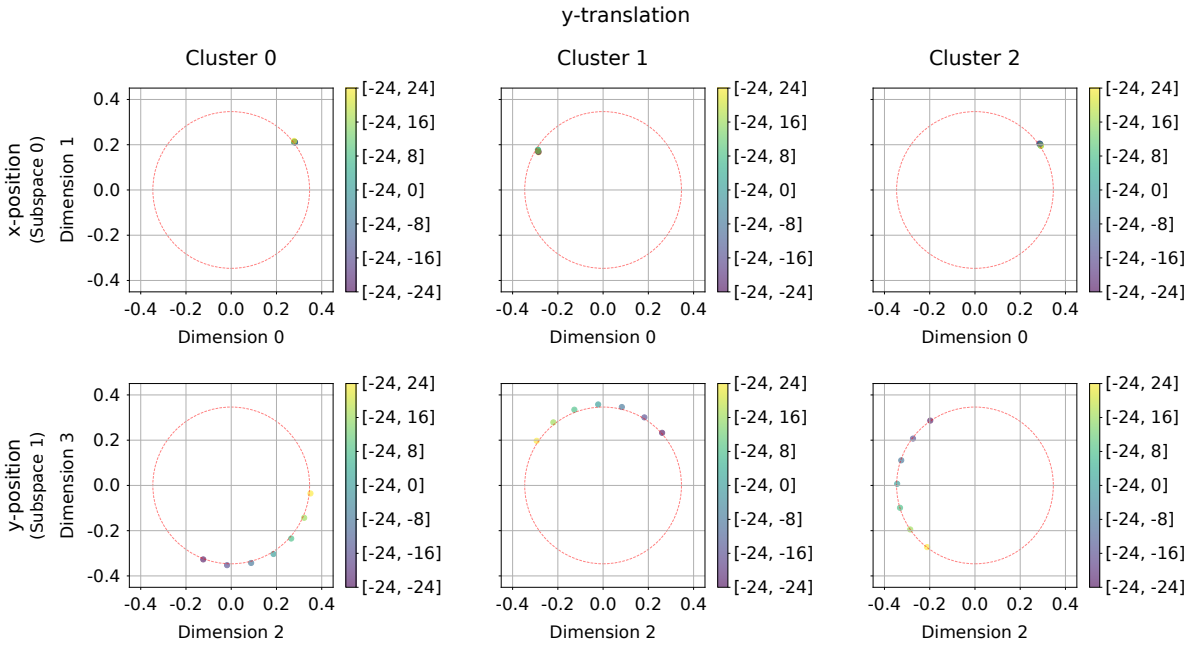


Figure 15: Continued: projections under y-translation, with the same cluster and rotation-plane layout.

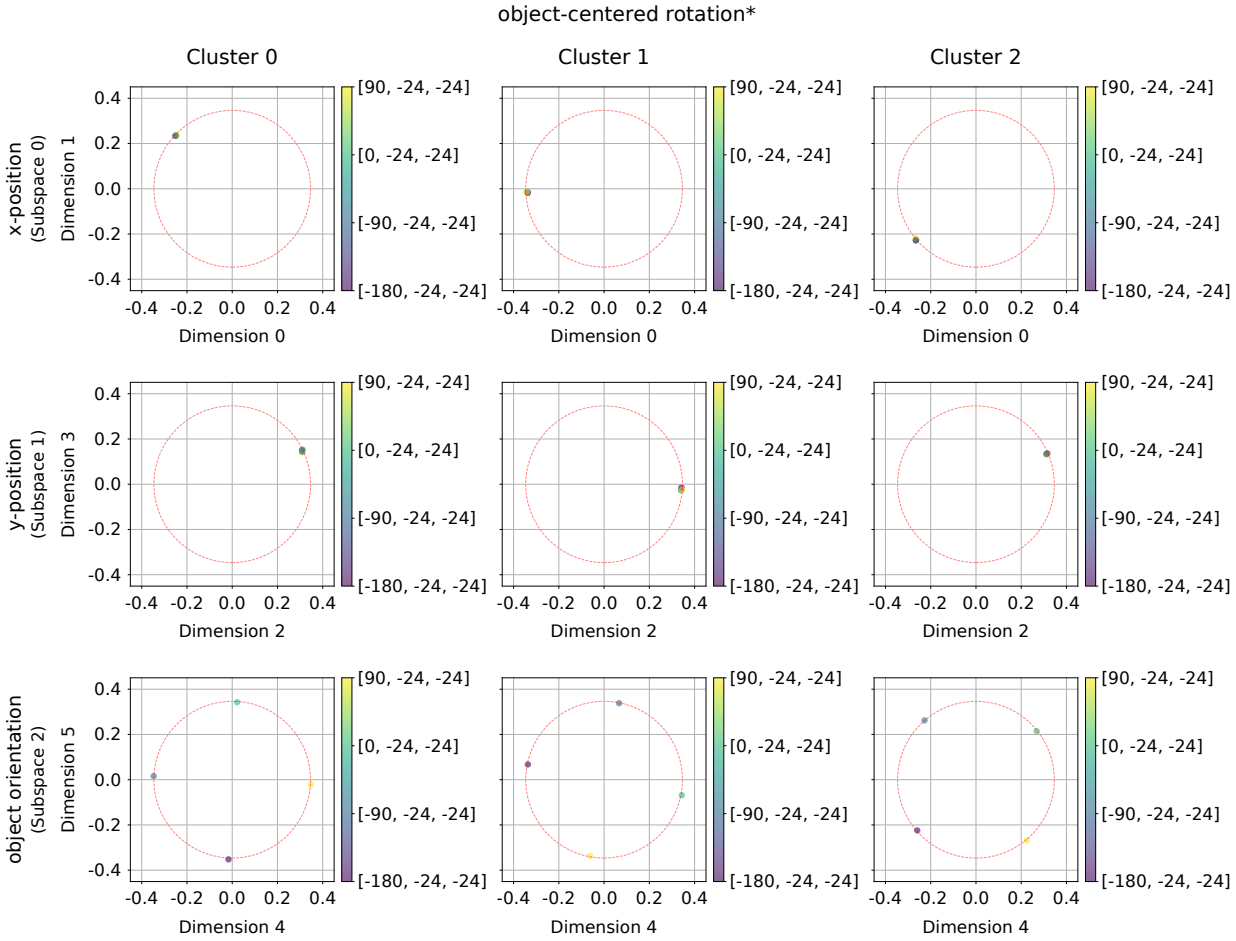


Figure 16: Full results for the experiment in Figure 3, showing feature projections across three clusters. Training uses x-translation, y-translation, and image-centered rotation. This page shows projections under object-centered rotation. Columns correspond to clusters; rows show planes for x-position, y-position, and object orientation, from top to bottom. Colorbar labels give (object orientation, x-position, y-position). Object-centered rotation acts about the object's own center; $*$ indicates that this rotation is virtual and is not supplied as a separate training augmentation. The following pages show projections under x-translation and y-translation. CRED disentangles changes in x-position, y-position, and object orientation into separate feature planes within each cluster.

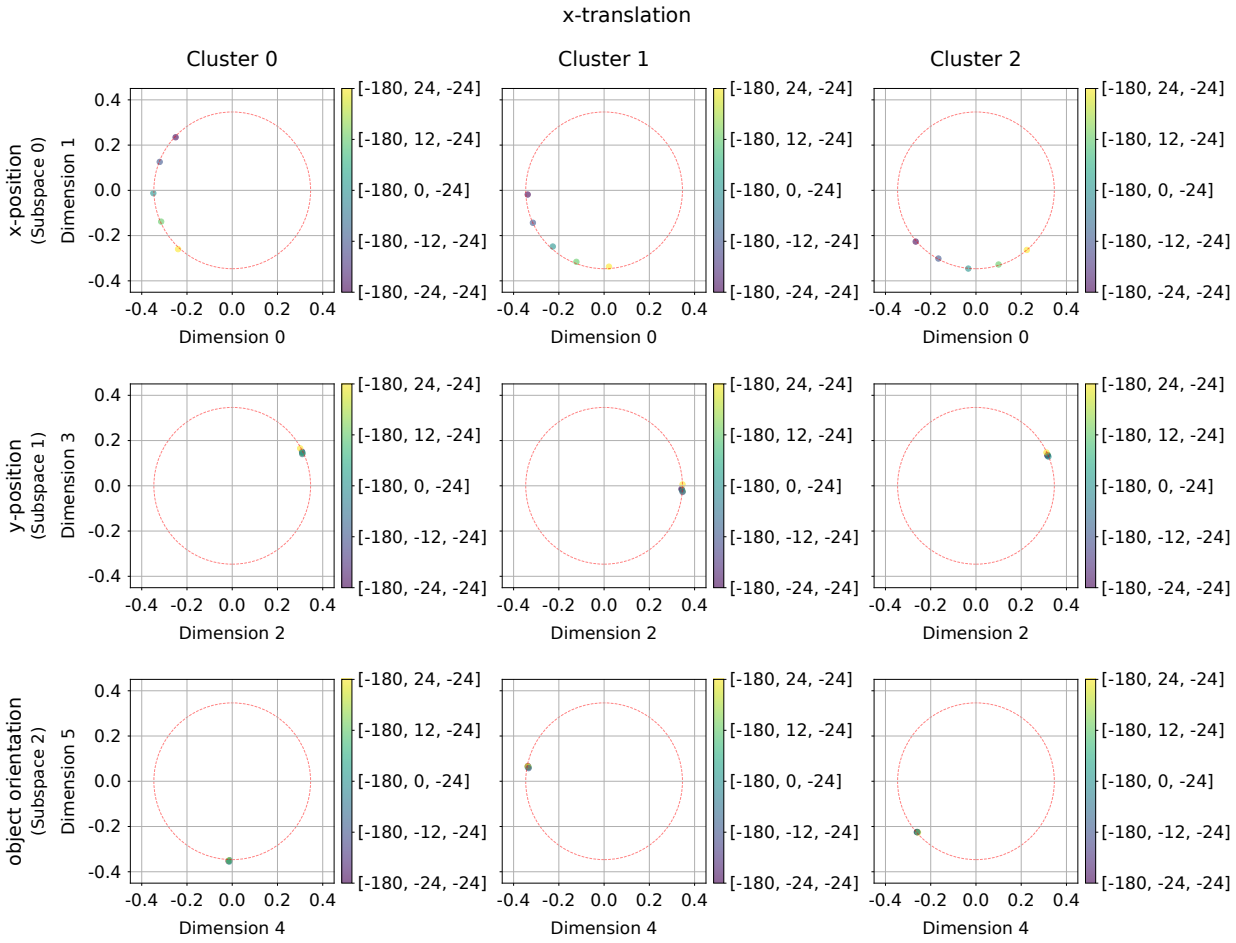


Figure 16: Continued: projections under x-translation, with the same cluster and rotation-plane layout.

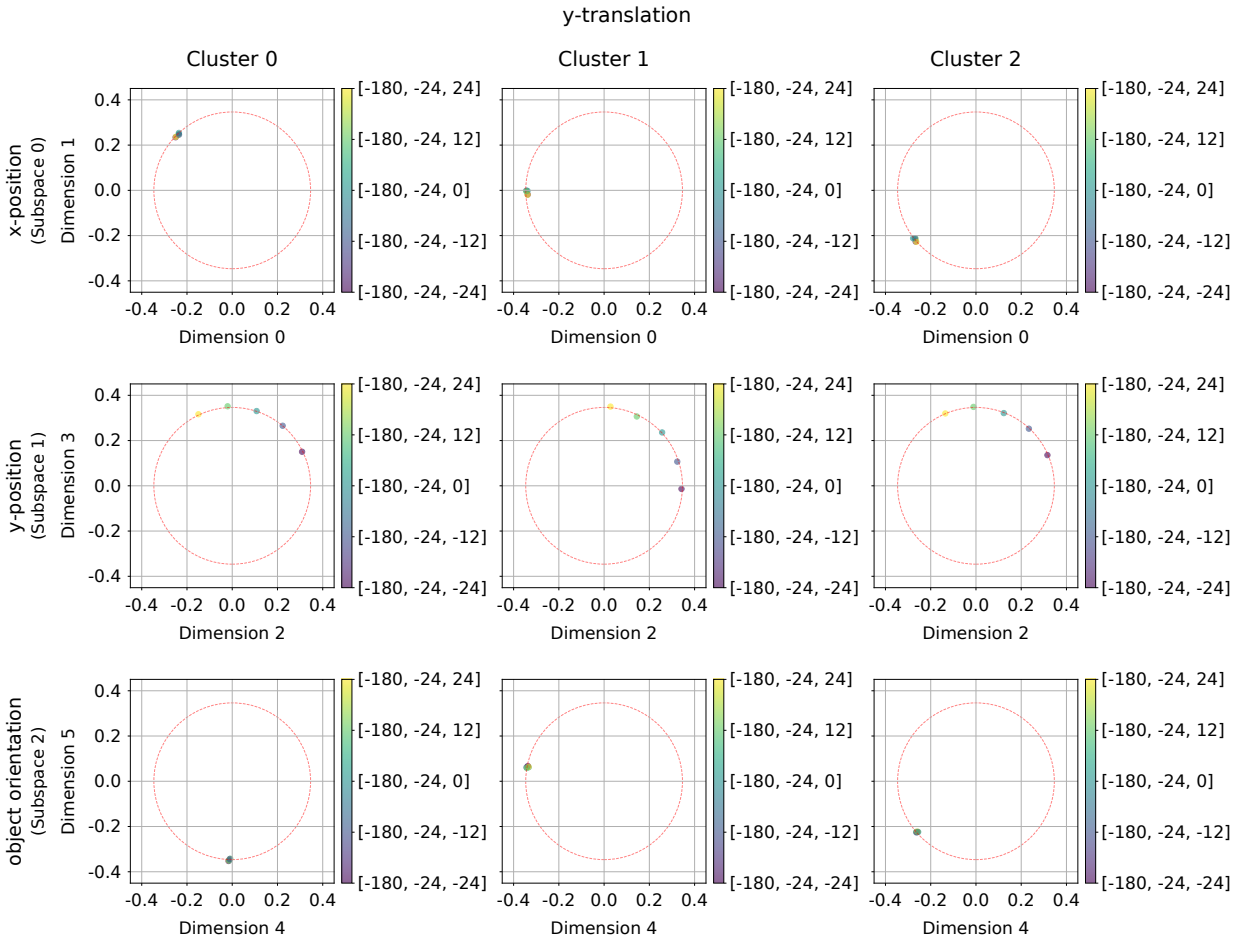


Figure 16: Continued: projections under y-translation, with the same cluster and rotation-plane layout.

C Extended Related Work

Self-supervised Representation Learning. Contrastive and non-contrastive objectives are widely used in self-supervised visual representation learning. Most contrastive methods (Yeh et al., 2022; He et al., 2020; Chen et al., 2020b;a) use an InfoNCE-type loss (van den Oord et al., 2018). HaoChen et al. (2021) study a spectral contrastive loss. Clustering-based methods use cluster assignments or prototypes to guide representation learning (Caron et al., 2018; 2020; Li et al., 2021). Non-contrastive methods learn without explicit negative pairs, using architectural design (Grill et al., 2020; Chen & He, 2021), feature whitening (Ermolov et al., 2021), cross-correlation objectives (Zbontar et al., 2021), or variance and covariance regularization (Bardes et al., 2022a;b). Garrido et al. (2023a) establish theoretical connections between contrastive and covariance-based non-contrastive objectives. In a sparse-coding setting, Pokle et al. (2022) identify non-collapsed global optima of a non-contrastive objective that fail to recover the ground-truth features. Our first stage builds on SimCLR (Chen et al., 2020a). Theoretical analyses characterize contrastive learning through positive-pair alignment and feature uniformity on the hypersphere (Wang & Isola, 2020), and provide guarantees for preserving data clusters under assumptions on augmentation structure and the encoder class (Wang et al., 2022; Parulekar et al., 2023).

Group-equivariant Representation Learning. Group-equivariant representation learning aims to learn feature maps whose response to structured transformations of the input follows a group representation, i.e., $f(T_{\mathsf{g}}(x)) = \rho(\mathsf{g})f(x)$ for a group element $\mathsf{g} \in \mathcal{G}$. Classical group-equivariant CNNs instantiate this idea by fixing both the group $\mathcal{G}$ and its action on the input domain, then designing convolutional layers so that feature channels transform according to a prescribed representation ρ (Cohen & Welling, 2016; 2017; Cohen et al., 2018; Weiler et al., 2018; Weiler & Cesa, 2019; Finzi et al., 2020; Lengyel et al., 2023; Yang et al., 2025). Equivariance has also been incorporated into message-passing networks (Anderson et al., 2019; Satorras et al., 2021), spatial transformation modules (Tai et al., 2019), and self-attention models (Fuchs et al., 2020). These approaches enforce equivariance through architectural constraints tied to a specified group action. CRED instead learns cluster-dependent subspace rotational equivariance through a training objective, without requiring a group-equivariant backbone.

Disentanglement in Generative Models. Generative approaches to disentanglement include VAE-based methods (Higgins et al., 2017; Kim & Mnih, 2018; Chen et al., 2018; Li et al., 2019; Mathieu et al., 2019; Esmaeili et al., 2019), GAN-based methods (Chen et al., 2016; Jeon et al., 2021; Ojha et al., 2020), diffusion models (Yang et al., 2023; Wu & Zheng, 2024), and generative models with structured latent flows (Song et al., 2023). Related work also studies controllable latent-space editing in flow-matching models (Hu et al., 2024). Weakly supervised approaches exploit grouped observations with shared content (Bouchacourt et al., 2018; Hosoya, 2019), paired observations sharing some factors (Locatello et al., 2020), or pairwise similarity annotations (Chen & Batmanghelich, 2020). Shu et al. (2020) analyze disentanglement guarantees under different forms of weak supervision. CRED adopts a self-supervised formulation that explicitly organizes augmentation-induced variations through rotations in mutually orthogonal two-dimensional subspaces.

D Theoretical Analysis and Proofs

The proofs of the five main-text propositions are organized as follows:

- Proposition 1: exact equivariance (Appendix D.3). p. 39
- Proposition 2: weak disentanglement (Appendix D.4). p. 42
- Proposition 3: strong disentanglement (Appendix D.5). p. 54
- Proposition 4: simple composite augmentations (Appendix D.7). p. 60
- Proposition 5: anchored complex augmentations (Appendix D.8). p. 61

Appendices D.1 and D.2 provide the parallel-configuration analysis and theoretical preliminaries. Appendix D.6 establishes the post-processing guarantees.

D.1 Parallel Configurations under Global Rotational Equivariance

Let $\{C_1, \ldots, C_K\}$ be semantic clusters and $f : \mathcal{X} \to \mathbb{S}^{d-1}$ a feature extractor satisfying exact global rotational equivariance on the considered augmentation pairs:

$$f(a(x)) = M_a f(x), \qquad M_a \in \mathrm{SO}(d), \tag{24}$$

where each M_a is shared across samples. We define the *shared variation subspace* as the smallest linear subspace containing all changes induced by the global rotation operators:

$$V := \mathrm{span}\big\{M_a v - v : \ a \in \mathcal{A}, \ v \in \mathbb{R}^d\big\}. \tag{25}$$

Here, v ranges over the ambient feature space. Under exact equivariance, every augmentation-induced feature difference lies in V, since

$$f(a(x)) - f(x) = (M_a - I)f(x) \in V.$$

For each cluster, we consider finite chains of pairs $(x, a(x))$ satisfying the above equivariance relation, with both points of every pair belonging to that cluster. Each pair may be traversed in either direction; this does not require an inverse augmentation to be available.

Proposition 6 (Parallel Affine Subspace Constraint)**.** *Suppose $0 < \dim V < d$, and any two samples in each cluster can be connected by such a chain. Then, for each cluster C_k, there exists $b_k \in V^\perp$ such that*

$$f(C_k) \subseteq (b_k + V) \cap \mathbb{S}^{d-1}. \tag{26}$$

These affine subspaces share the same direction space V, forming a parallel configuration*. Their offsets need not be distinct.*

Proof. For each considered augmentation pair, the identity above shows that the feature difference lies in V. Fix $x_k \in C_k$. Summing these differences, with the appropriate signs, along a connecting chain from x_k to any $x \in C_k$ yields $f(x) - f(x_k) \in V$. Thus, all features within the cluster have the same orthogonal projection onto $V^\perp$. Taking $b_k = P_{V^\perp} f(x_k)$ proves the result. □

Three-dimensional example. For nontrivial rotations about a common axis in three dimensions, V is the plane perpendicular to that axis. Proposition 6 therefore places augmentation-connected clusters in parallel planes. These planes intersect the sphere in circles, possibly degenerate, while the observed features may occupy only arcs or finitely many points. This geometry accounts for the parallel trajectories shown for CARE in Figures 1 and 5. The offsets b_k are fixed projections and need not equal the cluster centroids; the argument does not require augmentation-invariant sample means or within-cluster distributions.

Scope and relation to CRED. The condition $\dim V < d$ is essential for a nontrivial geometric restriction and is not guaranteed by arbitrary collections of global rotations. In CRED's aligned coordinates, the m rotation planes span a $2m$-dimensional subspace, with $d \geq 2m + 1$ leaving non-rotating coordinates. Cluster-dependent coordinate transformations allow the corresponding variation subspaces to have different orientations across clusters. This removes the common-orientation constraint characterized above.

D.2 Preliminaries

We introduce the definitions and assumptions used in the theoretical analysis.

Definition 1 (Augmentation Connectivity)**.** *Let $\mathcal{A}$ be a collection of augmentations acting on $\mathcal{X}$. Two samples x, y are adjacent if $a(x) = b(y)$ for some $a, b \in \mathcal{A}$. We write $x \sim_{\mathcal{A}} y$ if there exists a finite sequence $x = x_0, \ldots, x_n = y$ in which each consecutive pair is adjacent. A sequence of length zero is allowed. This defines an equivalence relation.*

A subset $D \subseteq \mathcal{X}$ is $\mathcal{A}$-connected if any two points in D can be joined by such a sequence contained in D. For latent points, $z \sim_{\mathcal{A}} z'$ means $\varphi(z) \sim_{\mathcal{A}} \varphi(z')$.

Assumption 1 (Intra-cluster Connectivity)**.** *Fix a collection of augmentations $\mathcal{A}$ acting on the observation space $\mathcal{X}$ and let $\sim_{\mathcal{A}}$ denote the augmentation-connectivity relation defined in Definition 1. We work in latent coordinates $z = (s, u)$, allowing an invertible reparameterization of the original generative factors. The coordinate change is absorbed into φ. Let $\mathcal{X}_{\text{train}} \subset \mathcal{X}$ be the training set, and write*

$$\mathcal{Z}_{\text{train}} := \varphi^{-1}(\mathcal{X}_{\text{train}}) \subseteq \mathcal{S}_{\mathcal{A}} \times \mathcal{U}_{\mathcal{A}},$$

where $\mathcal{S}_{\mathcal{A}}$ and $\mathcal{U}_{\mathcal{A}}$ are the style and cluster latent spaces, respectively.

1. ***Augmentation-wise factorization.*** *Each $a \in \mathcal{A}$ lifts to a map*

$$\tilde{a} : \mathcal{S}_{\mathcal{A}} \times \mathcal{U}_{\mathcal{A}} \longrightarrow \mathcal{S}_{\mathcal{A}} \times \mathcal{U}_{\mathcal{A}}, \qquad \tilde{a}(s, u) = (\tilde{s}, u),$$

   Thus, augmentations preserve the cluster coordinates u, while any induced change is confined to the style coordinates s.

2. ***Intra-cluster $\mathcal{A}$-connectivity.*** *Define the cluster partitions*

$$\mathcal{C}_{\mathcal{A}} := \big\{c_u := \big\{x \in \mathcal{X}_{\text{train}} \,:\, \varphi^{-1}(x) = (s, u),\ s \in \mathcal{S}_{\mathcal{A}}\big\}, \forall u \in \mathcal{U}_{\mathcal{A}}\big\}.$$

   We assume that every pair of samples in c_u can be joined by a finite chain of adjacent samples contained in c_u.

Because augmentations never modify the u-coordinate, no sequence of augmentations can connect points with different cluster coordinates, i.e., if $x \in c_u$ and $y \in c_{u'}$ with $u \neq u'$ then $x \not\sim_{\mathcal{A}} y$.

Definition 2 (Intrinsic cyclicity and observed range)**.** *Let $z^{(j)}$ denote a one-dimensional variation component. It is intrinsically cyclic with period $p > 0$ if coordinate values differing by p represent the same state, so its full domain is a circle. A non-cyclic component has an interval as its full domain, with distinct coordinate values representing distinct states.*

Even when a component is intrinsically cyclic, the data may cover only part of its period. For example, object orientation is cyclic, but orientations restricted to $[-45°, 45°]$ vary over an interval. Within this restricted range, orientation can therefore be analyzed using an interval coordinate.

Definition 3 (Cluster-wise axis-aligned representation)**.** *Let $f : \mathcal{X} \to \mathbb{S}^{d-1}$ be a unit-norm representation. Fix an augmentation set $\mathcal{A}$ that alters exactly d_{sty} style variation components and induces cluster partitions $\mathcal{C}_{\mathcal{A}}$ of $\mathcal{X}_{\text{train}}$. We say f is **cluster-wise axis-aligned** if there exist orthogonal matrices $\{H_c \in \mathrm{O}(d)\}_{c \in \mathcal{C}_{\mathcal{A}}}$ (one per cluster) satisfying the conditions below. For the aligned representation, write*

$$\tilde{f}(x) := H_{c(x)}\, f(x) \;=\; \big(\tilde{f}^{(1)}_{\text{sty}}(x), \ldots, \tilde{f}^{(d_{\text{sty}})}_{\text{sty}}(x), \tilde{f}_{\text{res}}(x)\big) \in \mathbb{S}^{d-1},$$

where $\tilde{f}^{(i)}_{\text{sty}}(x) \in \mathbb{R}^2$ is the feature block associated with the i-th rotation plane, and $\tilde{f}_{\text{res}}(x) \in \mathbb{R}^{d-2d_{\text{sty}}}$ contains the remaining coordinates. These blocks satisfy

$$\sum_{i=1}^{d_{\text{sty}}} \big\|\tilde{f}^{(i)}_{\text{sty}}(x)\big\|_2^2 + \big\|\tilde{f}_{\text{res}}(x)\big\|_2^2 = 1.$$

1. ***Center synchronization.*** *For each cluster c, H_c maps its rotation center r_c to the north pole $e_d = (0, \ldots, 0, 1)^T$ (i.e., $H_c r_c = e_d$).*

2. ***Style-plane separation.*** *For each $1 \le i \le d_{\text{sty}}$, the two-dimensional block $\tilde{f}^{(i)}_{\text{sty}}(x)$ lies in the i-th canonical rotation plane $\pi_i := \operatorname{span}\{e_{2i-1}, e_{2i}\}$. Equivalently, for any angle $\theta \in \mathbb{R}$, define the block-diagonal rotation*

$$\mathcal{R}_i(\theta) := \operatorname{diag}\big(I_2, \ldots, \underbrace{\begin{bmatrix}\cos\theta & -\sin\theta \\ \sin\theta & \cos\theta\end{bmatrix}}_{i\text{-th block}}, \ldots, I_2,\ I_{d-2d_{\text{sty}}}\big).$$

Then $\mathcal{R}_i(\theta)$ rotates only $\tilde{f}_{\text{sty}}^{(i)}$ and leaves $\tilde{f}_{\text{sty}}^{(j)}$ $(j \neq i)$ and $\tilde{f}_{\text{res}}$ unchanged. For an augmented sample $a(t, x)$, we use the same coordinate map $H_{c(x)}$ as for its original sample x. Under its stated conditions, Lemma 1 guarantees that the augmented sample has the same nearest-center assignment.

Thus, we focus solely on disentangling the augmentation-induced style latents in the aligned representation.

Definition 4 (Linearizing Reparametrization)**.** *For each single-axis augmentation family $\{a(t)\}_{t\in\mathcal{T}_a}$, we use its natural scalar parameter t, such as translation displacement or scaling factor. A* linearizing reparametrization *$\tilde{q}_a : \mathcal{T}_a \to \mathbb{R}$ is a continuous map that converts admissible parameter compositions into addition. If t_3 is obtained by composing t_1 and t_2 in the chosen parameterization, then*

$$\tilde{q}_a(t_3) = \tilde{q}_a(t_1) + \tilde{q}_a(t_2).$$

We assume that such a map exists and is strictly monotonic in t on the considered parameter range. For translations, $\tilde{q}_a(t) = t$; for positive scaling, $\tilde{q}_a(t) = \log t$.

For cyclic components, parameter addition is performed before reduction modulo the period, or within a chosen interval in which the considered compositions do not wrap around.

Definition 5 (Augmentation-Induced Style Displacement)**.** *Let $a_i(t_i) \in \mathcal{A}$ be an augmentation with parameter $t_i \in \mathcal{T}_{a_i}$. For $z = (s, u) \in \mathcal{Z}_{\text{train}}$, define the* induced style displacement *$\Delta_{a_i}(t_i, z)$ by*

$$a_i(t_i, \varphi(s, u)) = \varphi\big(s + \Delta_{a_i}(t_i, z), u\big).$$

Here, u is preserved by Assumption 1, and addition is taken in the chosen style coordinates. For a cyclic component with period p, the endpoint relation determines the displacement only modulo p. For each augmentation call, we fix one real-valued displacement consistent with this relation. The coordinate update adds this displacement to the original coordinate and reduces the result modulo p.

D.3 Proof of Proposition 1

Setup. Stage 1 yields K normalized rotation centers $\{r_\ell\}_{\ell=1}^K \subset \mathbb{S}^{d-1}$. In Stage 2 we fix these centers and the Householder maps $H_\ell r_\ell = e_d$, while recomputing $c(x) = \arg\min_\ell \|f(x) - r_\ell\|_2$ from the current unit feature. The formula for the assignment rule is fixed, not its value on each sample. We optimize $\mathcal{L}_{\text{equiv}}, \mathcal{L}_{\text{radius}}$ and $\mathcal{L}_{\text{theta}}$ from the main text.

Lemma 1 (Cluster-assignment preservation)**.** *Assume Assumption 1 and use the fixed unit rotation centers, orthogonal coordinate maps, and nearest-center assignment defined above. Suppose the radius and equivariance residuals are zero on every original–augmented pair under consideration, with*

$$d \geq 2m + 1, \qquad 0 < m\omega^2 < 1.$$

Define

$$\rho_\omega := \sqrt{2 - 2\sqrt{1 - m\omega^2}},$$

and, for $K \geq 2$, let

$$d_{\min} := \min_{\ell \neq o} \|r_\ell - r_o\|_2.$$

For $K = 1$, set $d_{\min} = +\infty$. If $d_{\min} > 2\rho_\omega$, then every pair $(x, a(t, x))$ under consideration satisfies

$$c(a(t, x)) = c(x).$$

Proof. Fix a pair $(x, a(t, x))$ under consideration, and write $\ell = c(x)$. The zero radius residual gives

$$\|[H_\ell f(x)]_i\|_2 = \|[H_\ell f(a(t, x))]_i\|_2 = \omega, \qquad i = 1, \ldots, m.$$

The zero equivariance residual gives

$$[H_\ell f(x)]_{\text{extra}} = [H_\ell f(a(t, x))]_{\text{extra}} = e_{\text{extra}},$$

where

$$e_{\text{extra}} := \left(0, \ldots, 0, \sqrt{1 - m\omega^2}\right)^\top \in \mathbb{R}^{d-2m}$$

is the fixed residual target in the equivariance loss.

Since H_ℓ is orthogonal and $H_\ell r_\ell = e_d$, for either $y = x$ or $y = a(t, x)$ we have

$$\begin{aligned}
\|f(y) - r_\ell\|_2^2 &= \|H_\ell f(y) - e_d\|_2^2 \\
&= \sum_{i=1}^{m} \|[H_\ell f(y)]_i\|_2^2 + \|[H_\ell f(y)]_{\text{extra}} - [e_d]_{\text{extra}}\|_2^2 \\
&= m\omega^2 + \left(\sqrt{1 - m\omega^2} - 1\right)^2 \\
&= 2 - 2\sqrt{1 - m\omega^2} \\
&= \rho_\omega^2.
\end{aligned}$$

Consequently,

$$\|f(x) - r_\ell\|_2 = \|f(a(t, x)) - r_\ell\|_2 = \rho_\omega.$$

For $K = 1$, the conclusion is immediate. For $K \geq 2$, suppose another center r_o, with $o \neq \ell$, were at least as close to $f(a(t, x))$ as r_ℓ. Then

$$\|f(a(t, x)) - r_o\|_2 \leq \|f(a(t, x)) - r_\ell\|_2 = \rho_\omega.$$

The triangle inequality would give

$$d_{\min} \leq \|r_\ell - r_o\|_2 \leq \|r_\ell - f(a(t, x))\|_2 + \|f(a(t, x)) - r_o\|_2 \leq 2\rho_\omega,$$

contradicting $d_{\min} > 2\rho_\omega$. Thus r_ℓ is the unique nearest center of $f(a(t, x))$, and

$$c(a(t, x)) = c(x) = \ell.$$

The same geometric calculation applies to any candidate feature $v \in \mathbb{R}^d$ satisfying

$$\|[H_\ell v]_i\|_2 = \omega \quad (i = 1, \ldots, m), \qquad [H_\ell v]_{\text{extra}} = e_{\text{extra}}.$$

It gives $\|v - r_\ell\|_2 = \rho_\omega$, and the same center-separation condition makes r_ℓ its unique nearest center. □

Proposition 7 (Perfect equivariance with three losses)**.** *Use the generative setting and notation above, and assume the dimension, radius, and center-separation conditions of Lemma 1. The rotation centers and their coordinate maps are fixed. The angle bounds are given by Equation* (17) *and lie within* $(-\pi, \pi)$ *by default. If a bound is slightly enlarged to include exact half-turns, the allowed angles must still satisfy* $|\theta_{a(t)}^{(i)}| < 2\pi$ *for components included in* $\mathcal{L}_{\text{theta}}$*.*

For each admissible augmentation, let $\Delta_a(t, z)$ *denote its style displacement, so that*

$$a(t, \varphi(s, u_k)) = \varphi(s + \Delta_a(t, z), u_k), \qquad z = (s, u_k).$$

For the construction below, component i *is represented in plane* i*, and* $\Delta_a^{(i)}(t, z) = 0$ *whenever* $i \notin \mathcal{S}_a$*. Cyclic coordinates are interpreted according to their intrinsic periods, with* $\Delta_a(t, z)$ *denoting a chosen real displacement.*

Suppose there exist nonzero constants $c_1, \ldots, c_m$ *such that:*

1. ***Head inputs.*** *Each value* $c_i \Delta_a^{(i)}(t, z)$ *is determined by the inputs available to the corresponding angle head* g*. For a parameter-only head, this condition requires the prescribed angles in this construction to be determined by* t *alone. When the head also receives* $x = \varphi(z)$*, invertibility of* φ *ensures that these angles are determined by* (t, x)*.*

2. ***Allowed angles.*** *Each candidate component* $g_a^{\star,(i)} := c_i \Delta_a^{(i)}(t,z)$ *lies in the prescribed output range. On every augmentation call covered by the angular margin, the candidate components satisfy*

$$\sum_{i \in \mathcal{S}_a} |g_a^{\star,(i)}| \geq \epsilon_\theta.$$

For each cyclic component with period p_i*, also require* $c_i p_i \in 2\pi\mathbb{Z}$*.*

Then, in a function class containing the construction below, the three losses admit a common zero:

$$\mathcal{L}_{\text{equiv}} = \mathcal{L}_{\text{radius}} = \mathcal{L}_{\text{theta}} = 0.$$

With positive loss weights, this is a global minimum of the three-loss objective.

For empirical losses, the following conclusions apply to each positively weighted pair; for population losses, they hold almost surely. Every common zero-loss solution satisfies perfect cluster-dependent equivariance on the evaluated pairs:

$$H_{c(x)} f(a(t,x)) = \mathcal{R}(\theta_{a(t)}) H_{c(x)} f(x), \qquad c(a(t,x)) = c(x).$$

Every augmentation call covered by the angular margin induces a non-identity block rotation and a nontrivial feature change.

Proof. A common zero-loss solution. Using the residual target e_{extra} from Lemma 1, choose fixed phases $\beta_{k,i}$ and set

$$h_{k,i}^\star(s_i) := \omega \begin{pmatrix} \cos(c_i s_i + \beta_{k,i}) \\ \sin(c_i s_i + \beta_{k,i}) \end{pmatrix}, \qquad f^\star(\varphi(s,u_k)) := H_k^\top \begin{pmatrix} h_{k,1}^\star(s_1) \\ \vdots \\ h_{k,m}^\star(s_m) \\ e_{\text{extra}} \end{pmatrix}.$$

The invertibility of φ makes this a function of the observation. For cyclic components, $c_i p_i \in 2\pi\mathbb{Z}$ ensures that the feature is unchanged by a full latent period. Moreover,

$$\|f^\star(\varphi(s,u_k))\|_2^2 = m\omega^2 + \|e_{\text{extra}}\|_2^2 = 1.$$

The geometric calculation in the proof of Lemma 1 applies to this candidate and gives

$$\|f^\star(\varphi(s,u_k)) - r_k\|_2 = \rho_\omega.$$

The center-separation condition makes r_k its unique nearest center. Hence $c(\varphi(s,u_k)) = k$, and content preservation gives the same assignment for its admissible augmentations.

Set

$$g_a^{\star,(i)} = c_i \Delta_a^{(i)}(t,z).$$

These outputs are zero on unassigned planes. The head-input condition makes them functions of the allowed inputs, and the allowed-angle condition ensures that they satisfy the prescribed output constraints.

By the sine–cosine addition formulas,

$$h_{k,i}^\star\big(s_i + \Delta_a^{(i)}(t,z)\big) = R\big(g_a^{\star,(i)}\big) h_{k,i}^\star(s_i).$$

Together with the fixed residual, this yields

$$H_k f^\star(a(t,\varphi(z))) = \mathcal{R}(g_a^\star) H_k f^\star(\varphi(z)).$$

Every style block has radius ω, and both residual blocks equal e_{extra}. Thus the radius loss and both terms of the equivariance loss vanish. The angular-margin condition also gives

$$\max\left(0, \epsilon_\theta - \sum_{i \in \mathcal{S}_a} |g_a^{\star,(i)}|\right) = 0$$

on every augmentation call covered by the angular margin. All three losses are therefore zero, attaining their common global minimum.

Now consider an arbitrary common zero-loss solution. Nonnegativity gives zero residuals on each positively weighted empirical pair, or almost surely in the population setting. Fix such a pair and write $k = c(x)$.

(i) Cluster preservation and nonzero blocks. Lemma 1 gives $c(a(t,x)) = c(x) = k$. Zero radius loss gives

$$\|[H_k f(x)]_i\|_2 = \|[H_k f(a(t,x))]_i\|_2 = \omega > 0$$

for every style plane.

(ii) Perfect equivariance. Zero equivariance loss gives

$$[H_k f(a(t,x))]_{1:m} = [\mathcal{R}(\theta_{a(t)}) H_k f(x)]_{1:m},$$

and sets both residual blocks to e_{extra}. Since $\mathcal{R}$ acts as the identity on the residual block,

$$H_k f(a(t,x)) = \mathcal{R}(\theta_{a(t)}) H_k f(x).$$

(iii) Nontrivial feature changes. For every augmentation call covered by the angular margin, zero angular loss gives

$$\sum_{i \in \mathcal{S}_a} |\theta^{(i)}_{a(t)}| \geq \epsilon_\theta > 0.$$

Thus at least one $i \in \mathcal{S}_a$ has $0 < |\theta^{(i)}_{a(t)}| < 2\pi$, by the angle bound. For this plane, $R(\theta^{(i)}_{a(t)}) \neq I$, and

$$\begin{aligned} \|[H_k f(a(t,x))]_i - [H_k f(x)]_i\|_2^2 &= \|R(\theta^{(i)}_{a(t)})[H_k f(x)]_i - [H_k f(x)]_i\|_2^2 \\ &= 4\omega^2 \sin^2\left(\frac{\theta^{(i)}_{a(t)}}{2}\right) > 0. \end{aligned}$$

Thus at least one feature block changes. The full block rotation is non-identity and, since H_k is orthogonal, $f(a(t,x)) \neq f(x)$. □

Remark 1 (Dynamic cluster assignments)**.** *During Stage 2, cluster assignments are recomputed from the current features, allowing the model to revise the initial assignments. Under the geometric margin and pointwise zero-residual conditions of Lemma 1, an augmented sample has the same current assignment as its original sample. Empirical classification and clustering results are reported in Table 2.*

For the single-axis augmentation families analyzed in Appendices D.4 and D.5, we assume a common system of style coordinates in which

$$\Delta_{a_i}(t_i, z) = \tilde{q}_{a_i}(t_i) e_i,$$

independently of z. The scalar map $\tilde{q}_{a_i}$ reparameterizes augmentation parameters so that admissible compositions are additive (Definition 4). Its strict monotonicity in t_i ensures that a nonzero linear function of $\tilde{q}_{a_i}(t_i)$ is strictly monotonic in t_i.

D.4 Proof of Proposition 2

Definition 6 (Weak Disentanglement)**.** *Let $i \in \{1, \ldots, d_{\text{sty}}\}$ index a style variation component assigned to rotation subspace i. For an augmentation $a(t)$ and a sample $x = \varphi(z)$, let $\Delta^{(i)}_a(t,z)$ denote the induced displacement in this component (Definition 5). The representation f is said to be* weakly disentangled with respect to component i under a *if:*

1. ***Subspace correspondence**: for every considered call (t,x), the feature block $[H_{c(x)} f(x)]_i$ is nonzero and transforms as*

$$[H_{c(x)} f(a(t,x))]_i = R(\theta^{(i)}_{a(t)})\, [H_{c(x)} f(x)]_i.$$

2. ***Monotonic correspondence***: *the rotation* $R(\theta^{(i)}_{a(t)})$ *admits a common real-valued angle function of* $\Delta^{(i)}_a(t,z)$ *that is strictly monotonic on the considered displacement range. For every considered call with* $\Delta^{(i)}_a(t,z)=0$*, the corresponding rotation is the identity.*

If, additionally, $\Delta^{(j)}_a(t,z)=0$ *for every* $j\neq i$ *and every considered call, we further require:*

3. ***Subspace exclusivity***: $[H_{c(x)}f(a(t,x))]_j=[H_{c(x)}f(x)]_j$ *for all* $j\neq i$.

D.4.1 Discrete Augmentations

Lemma 2. *Let* $\mathcal{A}=\{a_i(t_i)\}_{i=1}^{d_{\text{sty}}}$ *be parametric augmentation families, with* a_i *assigned to the* i*-th canonical rotation plane. Assume that, in a common system of style coordinates, their induced displacements satisfy*

$$\Delta_{a_i}(t_i,z)=\tilde{q}_{a_i}(t_i)\,e_i,\ \text{ where } |\tilde{q}_{a_i}(t_i)|\le M_i\tilde{q}_0^{(i)},$$

for every $z\in\mathcal{Z}_{\text{train}}$ *and* $t_i\in\mathcal{T}_{a_i}$*. Here,* e_i *is the* i*-th standard basis vector of* $\mathbb{R}^{d_{\text{sty}}}$*,* $\tilde{q}_0^{(i)}>0$ *the canonical step and* $M_i\in\mathbb{N}_+$ *a coordinate bound. For* $k\in\mathbb{Z}^{d_{\text{sty}}}$ *abbreviate* $k\odot\tilde{q}_0:=\sum_j k^{(j)}\tilde{q}_0^{(j)}e_j$*. Let* $\mathcal{K}=\{k\in\mathbb{Z}^{d_{\text{sty}}}:|k^{(i)}|\le M_i\}$*. For each* $k\in\mathcal{K}$*, assume that a parameter tuple* $t(k)$ *realizing* $k\odot\tilde{q}_0$ *is available whenever that call is used below. Zero coordinates are implemented by omitting their augmentations.*

A. Construction of $\mathcal{D}$.

R1 Seeds. $\mathcal{D}^0:=\{0,e_1,\dots,e_{d_{\text{sty}}}\}$.

R2 Two-of-three closure. *Let* $k,k_1,k+k_1\in\mathcal{K}$*. Use this triple only if there exist* $z_A,z_B\in\mathcal{Z}_{\text{train}}$ *and* $z_C\in\mathcal{Z}$ *such that the augmentation calls using* $t(k)$*,* $t(k_1)$*, and* $t(k+k_1)$*, respectively, have the following endpoints:*

$$z_A\xrightarrow{k\odot\tilde{q}_0}z_B,\qquad z_B\xrightarrow{k_1\odot\tilde{q}_0}z_C,\qquad z_A\xrightarrow{(k+k_1)\odot\tilde{q}_0}z_C.$$

All three calls must be admissible and covered by ***E1****. Whenever two entries of the ordered triple* $(k,k_1,k+k_1)$ *belong to* $\mathcal{D}^n$*, insert the remaining entry. The entries need not be distinct.*

Iterating ***R2*** *yields an ascending chain* $\mathcal{D}^0\subset\mathcal{D}^1\subset\cdots$*; abuse* $\mathcal{D}:=\bigcup_{n\ge0}\mathcal{D}^n$.

B. Training-time equivariance. *For each augmentation* $a_i(t_i)$*, there is an angle network* $g_i:\mathcal{T}_{a_i}\to[-\pi,\pi]$*. Assume the training satisfies:*

E1 Cluster-wise perfect equivariance. *All the data are assigned to the corresponding cluster determined by* $\mathcal{A}$*. For any index set* $S\subseteq\{1,\dots,d_{\text{sty}}\}$*, parameters* $(t_i)_{i\in S}$*, and all* $z\in\mathcal{Z}_{\text{train}}$*, calling* only $\{a_i(t_i)\}_{i\in S}$ *yields*

$$\tilde{f}\big(a_S(t,\varphi(z))\big)=\mathcal{R}\Big(\sum_{i\in S}g_i(t_i)\,e_i\Big)\tilde{f}(\varphi(z)),$$

E2 Non-zero canonical angle. $\exists\, t_+^{(i)}\in\mathcal{T}_{a_i}$ *such that* $\tilde{q}_{a_i}(t_+^{(i)})=\tilde{q}_0^{(i)}$ *and* $g_i(t_+^{(i)})=\theta_0^{(i)}\neq0$.

E3 Nonzero style blocks. $\|\tilde{f}^{(i)}_{\text{sty}}(\varphi(z))\|_2>0$ *for all* $z\in\mathcal{Z}_{\text{train}}$ *and all* $i=1,\cdots,d_{\text{sty}}$.

C. Conclusion. (i) $\mathcal{D}$ *is well-defined and is the unique minimal subset satisfying* ***Seeds + Two-of-three closure****.* (ii) *For any* $k\in\mathcal{D}$*, choose parameters* $t_i(k)$ *for the coordinates with* $k^{(i)}\neq0$*, satisfying*

$$\tilde{q}_{a_i}(t_i(k))=k^{(i)}\tilde{q}_0^{(i)},$$

and omit all other augmentations. Let $t(k)$ collect these parameters and define the call's angle vector by

$$g(t(k)) := \sum_{i:\, k^{(i)} \neq 0} g_i(t_i(k))\, e_i.$$

Write $\theta_0 = (\theta_0^{(1)}, \ldots, \theta_0^{(d_{\text{sty}})})^T$. Then

$$\mathcal{R}\big(g(t(k))\big) = \mathcal{R}(k \odot \theta_0).$$

Consequently, for each plane i there exists a strictly monotonic function $\bar{g}_{a_i}$ on the corresponding parameter range such that, for $k \in \mathcal{D}$ with $k^{(i)} \neq 0$,

$$R\big(g_i(t_i(k))\big) = R\big(\bar{g}_{a_i}(t_i(k))\big).$$

Its construction from the canonical step is given in the proof.

Proof. (i) Apply **R2** until no further vector can be added. Since $\mathcal{K}$ is finite, this process terminates. Every subset of $\mathcal{K}$ containing the seeds and closed under **R2** contains all inserted vectors, by induction. Hence the final set is the unique smallest such subset and is independent of the order of rule applications.

(ii) We prove the linearity by induction over the depth at which k enters $\mathcal{D}$.

Base.

- *Zero vector.* For $k = 0$, no augmentation is called, so $g(t(0)) = 0$ and the rotation is the identity.
- *Unit vector e_i* Activate *only* the i-th augmentation with its canonical parameter $t_+^{(i)}$ and leave every other augmentation uncalled: $a_{\{i\}}(t) = \{a_i(t_+^{(i)})\}$. Because $\tilde{q}_{a_i}(t_+^{(i)}) = \tilde{q}_0^{(i)}$ while all other displacements are zero, the latent displacement equals $e_i \odot \tilde{q}_0$. Perfect equivariance therefore yields

  $$\tilde{f}\big(a_i(t_+^{(i)}, \varphi(z))\big) = \mathcal{R}_i\big(g_i(t_+^{(i)})\big)\, \tilde{f}\big(\varphi(z)\big),$$

  and **E2** gives $g_i(t_+^{(i)}) = \theta_0^{(i)} \neq 0$. Consequently

  $$g(t(e_i)) = (0, \ldots, \theta_0^{(i)}, \ldots, 0) = e_i \odot \theta_0.$$

Hence $g(t(k)) = k \odot \theta_0$ holds for every $k \in \mathcal{D}^0$.

Induction step. Assume $\mathcal{R}(g(t(\ell))) = \mathcal{R}(\ell \odot \theta_0)$ for $\forall \ell \in \mathcal{D}^n$. Let **R2** add the previously missing vector of the triple $(k, k_1, k + k_1)$. By the admissibility requirement in **R2**, the same three training-time equations are available regardless of which displacement is the newly inserted one. In the displays below z_A, z_B are the supervised base points, and z_C is their common endpoint.

$$\begin{aligned}
&\textbf{Path I:} && z_A \xrightarrow{k \odot \tilde{q}_0} z_B && \tilde{f}(\varphi(z_B)) = \mathcal{R}(g(t(k)))\, \tilde{f}(\varphi(z_A)),\\
&\textbf{Path II:} && z_B \xrightarrow{k_1 \odot \tilde{q}_0} z_C && \tilde{f}(\varphi(z_C)) = \mathcal{R}(g(t(k_1)))\, \tilde{f}(\varphi(z_B)),\\
&\textbf{Path III:} && z_A \xrightarrow{(k+k_1) \odot \tilde{q}_0} z_C && \tilde{f}(\varphi(z_C)) = \mathcal{R}(g(t(k+k_1)))\, \tilde{f}(\varphi(z_A)),
\end{aligned}$$

Composing Paths I and II and comparing with Path III gives

$$\mathcal{R}(g(t(k+k_1)))\tilde{f}(\varphi(z_A)) = \mathcal{R}(g(t(k_1)) + g(t(k)))\tilde{f}(\varphi(z_A)).$$

In each 2D plane the feature block is nonzero. Two planar rotations that agree on a nonzero vector are the same rotation. Hence

$$\mathcal{R}(g(t(k+k_1))) = \mathcal{R}(g(t(k_1)))\mathcal{R}(g(t(k))).$$

If the new vector is $k + k_1$, multiply the two known rotations. If it is k or k_1, multiply by the inverse of the other known rotation. In all three cases the induction hypothesis gives

$$\mathcal{R}(g(t(\ell))) = \mathcal{R}(\ell \odot \theta_0)$$

for the new vector ℓ. This completes the induction.

For each component, define one function over its whole parameter range by

$$\bar{g}_{a_i}(t_i) := \frac{\theta_0^{(i)}}{\tilde{q}_0^{(i)}}\, \tilde{q}_{a_i}(t_i).$$

The denominator is positive and the numerator is nonzero. Since $\tilde{q}_{a_i}$ is strictly monotonic in t_i, $\bar{g}_{a_i}$ is strictly monotonic as well. For $k \in \mathcal{D}$ with $k^{(i)} \neq 0$, the rotation identity above gives

$$R(g_i(t_i(k))) = R\big(\bar{g}_{a_i}(t_i(k))\big).$$

This proves the claimed existence. □

Using the lemma, we now relate any two training features in the same cluster by rotations in their separate style subspaces.

Proposition 7a (Discrete weak disentanglement)**.** *Assume **E1**–**E3** of Lemma 2 and the following conditions:*

- ***A1 Rotation correspondence*** *(Lemma 2): for all $k \in \mathcal{D}$, there exists a parameter collection $t(k)$ as defined in Lemma 2 such that*
$$\mathcal{R}(g(t(k))) = \mathcal{R}(k \odot \theta_0).$$
- ***A2 Chain connectivity****: For any $z_A, z_B \in \mathcal{Z}_{\text{train}}$ with $z_A \sim_{\mathcal{A}} z_B$, there exists a finite chain*
$$z_A = z_0,\ z_1, \ldots, z_n = z_B \quad (\text{all in } \mathcal{Z}_{\text{train}})$$
such that for each edge (z_j, z_{j+1}) there exist augmentations $u_j^+(t_j^+)$ and $u_j^-(t_j^-)$ with
$$\varphi^{-1}\big(u_j^+(t_j^+, \varphi(z_j))\big) = \varphi^{-1}\big(u_j^-(t_j^-, \varphi(z_{j+1}))\big) =: z_j^\star \in \mathcal{Z},$$
*and associated style displacements $k_j^+ \odot \tilde{q}_0$ and $k_j^- \odot \tilde{q}_0$ with $k_j^+, k_j^- \in \mathcal{D}$. Each call $u_j^\pm$ applies a subset of the families $\{a_i\}$, uses the parameter tuple $t(k_j^\pm)$ from **A1**, and is covered by **E1**.*

*Then the representation f satisfies **weak disentanglement** (Definition 6) on the discrete grid. For any two training points $x_A = \varphi(z_A)$ and $x_B = \varphi(z_B)$ in the same cluster, choose a chain satisfying **A2** and set*

$$k := \sum_{j=0}^{n-1} (k_j^+ - k_j^-) \in \mathbb{Z}^{d_{\text{sty}}}.$$

Then

$$\tilde{f}(x_B) = \mathcal{R}(k \odot \theta_0)\tilde{f}(x_A).$$

The accumulated style displacement is $k \odot \tilde{q}_0$. Writing $z_A = (s_A, u)$ and $z_B = (s_B, u)$, its i-th component equals $s_B^{(i)} - s_A^{(i)}$ for a non-cyclic coordinate, and is congruent to this difference modulo p_i for a cyclic coordinate with period p_i. Only the call indices $k_j^\pm$ are required to lie in $\mathcal{D}$; the accumulated index k need not. Different admissible chains induce the same block-diagonal rotation.

*Subspace correspondence follows from **E1**. The remaining requirements of Definition 6 hold as follows:*

1. ***Subspace exclusivity****: all feature blocks other than the i-th remain unchanged under $a_i(t)$;*

2. ***Monotonic correspondence****: there exists a strictly monotonic function $\bar{g}_{a_i}$ on the considered parameter range such that*

$$R\big(\theta^{(i)}_{a_i(t)}\big) = R\big(\bar{g}_{a_i}(t)\big).$$

Proof. (Cluster-wise alignment). Fix an edge (z_j, z_{j+1}) of the chain in **A2**. By perfect equivariance and **A1** (provides $\mathcal{R}(g(t(k))) = \mathcal{R}(k \odot \theta_0)$ on $\mathcal{D}$),

$$\tilde{f}(\varphi(z_j^\star)) = \mathcal{R}(k_j^+ \odot \theta_0)\, \tilde{f}(\varphi(z_j)), \qquad \tilde{f}(\varphi(z_j^\star)) = \mathcal{R}(k_j^- \odot \theta_0)\, \tilde{f}(\varphi(z_{j+1})).$$

Equating and left-multiplying by the inverse rotation yields

$$\tilde{f}(\varphi(z_{j+1})) = \mathcal{R}\big((k_j^+ - k_j^-) \odot \theta_0\big)\, \tilde{f}(\varphi(z_j)).$$

Composing along $j = 0, \ldots, n-1$ and using commutativity of block-diagonal rotations,

$$\tilde{f}(\varphi(z_B)) = \mathcal{R}\Big(\sum_{j=0}^{n-1} (k_j^+ - k_j^-) \odot \theta_0\Big)\, \tilde{f}(\varphi(z_A)).$$

With k defined above, this is the claimed feature relation. The endpoint displacement identities follow by summing the style-coordinate changes, using ordinary equality for non-cyclic coordinates and equality modulo the period for cyclic coordinates.

For another admissible chain with accumulated index k', both rotations map each nonzero style block of x_A to the corresponding block of x_B. By **E3**, the planar rotations must coincide, so

$$(k^{(i)} - k'^{(i)})\theta_0^{(i)} \in 2\pi\mathbb{Z} \qquad \text{for every } i.$$

Thus $\mathcal{R}(k \odot \theta_0) = \mathcal{R}(k' \odot \theta_0)$.

(Subspace exclusivity). Since the rotation $\mathcal{R}(k \odot \theta_0)$ is block-diagonal with the i-th block equal to $R(k^{(i)}\theta_0^{(i)})$, an orthogonal augmentation $a_i(t)$ that contributes only to $k^{(i)}$ leaves all blocks $j \neq i$ unchanged, establishing subspace exclusivity.

(Monotonic correspondence). Use the function constructed in Lemma 2:

$$\bar{g}_{a_i}(t) := \frac{\theta_0^{(i)}}{\tilde{q}_0^{(i)}}\, \tilde{q}_{a_i}(t).$$

It is strictly monotonic because $\theta_0^{(i)} \neq 0$ and the reparametrization is strictly monotonic. Condition **A1** gives

$$R\big(\theta^{(i)}_{a_i(t)}\big) = R\big(\bar{g}_{a_i}(t)\big)$$

on the considered grid. This proves monotonic correspondence in Definition 6. □

Remark 2 (Intuition)**.** *On the displacement grid covered by training, each learned rotation admits an angle representation that is linear in $\tilde{q}$. The linearity can be proved by propagating rotation composition relations from single-step moves to the covered displacement grid.*

D.4.2 Continuous Augmentations

Definition 7 (Realized displacement)**.** *A scalar $s \in [-M_i, M_i]$ is a* realized displacement *along coordinate i if there exist $z_0, z_1 \in \mathcal{Z}_{\text{train}}$ and an admissible call with parameter t_i such that*

$$\tilde{q}_{a_i}(t_i) = s, \qquad a_i(t_i, \varphi(z_0)) = \varphi(z_1).$$

Here, s is the real displacement in the chosen linearizing style coordinate; for a cyclic component, the coordinate update is interpreted modulo its period.

Lemma 3. *Let $\mathcal{A} = \{a_i(t_i)\}_{i=1}^{d_{\text{sty}}}$ be parametric augmentation families, with a_i assigned to the i-th canonical rotation plane. Assume that, in a common system of style coordinates, their induced displacements satisfy*

$$\Delta_{a_i}(t_i, z) = \tilde{q}_{a_i}(t_i)\, e_i, \ \text{where } \tilde{q}_{a_i}(t_i) \in [-M_i, M_i],$$

for every $z \in \mathcal{Z}_{\text{train}}$ and $t_i \in \mathcal{T}_{a_i}$. Here, e_i is the i-th standard basis vector of $\mathbb{R}^{d_{\text{sty}}}$ and $[-M_i, M_i] \subset \mathbb{R}$ is coordinate bound.

*Assume conditions **E1** and **E3** of Lemma 2. For each i, assume that $g_i : \mathcal{T}_{a_i} \to [-\pi, \pi]$ is continuous and that $\tilde{q}_{a_i} : \mathcal{T}_{a_i} \to [-M_i, M_i]$ is a continuous bijection with continuous inverse.*

For every realized step s used in this lemma and every

$$d \in I_s := [-M_i, M_i] \cap (s + [-M_i, M_i]),$$

*assume the three-path comparison below is admissible and covered by **E1**, with $z_0, z_1 \in \mathcal{Z}_{\text{train}}$ and a common endpoint $z_2 \in \mathcal{Z}$. The witnessing pair may depend on d; sample-independence of g_i makes the resulting angular identity common to all witnesses. Write*

$$\widehat{g}_i(d) := g_i\big(\tilde{q}_{a_i}^{-1}(d)\big), \qquad d \in [-M_i, M_i],$$

and assume $\widehat{g}_i(d) \neq 0$ for any nonzero displacement d. We have

1. *$\widehat{g}_i(0) = 0$;*
2. *For every non-zero realized displacement s satisfying the coverage condition, the one-step identity holds on admissible pairs:*

$$\widehat{g}_i(d) - \widehat{g}_i(d-s) = \theta_s \quad \text{whenever } d, d-s \in [-M_i, M_i], \qquad \theta_s := \widehat{g}_i(s) \neq 0. \tag{27}$$

Proof. For any $z \in \mathcal{Z}_{\text{train}}$, Condition **E1** gives $\tilde{f}(\varphi(z)) = \mathcal{R}_i(\widehat{g}_i(0))\tilde{f}(\varphi(z))$. By **E3**, the i-th feature block is nonzero, so the identity implies $\widehat{g}_i(0) \in 2\pi\mathbb{Z}$. Since $\widehat{g}_i(0) \in [-\pi, \pi]$, it follows that $\widehat{g}_i(0) = 0$. This proves (1).

For a realized displacement s and any $d \in I_s$, choose the witnessing training pair and common endpoint supplied by the coverage assumption. The three calls give:

$$\begin{aligned}
&\textbf{Path I:} && z_0 \xrightarrow{se_i} z_1 && \tilde{f}(\varphi(z_1)) = \mathcal{R}_i(\widehat{g}_i(s))\, \tilde{f}(\varphi(z_0)),\\
&\textbf{Path II:} && z_1 \xrightarrow{(d-s)e_i} z_2 && \tilde{f}(\varphi(z_2)) = \mathcal{R}_i\big(\widehat{g}_i(d-s)\big)\, \tilde{f}(\varphi(z_1)),\\
&\textbf{Path III:} && z_0 \xrightarrow{de_i} z_2 && \tilde{f}(\varphi(z_2)) = \mathcal{R}_i\big(\widehat{g}_i(d)\big)\, \tilde{f}(\varphi(z_0)).
\end{aligned}$$

Composing Paths I and II and comparing with Path III gives

$$\mathcal{R}_i\big(\widehat{g}_i(d)\big)\tilde{f}(\varphi(z_0)) = \mathcal{R}_i\big(\widehat{g}_i(d-s)\big)\mathcal{R}_i(\widehat{g}_i(s))\tilde{f}(\varphi(z_0))$$

By **E3**, the i-th style block of $\tilde{f}(\varphi(z_0))$ is nonzero. The two planar rotations agree on this block and therefore coincide:

$$\mathcal{R}_i\big(\widehat{g}_i(d)\big) = \mathcal{R}_i\big(\widehat{g}_i(d-s)\big)\mathcal{R}_i(\widehat{g}_i(s)) = \mathcal{R}_i\big(\widehat{g}_i(d-s) + \widehat{g}_i(s)\big).$$

This first gives only $\widehat{g}_i(d) - \widehat{g}_i(d-s) - \widehat{g}_i(s) \in 2\pi\mathbb{Z}$. Define the left-hand side as $F_s(d)$. By the coverage assumption this holds for every $d \in I_s$, and F_s is continuous there. Since I_s is connected and $2\pi\mathbb{Z}$ is discrete, F_s is constant. At $d = s$ it equals $-\widehat{g}_i(0) = 0$. Hence $F_s \equiv 0$ and $\widehat{g}_i(d) - \widehat{g}_i(d-s) = \theta_s$ on I_s, with $\theta_s = \widehat{g}_i(s) \neq 0$. □

Lemma 4. *Under the assumptions of Lemma 3, let $S \subset (0, M_i]$ be a nonempty family of realized steps such that, for every $r, s \in S$, there exist $m, n \in \mathbb{N}_+$ satisfying $L := mr = ns \in (0, M_i]$. Then the quantity*

$$c_S := \frac{\theta_s}{s} \neq 0 \qquad (s \in S)$$

is well-defined (independent of $s \in S$), and

$$\widehat{g}_i(d) = c_S\, d + H_S(d), \qquad H_S(d - s) = H_S(d) \quad (\forall s \in S),$$

on every subinterval where the pairs in Equation (27) are admissible. In particular, H_S is s-periodic for each $s \in S$.

Proof. Apply the three paths in Lemma 3 at $d_j := jr$ $(j = 1, \ldots, m)$, all of which lie in $[-M_i, M_i]$:

$$\widehat{g}_i(d_j) - \widehat{g}_i(d_{j-1}) = \theta_r \quad \Rightarrow \quad \widehat{g}_i(L) - \widehat{g}_i(0) = \sum_{j=1}^{m} \theta_r = m\,\theta_r.$$

Similarly, using step s at $e_\ell := \ell s$ $(\ell = 1, \ldots, n)$,

$$\widehat{g}_i(L) - \widehat{g}_i(0) = n\,\theta_s.$$

Therefore

$$\frac{\theta_r}{r} = \frac{\theta_s}{s} =: c_S.$$

So the slope is independent of the realized step. Define the common remainder $H_S(d) := \widehat{g}_i(d) - c_S d$.

$$H_S(d) - H_S(d - s) = \big(\widehat{g}_i(d) - \widehat{g}_i(d - s)\big) - \frac{\widehat{g}_i(s)}{s}\big(d - (d - s)\big) = \widehat{g}_i(s) - \widehat{g}_i(s) = 0.$$

Therefore

$$\widehat{g}_i(d) = c_S d + H_S(d), \qquad H_S(d - s) = H_S(d) \quad (\forall\, s \in S). \tag{28}$$

□

Lemma 5. *Under the conditions of Lemma 3, let $S \subset (0, M_i]$ be a nonempty family of realized displacements satisfying the condition on common multiples in Lemma 4, so that $c_S = \theta_s/s$ is well defined for all $s \in S$ and $H_S(d) := \widehat{g}_i(d) - c_S d$ is s-periodic on admissible pairs for every $s \in S$. Suppose moreover that $\widehat{g}_i : [-M_i, M_i] \to \mathbb{R}$ be L-Lipschitz:*

$$|\widehat{g}_i(x) - \widehat{g}_i(y)| \leq L\,|x - y| \qquad \forall\, x, y \in [-M_i, M_i].$$

(a) ***Uniform small remainder from the smallest step in a family.*** *Assume S contains the smallest magnitude step*

$$s_* := \arg\min_{s \in S} |s|.$$

Then, using only the periodicity provided by Lemma 4 for the family S,

$$|H_S(d)| \leq\ 2L\,|s_*|.$$

(b) ***Near-loop slope matching across two families.*** *Let $S_1, S_2 \subset (0, M_i]$ be two families satisfying the condition on common multiples in Lemma 4, with slopes c_{S_1}, c_{S_2}. Suppose there exist $r \in S_1$, $s \in S_2$ and integers $m, n \neq 0$ such that, for some basepoint $d \in [-M_i, M_i]$, the chains*

$$d \to d + mr, \qquad d \to d + ns$$

are admissible and the endpoints are δ-close: $|mr - ns| \le \delta$. Then

$$|c_{S_1} - c_{S_2}| \;\le\; \frac{2L\,\delta}{\max\{\,|mr|,\,|ns|\,\}}$$

and, for any $x \in [-M_i, M_i]$,

$$\big|H_{S_1}(x) - H_{S_2}(x)\big| \;\le\; M_i\,|c_{S_1} - c_{S_2}| \;\le\; \frac{2L\,M_i\,\delta}{\max\{\,|mr|,\,|ns|\,\}}.$$

Each remainder is $2L\delta$-almost periodic across the stated multiple of the other family's step:

$$|H_{S_1}(d + ns) - H_{S_1}(d)| \le 2L\delta, \qquad |H_{S_2}(d + mr) - H_{S_2}(d)| \le 2L\delta,$$

Proof. By Lemma 4 the quotient $c_S = \theta_s/s$ is fixed for any $s \in S$, and $H_S(d) = \widehat{g}_i(d) - c_S d$ is s-periodic for each $s \in S$ on admissible pairs. Moreover, since $\widehat{g}_i$ is L-Lipschitz and $\theta_s = \widehat{g}_i(s) - 0 = \widehat{g}_i(s) - \widehat{g}_i(0)$,

$$|c_S| = \Big|\frac{\theta_s}{s}\Big| = \frac{|\widehat{g}_i(s) - \widehat{g}_i(0)|}{|s|} \le L \qquad \text{for any } s \in S. \tag{$*$}$$

(a) Uniform bound from the smallest step. Fix $d \in [-M_i, M_i]$. By s_*-periodicity (iterated along the segment joining d to $d - ks_* \in [-M_i, M_i]$), choose $k \in \mathcal{Z}$ such that

$$d^\circ := d - ks_* \in [-|s_*|,\, |s_*|] \cap [-M_i, M_i] \qquad \text{and} \qquad H_S(d) = H_S(d^\circ).$$

Then, using $H_S(0) = \widehat{g}_i(0) = 0$ and Equation $(*)$,

$$\begin{aligned} |H_S(d)| = |H_S(d^\circ)| = \big|\widehat{g}_i(d^\circ) - c_S d^\circ\big| &\le |\widehat{g}_i(d^\circ) - \widehat{g}_i(0)| + |c_S|\,|d^\circ| \\ &\le L\,|d^\circ| + L\,|d^\circ| \;\le\; 2L\,|s_*|. \end{aligned}$$

This proves the claimed uniform bound.

(b) Near-loop slope matching and cross-almost-periodicity. By admissibility and additivity along the two chains,

$$\widehat{g}_i(d + mr) = \widehat{g}_i(d) + m\,\theta_r = \widehat{g}_i(d) + m\,c_{S_1}\,r, \qquad \widehat{g}_i(d + ns) = \widehat{g}_i(d) + n\,\theta_s = \widehat{g}_i(d) + n\,c_{S_2}\,s.$$

Subtract and use the L-Lipschitz property:

$$\big|m\,c_{S_1}\,r - n\,c_{S_2}\,s\big| = \big|\widehat{g}_i(d + mr) - \widehat{g}_i(d + ns)\big| \le L\,|mr - ns| \le L\,\delta.$$

By the triangle inequality,

$$\begin{aligned} \big|m\,c_{S_1}\,r - n\,c_{S_2}\,s\big| &= \big|mr\,(c_{S_1} - c_{S_2}) + c_{S_2}\,(mr - ns)\big| \\ &\ge |mr||c_{S_1} - c_{S_2}| - |c_{S_2}||(mr - ns)| \end{aligned}$$

Then,

$$\begin{aligned} |c_{S_1} - c_{S_2}| &\le \frac{\big|m\,c_{S_1}\,r - n\,c_{S_2}\,s\big| + |c_{S_2}|\,|mr - ns|}{|mr|} \\ &\le \frac{L\,\delta + |c_{S_2}|\,|mr - ns|}{|mr|} \\ &\le \frac{L\,\delta + L\,\delta}{|mr|} \quad \text{Using Equation } (*) \\ &= \frac{2L\,\delta}{|mr|} \end{aligned}$$

Interchanging the roles of (r, m) and (s, n) gives

$$|c_{S_1} - c_{S_2}| \le \frac{2L\,\delta}{|ns|}.$$

Combining the two bounds yields the displayed estimate for $|c_{S_1} - c_{S_2}|$,

$$|c_{S_1} - c_{S_2}| \;\le\; \frac{2L\,\delta}{\max\{\,|mr|,\,|ns|\,\}}$$

For the remainders,

$$|H_{S_1}(x) - H_{S_2}(x)| = |(c_{S_2} - c_{S_1})\,x| \le |x|\,|c_{S_1} - c_{S_2}| \le M_i\,|c_{S_1} - c_{S_2}|,$$

which gives the second display.

Finally, for the cross-almost-periodicity, note that

$$\begin{aligned} |H_{S_1}(d + ns) - H_{S_1}(d)| &= \big|\widehat{g}_i(d + ns) - \widehat{g}_i(d) - c_{S_1}\,ns\big| \\ &= \big|n\,c_{S_2}\,s - c_{S_1}\,ns\big| = |ns|\,|c_{S_2} - c_{S_1}|. \end{aligned}$$

Since H_{S_1} is r-periodic and $H_{S_1}(d + mr) = H_{S_1}(d)$

$$|H_{S_1}(d + ns) - H_{S_1}(d)| = \big|H_{S_1}(d + ns) - H_{S_1}(d + mr)\big|.$$

Using the previously derived bound with the triangle trick, we have

$$\begin{aligned} \big|H_{S_1}(d + ns) - H_{S_1}(d)\big| &= \big|\widehat{g}_i(d + ns) - \widehat{g}_i(d + mr) - c_{S_1}(ns - mr)\big| \\ &\le |\widehat{g}_i(d + ns) - \widehat{g}_i(d + mr)| + |c_{S_1}|\,|ns - mr| \\ &\le L\,\delta + L\,\delta = 2L\,\delta. \end{aligned}$$

The estimate for H_{S_2} is identical by symmetry. This completes the proof. □

Remark 3 (Intuition)**.** *Lemma 3 simply shows that along a fixed style coordinate, each realized step s must always induce the same angular increment θ_s, independent of where you start. Lemma 4 then says that, within any family S of steps that share common multiples, these increments are all compatible: they define a single common slope c_S, and $\widehat{g}_i$ decomposes into a linear part $c_S d$ plus an S-periodic remainder H_S. Under the stated coverage conditions, Lipschitz regularity limits how much the angle can change between nearby displacements. With a fixed Lipschitz bound, smaller steps tighten the bound on the periodic remainder, while paths that end close together relative to their accumulated displacements constrain their inferred slopes to be close (Lemma 5). The following corollary gives sufficient conditions for the remainder to vanish, making the angle exactly linear in displacement.*

Corollary 1 (Linearity from either an irrational pair or sub-step refinement)**.** *Fix a coordinate i and assume Lemma 3, including its continuity, non-degeneracy and overlap-coverage conditions. Suppose either*

- *(i) there are realized steps $r_i \in (0, M_i]$ and $\beta_i r_i \in (0, M_i]$, with $\beta_i \in (0, 1) \setminus \mathbb{Q}$; or*
- *(ii) there is a realized step $r_0 \in (0, M_i]$ and every r_0/n, $n \in \mathbb{N}_+$, is realized with the same overlap coverage.*

In case (i), set $c_i = \widehat{g}_i(r_i)/r_i$; in case (ii), set $c_i = \widehat{g}_i(r_0)/r_0$. Then $c_i \neq 0$ and

$$H_i(d) := \widehat{g}_i(d) - c_i d \equiv 0 \quad (d \in [-M_i, M_i]), \qquad g_i(t_i) = c_i \tilde{q}_{a_i}(t_i).$$

Proof. We give separate arguments.

Case (i): incommensurate pair r_i and $\beta_i r_i$. Fix $r_i \in (0, M_i]$ realized and $\beta_i \in (0,1) \setminus \mathbb{Q}$ such that $\beta_i r_i$ is realized. Write $\theta_0^{(i)} := \widehat{g}_i(r_i)$. The one-step identity with $s = r_i$ and the definition $c_i = \theta_0^{(i)}/r_i$ give

$$H_i(d) = H_i(d - r_i) \qquad \text{whenever } d, d - r_i \in [-M_i, M_i].$$

We next compute the change in H_i when its argument decreases by $\beta_i r_i$. For any l with $l, l - \beta_i r_i \in [-M_i, M_i]$,

$$\begin{aligned} H_i(l - \beta_i r_i) &= \widehat{g}_i(l - \beta_i r_i) - \frac{\theta_0^{(i)}}{r_i}(l - \beta_i r_i) \\ &= \big(\widehat{g}_i(l) - \widehat{g}_i(\beta_i r_i)\big) - \frac{\theta_0^{(i)}}{r_i} l + \theta_0^{(i)} \beta_i \\ &= \Big(\widehat{g}_i(l) - \frac{\theta_0^{(i)}}{r_i} l\Big) - \Big(\widehat{g}_i(\beta_i r_i) - \theta_0^{(i)} \beta_i\Big) \\ &= H_i(l) - C, \end{aligned} \tag{29}$$

where we set

$$C := \widehat{g}_i(\beta_i r_i) - \theta_0^{(i)} \beta_i \quad \text{(a constant independent of } l\text{)}.$$

Since $r_i \le M_i$, it suffices to analyze H_i on $I := (0, r_i] \subset [-M_i, M_i]$ and extend by r_i-periodicity. For $m \ge 1$, define

$$d_m := r_i\big(1 - \{m\beta_i\}\big) \in (0, r_i), \tag{30}$$

where $\{x\} = x - \lfloor x \rfloor$ is the fractional part. Since β_i is irrational, $d_m \ne \beta_i r_i$: equality would imply $(m+1)\beta_i \in \mathbb{Z}$. The sequence satisfies $d_1 = r_i - \beta_i r_i$ and

$$d_{m+1} = \begin{cases} d_m - \beta_i r_i, & d_m > \beta_i r_i, \\ d_m - \beta_i r_i + r_i, & d_m < \beta_i r_i. \end{cases}$$

Claim 1. For all $m \ge 1$,

$$H_i(d_m) = -m\,C, \qquad d_m \in (0, r_i). \tag{31}$$

Proof of Claim 1. Base case $m = 1$: taking $l = r_i$ in Equation (29) (both r_i and $r_i - \beta_i r_i$ lie in I),

$$H_i(d_1) = H_i(r_i - \beta_i r_i) = H_i(r_i) - C = 0 - C = -C.$$

Induction step $m \to m + 1$: If $d_m > \beta_i r_i$, then $d_{m+1} = d_m - \beta_i r_i \in (0, r_i)$ and Equation (29) with $l = d_m$ gives $H_i(d_{m+1}) = H_i(d_m) - C$. If instead $d_m < \beta_i r_i$, then $d_{m+1} = d_m - \beta_i r_i + r_i \in (0, r_i)$ and Equation (29) yields $H_i(d_m) - H_i(d_m - \beta_i r_i) = C$; by r_i-periodicity, $H_i(d_m - \beta_i r_i) = H_i(d_m - \beta_i r_i + r_i) = H_i(d_{m+1})$, hence again $H_i(d_{m+1}) = H_i(d_m) - C$. This proves Equation (31).

Claim 2. The sequence $\{d_m\}$ is dense in $(0, r_i]$.

Proof of Claim 2. Since β_i is irrational, the fractional parts $\{\{m\beta_i\} : m \ge 1\}$ are dense in $[0, 1)$ (Kronecker's theorem). Equation (30) therefore implies that $\{d_m\}$ is dense in $(0, r_i]$. In particular, there is a subsequence $d_{m_\ell} \to 0^+$.

From Equation (31), $H_i(d_{m_\ell}) = -m_\ell C$. Since $d_{m_\ell} \to 0^+$ and H_i is continuous with $H_i(0) = 0$, we have $H_i(d_{m_\ell}) \to 0$, hence $-m_\ell C \to 0$. As $m_\ell \to \infty$, the only possibility is $C = 0$. With $C = 0$, Equation (29) gives $H_i(l - \beta_i r_i) = H_i(l)$ on admissible pairs, so H_i is both r_i- and $\beta_i r_i$-periodic. From Equation (31) we also have $H_i(d_m) = 0$ for all m, so H_i vanishes on the dense set $\{d_m\} \subset (0, r_i]$; by continuity, $H_i \equiv 0$ on $(0, r_i]$, and then by r_i-periodicity $H_i \equiv 0$ on $[-M_i, M_i]$.

Case (ii): periods accumulating at 0. By hypothesis, $r_N := r_0/N$ is realized for every $N \ge 1$. Each pair of these steps has the admissible common multiple r_0, so Lemma 4 gives the same slope c_i and makes H_i r_N-periodic on all admissible pairs.

Fix $d \in [-M_i, M_i]$ and define

$$k_N := \operatorname{sgn}(d) \left\lfloor \frac{|d|}{r_N} \right\rfloor, \qquad d_N := d - k_N r_N.$$

Then d_N lies between 0 and d, and $|d_N| < r_N$. All displacement arguments in the stepwise comparison from d to d_N remain in $[-M_i, M_i]$. The overlap-coverage assumption therefore allows us to iterate the r_N-periodicity identity, giving $H_i(d) = H_i(d_N)$. Since $d_N \to 0$, continuity and $H_i(0) = 0$ yield

$$H_i(d) = \lim_{N\to\infty} H_i(d_N) = 0.$$

As d was arbitrary, $H_i \equiv 0$ on $[-M_i, M_i]$.

In both cases, $H_i \equiv 0$ on $[-M_i, M_i]$, hence $\widehat{g}_i(d) = c_i d$ for all $d \in [-M_i, M_i]$. Substituting $d = \tilde{q}_{a_i}(t_i)$ yields

$$g_i(t_i) = c_i\, \tilde{q}_{a_i}(t_i) \quad \text{for all } t_i \in \mathcal{T}_{a_i}.$$

Applying the same argument to all $i \in \{1, \dots, d_{\text{sty}}\}$ gives the linearity of g_i in $\tilde{q}_{a_i}$ for every i.

Consequently, there exists a strictly monotonic function $\bar{g}_{a_i}$ on the considered parameter range such that

$$g_i(t_i) = \bar{g}_{a_i}\big(t_i\big). \qquad \square$$

Remark 4 (Intuition)**.** *The previous lemmas show that, along a single style coordinate, the angle can be written as a linear function plus a remainder periodic within each admissible common-multiple family. The irrational-pair argument independently matches the two increments before eliminating the remainder. Corollary 1 uses an extra richness assumption on the realized steps (either an irrational pair of step sizes, or steps that can be refined to arbitrarily small increments) to show that such a continuous, multi-periodic remainder must in fact be constant. Since the remainder is zero at zero displacement, it must be zero everywhere, so the angle becomes exactly linear in the latent displacement.*

Remark 5 (Empirical near-linearity)**.** *Figures 10 and 11 show that predicted rotation angles vary approximately linearly with augmentation parameters over the sampled ranges on 3DIdent and 3DIEBench. With the subspace radii kept near $\omega > 0$, a near-zero equivariance loss indicates close agreement between the predicted and actual feature rotations.*

In additive displacement coordinates, Lemma 5 bounds the periodic remainder and differences between slopes under its stated coverage and Lipschitz assumptions. Corollary 1 gives sufficient conditions for exact linearity.

Proposition 7b (Continuous weak disentanglement)**.** *Assume cluster-wise perfect equivariance and nonzero style blocks on all pairs used below, as in **E1** and **E3** of Lemma 2 and assume*

***B1 Angle functions** (Corollary 1): for each coordinate i, there exist a nonzero constant c_i and a strictly monotonic function $\bar{g}_{a_i}$ on $\mathcal{T}_{a_i}$ such that*

$$g_i(t_i) = \bar{g}_{a_i}\big(t_i\big) = c_i \tilde{q}_{a_i}(t_i), \qquad t_i \in \mathcal{T}_{a_i}.$$

Write $g(t) := (g_1(t_1), \dots, g_{d_{\text{sty}}}(t_{d_{\text{sty}}}))^\top$ and $c_0 := (c_1, \dots, c_{d_{\text{sty}}})^\top$.

***B2 Chain connectivity**: For any $z_A, z_B \in \mathcal{Z}_{\text{train}}$ with $z_A \sim_{\mathcal{A}} z_B$, there exists a finite chain*

$$z_A = z_0, z_1, \dots, z_n = z_B \quad (\text{all in } \mathcal{Z}_{\text{train}}),$$

such that for each edge (z_j, z_{j+1}) there exist augmentations $u_j^+(t_j^+)$, $u_j^-(t_j^-)$ and a latent point $z_j^\star \in \mathcal{Z}$ with

$$\varphi^{-1}\big(u_j^+(t_j^+, \varphi(z_j))\big) = \varphi^{-1}\big(u_j^-(t_j^-, \varphi(z_{j+1}))\big) =: z_j^\star.$$

*Each call $u_j^\pm$ applies a subset of the augmentation families $\{a_i\}_{i=1}^{d_{\text{sty}}}$ and is covered by **E1**. The corresponding linearizing displacements $m_j^+, m_j^- \in \mathbb{R}^{d_{\text{sty}}}$ are defined coordinatewise by $m_j^{\pm,(i)} := \tilde{q}_{a_i}(t_j^{\pm,(i)}) \in [-M_i, M_i]$ for each applied a_i, and $m_j^{\pm,(i)} = 0$ otherwise.*

*Then f satisfies **weak disentanglement** (Definition 6) globally on each cluster. For any two training points $x_A = \varphi(z_A)$ and $x_B = \varphi(z_B)$ in the same cluster, write $z_A = (s_A, u)$ and $z_B = (s_B, u)$ in the chosen linearizing coordinates. For any chain satisfying **B2** that connects these points, define its accumulated style displacement by*

$$m := \sum_{j=0}^{n-1} (m_j^+ - m_j^-) \in \mathbb{R}^{d_{\text{sty}}}.$$

Then

$$\tilde{f}(x_B) = \mathcal{R}(m \odot c_0)\tilde{f}(x_A).$$

The resulting rotation is independent of the chosen chain. For a non-cyclic coordinate i, $m^{(i)} = s_B^{(i)} - s_A^{(i)}$. For a cyclic coordinate with period p_i, this equality holds modulo p_i.

*Subspace correspondence follows from **E1**. The remaining requirements of Definition 6 hold as follows:*

1. ***Subspace exclusivity**: all feature blocks other than the i-th remain unchanged under $a_i(t)$;*
2. ***Monotonic correspondence**: there exists a strictly monotonic function $\bar{g}_{a_i}$ on the considered parameter range such that*

$$R\big(\theta^{(i)}_{a_i(t)}\big) = R\big(\bar{g}_{a_i}(t)\big).$$

Proof. (Cluster alignment.) Since perfect equivariance is achieved, for a fixed edge (z_j, z_{j+1}) we have

$$\tilde{f}\big(\varphi(z_j^\star)\big) = \mathcal{R}\big(g(u_j^+)\big)\,\tilde{f}\big(\varphi(z_j)\big) = \mathcal{R}(m_j^+ \odot c_0)\,\tilde{f}\big(\varphi(z_j)\big),$$

where the second equality uses **B1**. Similarly,

$$\tilde{f}\big(\varphi(z_j^\star)\big) = \mathcal{R}(m_j^- \odot c_0)\,\tilde{f}\big(\varphi(z_{j+1})\big).$$

Equating and left-multiplying by the inverse rotation,

$$\tilde{f}\big(\varphi(z_{j+1})\big) = \mathcal{R}\big((m_j^+ - m_j^-) \odot c_0\big)\,\tilde{f}\big(\varphi(z_j)\big).$$

Composing over $j = 0, \ldots, n-1$ and using commutativity of block-diagonal rotations,

$$\tilde{f}\big(\varphi(z_B)\big) = \mathcal{R}\left(\sum_{j=0}^{n-1}(m_j^+ - m_j^-) \odot c_0\right)\tilde{f}\big(\varphi(z_A)\big).$$

Setting $m := \sum_{j=0}^{n-1}(m_j^+ - m_j^-)$ gives the claimed rotation relation. For two such chains with accumulated displacements m and m', both rotations map each nonzero style block at x_A to the same block at x_B. Hence

$$c_i\big(m^{(i)} - m'^{(i)}\big) \in 2\pi\mathbb{Z} \qquad \text{for every } i,$$

so $\mathcal{R}(m \odot c_0) = \mathcal{R}(m' \odot c_0)$. The coordinate identities follow by summing the shared-endpoint relations, using ordinary addition for non-cyclic coordinates and addition modulo p_i for cyclic ones.

(Subspace exclusivity.) The rotation $\mathcal{R}(m \odot c_0)$ is block-diagonal with the i-th block equal to $R(m^{(i)}c_i)$. An orthogonal augmentation $a_i(t)$ contributes only to coordinate i, hence leaves all blocks $j \neq i$ unchanged.

(Monotonic correspondence.) Define

$$\bar{g}_{a_i}(t) := c_i \tilde{q}_{a_i}(t).$$

By **B1**, this function equals the network output $g_i(t)$ as a real-valued angle. Since $c_i \neq 0$ and $\tilde{q}_{a_i}$ is strictly monotonic in t, the function $\bar{g}_{a_i}$ is strictly monotonic. In particular,

$$R\big(\theta^{(i)}_{a_i(t)}\big) = R\big(\bar{g}_{a_i}(t)\big).$$

Together with subspace exclusivity, this proves weak disentanglement. The continuous assumptions give the same conclusion as the discrete case and additionally identify the network output with $\bar{g}_{a_i}(t)$ itself. □

Remark 6 (Intuition)**.** *With the linearity established in Corollary 1, the chain-connectivity argument becomes a direct generalization of the discrete case (Proposition 7a). Each edge provides a rotation relation, and composing these relations gives the rotation associated with the accumulated style displacement. Different chains may accumulate different numbers of full periods in cyclic coordinates, but perfect equivariance and nonzero style blocks ensure that they give the same final rotation.*

D.5 Proof of Proposition 3

Setup. In the chosen linearizing style coordinates, write $z_A = (s_A, u)$ and $z_B = (s_B, u)$ for two points in the same cluster, and set $v := s_B - s_A$. Condition **S2** below requires each comparison chain used in this proposition to accumulate this displacement. We adopt the orientation convention $c_i > 0$ for every plane (which can be enforced by flipping plane axes if needed).

Definition 8 (Strong Disentanglement)**.** *A representation f is said to be* strongly disentangled *with respect to variation component i under augmentation a if it is weakly disentangled with respect to that component under a (Definition 6) and additionally satisfies:*

4. ***Injectivity**: within each cluster, the values of the corresponding variation component are in one-to-one correspondence with feature directions in plane i over the considered support. For a non-cyclic component, a suitable fixed reference direction yields a single-valued, strictly monotonic angular readout. For a cyclic component, values are identified modulo its intrinsic period.*

Definition 9 (Canonical nonnegative feature angle on a style plane)**.** *Fix the counterclockwise orientation of each style plane. For the i-th plane, write*

$$\tilde{f}^{(i)}_{\mathrm{sty}}(x) = (u_i(x), v_i(x)) \in \mathbb{R}^2 \setminus \{0\}.$$

We measure angles between the directions of these blocks, equivalently between their normalized points on the unit circle. For two connected points $x_A = \varphi(z_A) \sim_{\mathcal{A}} x_B = \varphi(z_B)$, let $v := s_B - s_A$ as above and define the principal relative angle by

$$\begin{aligned}\theta^{\mathrm{pr}}_i(\tilde{f}(x_B); \tilde{f}(x_A)) \;&:=\; \operatorname{atan2}\Big(\tilde{f}^{(i)}_{\mathrm{sty}}(x_A) \times \tilde{f}^{(i)}_{\mathrm{sty}}(x_B), \tilde{f}^{(i)}_{\mathrm{sty}}(x_A) \cdot \tilde{f}^{(i)}_{\mathrm{sty}}(x_B)\Big) \\ &= \operatorname{atan2}\Big(u_i(x_A)\, v_i(x_B) - v_i(x_A)\, u_i(x_B),\; u_i(x_A)\, u_i(x_B) + v_i(x_A)\, v_i(x_B)\Big) \\ &\in (-\pi, \pi].\end{aligned}$$

Define the canonical feature angle

$$\Theta^{\mathrm{ccw}}_i(\tilde{f}(x_B); \tilde{f}(x_A)) \;:=\; \begin{cases} \theta^{\mathrm{pr}}_i(\tilde{f}(x_B); \tilde{f}(x_A)), & \theta^{\mathrm{pr}}_i \geq 0, \\ \theta^{\mathrm{pr}}_i(\tilde{f}(x_B); \tilde{f}(x_A)) + 2\pi, & \theta^{\mathrm{pr}}_i < 0, \end{cases} \quad \in [0, 2\pi),$$

and the directed feature angle (follow the sign of $v^{(i)}$) by

$$\Theta^{\mathrm{dir}}_i(\tilde{f}(x_B); \tilde{f}(x_A); v^{(i)}) := \begin{cases} \Theta^{\mathrm{ccw}}_i(\tilde{f}(x_B); \tilde{f}(x_A)) & \in [0, 2\pi), \quad v^{(i)} \geq 0, \\ \Theta^{\mathrm{ccw}}_i(\tilde{f}(x_B); \tilde{f}(x_A)) - 2\pi & \in [-2\pi, 0), \quad v^{(i)} < 0. \end{cases}$$

The argument $v^{(i)}$ in Θ^{dir}_i selects an orientation using the latent displacement sign. Thus this is an analytical characterization, the feature-only readout is given in the post-processing subsection.

Proposition 8 (Strong disentanglement)**.** *Let $\mathcal{A} = \{a_i(t_i)\}_{i=1}^{d_{\mathrm{sty}}}$ be orthogonal augmentations, with a_i acting on component i and assigned only to rotation plane i. Fix a cluster C_k on the support covered by the corresponding weak-disentanglement result. Assume:*

S1 Monotonic angle correspondence. *Under the conditions of Corollary 1 (continuous case) or Lemma 2 (discrete case), the preceding results construct a strictly monotonic function $\bar{g}_{a_i}$ such that*

$$R\big(g_i(t_i)\big) = R\big(\bar{g}_{a_i}(t_i)\big), \qquad \bar{g}_{a_i}(t_i) = c_i \tilde{q}_{a_i}(t_i), \quad c_i \neq 0.$$

In the discrete case, these relations hold for the parameters $t(k)$ with $k \in \mathcal{D}$, and $c_i = \theta_0^{(i)}/\tilde{q}_0^{(i)}$. Only equality of the rotations is needed below; the discrete network output need not equal $c_i \tilde{q}_{a_i}(t_i)$.

S2 Cluster-wise reachability. *The perfect-equivariance, nonzero-block, and chain-connectivity conditions of Proposition 7b (continuous case) or Proposition 7a (discrete case) hold on C_k. Each comparison chain uses the admissible, supervised calls covered by **S1**, with total additive displacement $v = s_B - s_A$.*

For every connected pair in C_k, the directed angle satisfies

$$R\big(\Theta_i^{\mathrm{dir}}(\tilde{f}(x_B); \tilde{f}(x_A); v^{(i)})\big) = R\big(c_i v^{(i)}\big). \tag{32}$$

For a non-cyclic component i, let its values on this support lie in $[a_{k,i}, b_{k,i}]$, where $a_{k,i} < b_{k,i}$. Suppose the full-data no-wrap condition holds:

$$|c_i|(b_{k,i} - a_{k,i}) < 2\pi. \tag{33}$$

Then

$$\Theta_i^{\mathrm{dir}}(\tilde{f}(x_B); \tilde{f}(x_A); v^{(i)}) = c_i v^{(i)}. \tag{34}$$

Moreover, there is a fixed unit reference $e_{k,i} \in \mathbb{S}^1$ for which the angular readout

$$\widehat{f}^{(i)}(x) := \operatorname{atan2}\Big(e_{k,i} \times \tilde{f}_{\mathrm{sty}}^{(i)}(x),\, e_{k,i} \cdot \tilde{f}_{\mathrm{sty}}^{(i)}(x)\Big) \tag{35}$$

is single-valued, depends only on component i, and is strictly monotonic in that component within C_k. Thus f is strongly disentangled with respect to every non-cyclic component satisfying Equation (33). In the discrete case, this conclusion concerns the specified grid, not unobserved values between grid points.

Proof. (i) Rotation between two points. Fix a plane i and a connected pair $x_A = \varphi(z_A)$, $x_B = \varphi(z_B)$ in C_k. By **S1**, a call with additive displacement d on plane i has the same rotation as $R(c_i d)$. Along each shared-endpoint edge in the chain of **S2**, write the two displacements as d_j^+ and d_j^-. Their common endpoint gives

$$R(c_i d_j^+)\tilde{f}_{\mathrm{sty}}^{(i)}(x_j) = R(c_i d_j^-)\tilde{f}_{\mathrm{sty}}^{(i)}(x_{j+1}),$$

and hence

$$\tilde{f}_{\mathrm{sty}}^{(i)}(x_{j+1}) = R\big(c_i(d_j^+ - d_j^-)\big)\tilde{f}_{\mathrm{sty}}^{(i)}(x_j).$$

By **S2**, the edge displacements sum to $v^{(i)} = s_B^{(i)} - s_A^{(i)}$. Composing the rotations gives

$$\tilde{f}_{\mathrm{sty}}^{(i)}(x_B) = R(\theta)\tilde{f}_{\mathrm{sty}}^{(i)}(x_A), \qquad \theta := c_i v^{(i)}. \tag{36}$$

This argument uses the rotation equalities, not real-valued additivity of the discrete network outputs.

(ii) Directed angle under the no-wrap condition. Define the unit directions

$$a = \frac{\tilde{f}_{\mathrm{sty}}^{(i)}(x_A)}{\|\tilde{f}_{\mathrm{sty}}^{(i)}(x_A)\|_2}, \qquad b = \frac{\tilde{f}_{\mathrm{sty}}^{(i)}(x_B)}{\|\tilde{f}_{\mathrm{sty}}^{(i)}(x_B)\|_2}.$$

Since rotations preserve norms, Equation (36) gives $b = R(\theta)a$. Hence,

$$a \cdot b = \cos\theta, \qquad a \times b = \sin\theta. \tag{37}$$

Thus

$$\theta_i^{\mathrm{pr}} = \operatorname{atan2}(\sin\theta, \cos\theta),$$

which implies

$$R(\theta_i^{\mathrm{pr}}) = R(\theta). \tag{38}$$

The definitions of Θ_i^{ccw} and Θ_i^{dir} only add or subtract a full turn, so

$$R(\Theta_i^{\mathrm{ccw}}) = R(\Theta_i^{\mathrm{dir}}) = R(\theta). \tag{39}$$

This proves Equation (32).

Under Equation (33), $|\theta| = |c_i v^{(i)}| < 2\pi$ for every such pair. By the orientation convention, $c_i > 0$. If $v^{(i)} > 0$, then $0 < \theta < 2\pi$, so $\Theta_i^{\mathrm{ccw}} = \theta$ and $\Theta_i^{\mathrm{dir}} = \theta$. If $v^{(i)} < 0$, then $-2\pi < \theta < 0$, so $\Theta_i^{\mathrm{ccw}} = \theta + 2\pi$ and $\Theta_i^{\mathrm{dir}} = \theta$. If $v^{(i)} = 0$, both angles are zero. This proves Equation (34).

(iii) Distinct component values have distinct directions. Equation (36) first shows that two points with the same $z^{(i)}$ have the same direction on plane i, even if their other style coordinates differ. Conversely, suppose two points have the same direction on plane i. A planar rotation fixes a nonzero vector only at integer multiples of 2π, so Equation (36) gives

$$c_i(z_B^{(i)} - z_A^{(i)}) \in 2\pi\mathbb{Z}.$$

Its absolute value is less than 2π by Equation (33). It must therefore be zero, and $c_i \neq 0$ implies $z_B^{(i)} = z_A^{(i)}$. Thus the plane direction depends only on component i and is one-to-one in its value on the specified support.

(iv) A single-valued monotonic readout. Choose one point $x_* = \varphi(z_*) \in C_k$ and set

$$\mu_{k,i} := \frac{a_{k,i} + b_{k,i}}{2}, \qquad e_{k,i} := R\big(c_i(\mu_{k,i} - z_*^{(i)})\big)\frac{\tilde{f}_{\mathrm{sty}}^{(i)}(x_*)}{\|\tilde{f}_{\mathrm{sty}}^{(i)}(x_*)\|_2}.$$

This is a fixed unit reference for the cluster and plane; it need not be the feature direction of an observed sample. Normalizing Equation (36), applied between x_* and any $x = \varphi(z)$, gives

$$\frac{\tilde{f}_{\mathrm{sty}}^{(i)}(x)}{\|\tilde{f}_{\mathrm{sty}}^{(i)}(x)\|_2} = R\big(c_i(z^{(i)} - \mu_{k,i})\big)e_{k,i}.$$

Furthermore,

$$|c_i(z^{(i)} - \mu_{k,i})| \leq \frac{|c_i|(b_{k,i} - a_{k,i})}{2} < \pi.$$

Therefore the principal angle in Equation (35) equals this ordinary real angle:

$$\widehat{f}^{(i)}(x) = c_i(z^{(i)} - \mu_{k,i}). \tag{40}$$

It is single-valued and strictly increasing in $z^{(i)}$ under our orientation convention, and does not depend on other style coordinates. Without reversing a negative-slope plane, the same construction gives a strictly decreasing readout instead. In either orientation, it is strictly monotonic. Together with weak disentanglement, this proves the claim. □

Remark 7 (Intuition)**.** *The established rotation relation makes each feature direction depend only on its corresponding variation component within a cluster. For a non-cyclic component with an angular span below 2π, distinct values have distinct directions. Measuring angles from the midpoint of the covered arc then yields a single-valued, strictly monotonic readout.*

Remark 8 (Angle budget and the role of η)**.** *In Stage 2, the learned rotation is generally vector-valued:*

$$\theta_{a(t)} = \big(\theta_{a(t)}^{(1)}, \ldots, \theta_{a(t)}^{(m)}\big),$$

where each component corresponds to one $\mathrm{SO}(2)$ *subspace. Hence, the no-wrap condition in Proposition 8 is imposed component-wise rather than on the norm of the whole angle vector. Let $\mathcal{S}_a$ denote the set of rotation subspaces assigned to augmentation a. For every $i \in \mathcal{S}_a$, we use the angle bound in Equation* (17)*. The base range $(-\pi, \pi)$ represents all planar rotations except exact half-turns. If exact half-turns are needed, setting $\eta_a^{(i)}$ slightly above* 1 *places $\pm\pi$ inside the output range.*

For theoretical calibration, fix a variation component i and a cluster. Assume the within-cluster rotation relation in Equation (36) *holds throughout the cluster's support with a common nonzero slope c_i. Let $L_i > 0$ denote the length of the component's full coordinate range, taking $L_i = p_i$ when the support covers a full cyclic period p_i. Consider an augmentation family a for which the theoretical assumptions yield the following real-valued angle relation for $x = \varphi(z)$:*

$$\theta^{(i)}_{a(t)} = c_i \Delta^{(i)}_a(t, z).$$

Define

$$M^{(i)}_a := \sup_{(t,z)} \left| \Delta^{(i)}_a(t, z) \right|, \qquad \eta^{(i)}_{a,\star} := \frac{2M^{(i)}_a}{L_i},$$

where the supremum is over the considered augmentation calls and $M^{(i)}_a > 0$. These quantities describe an ideal calibration and are not assumed to be observed during training. The use of maximum absolute displacement also accommodates one-sided and fixed-step augmentations.

With the chosen angle-bound parameter $\eta^{(i)}_a$, Equation (17) *gives*

$$S_i := |c_i| L_i \le \pi \eta^{(i)}_a \frac{L_i}{M^{(i)}_a} = 2\pi \frac{\eta^{(i)}_a}{\eta^{(i)}_{a,\star}}.$$

The reference value $\eta^{(i)}_{a,\star}$ calibrates the angle bound to one full turn over the component's full variation range. For non-cyclic components, $0 < \eta^{(i)}_a < \eta^{(i)}_{a,\star}$ ensures an angular span below 2π, provided the angular-margin constraint remains feasible. For example, when compatible with the prescribed output range, $\eta^{(i)}_a = \frac{2}{3}\eta^{(i)}_{a,\star}$ gives an upper bound of $4\pi/3$.

For a component covering a full continuous period, periodic closure and a nonzero slope imply $c_i p_i = 2\pi n$ for a nonzero integer n. Choosing $\eta^{(i)}_{a,\star} < \eta^{(i)}_a < 2\eta^{(i)}_{a,\star}$ accommodates a single turn within the open angle bounds and ensures $S_i < 4\pi$, forcing $n = \pm 1$. A possible choice, when compatible with the prescribed output range, is $\eta^{(i)}_a = \frac{3}{2}\eta^{(i)}_{a,\star}$, which gives an upper bound of 3π.

To initialize $\eta^{(i)}_a$, we can use the known augmentation magnitude and a coarse estimate of the corresponding variation range in the data, when these can be compared in a common coordinate. These estimates provide an approximate reference value $\eta^{(i)}_{a,\star}$, which guides the initial choice of the angle bound. The bound can then be adjusted after inspecting the learned feature changes. When a comparable range estimate is unavailable, a provisional per-call angle bound can be selected and refined in the same way.

After training, we can visualize the feature distribution of each cluster in its aligned rotation planes. For non-cyclic components, an estimated occupied arc wider than a target such as $4\pi/3$ suggests tightening the bound and retraining. When wrap-around is unclear, we can track feature directions along sufficiently small augmentation steps to estimate the accumulated angle. This helps detect repeated turns that a static feature plot may hide. For full-period cyclic components, we can track these changes over a known period to check for one complete turn.

D.6 Post-processing

Having established strong disentanglement, we can now read out the latent coordinate on each style plane by a simple geometric post-processing of the feature angles.

Under the conditions below, the feature directions for a non-cyclic component lie within an arc of length less than 2π. When the sampling conditions in Corollary 2 hold, the complement of the selected largest gap identifies this arc. Taking its midpoint as the reference direction yields an affine angular readout centered at the midpoint of the latent support.

Under the full-period conditions below, periodic closure, a nonzero slope, and an angular span below 4π imply exactly one turn per latent period and distinct feature directions for distinct cyclic states. Its phase is linear in the cyclic coordinate, with a fixed phase offset for each cluster, as established in Corollary 3.

Corollary 2 (Principal-angle post-processing when the largest gap is the endpoint wrap gap)**.** *Fix the i-th style plane and an $\mathcal{A}$-connected cluster c_k with non-cyclic latent support $z^{(i)} \in [a_k, b_k]$ (the range may depend on the cluster c_k). For each $x = \varphi(z)$, write*

$$\tilde{f}^{(i)}_{\mathrm{sty}}(x) = (u_i(x), v_i(x)) \in \mathbb{R}^2 \setminus \{0\},$$

and define its direction angle by

$$\theta(x) \;:=\; \operatorname{atan2}\big(v_i(x), u_i(x)\big) \;\in (-\pi, \pi].$$

Assume the linear rotation relation in Equation (36) holds for every pair in the considered support of c_k, with a common slope $c_0^{(i)} \neq 0$. This follows from Proposition 8 or Corollary 4 under their respective assumptions. Suppose

$$|c_0^{(i)}|(b_k - a_k) < 2\pi.$$

Then there exists a phase $\beta^{(i)} \in \mathbb{R}/2\pi\mathbb{Z}$ such that

$$\theta(x) \;\equiv\; c_0^{(i)}\, z^{(i)} + \beta^{(i)} \pmod{2\pi} \tag{41}$$

for every $x = \varphi(z)$ in this support. Assume the sample used to determine the reference direction contains observations at both latent endpoints $a_k < b_k$. Let

$$-\pi < \theta_1 < \cdots < \theta_n \le \pi, \qquad n \ge 2,$$

be its distinct principal angles. Define

$$\delta_\ell := \theta_{\ell+1} - \theta_\ell \quad (\ell = 1, \dots, n-1), \qquad \delta_n := \theta_1 + 2\pi - \theta_n.$$

Choose $\ell^* \in \arg\max_{1 \le \ell \le n} \delta_\ell$. Assume the selected gap δ_{ℓ^*} is the gap outside the feature arc corresponding to $[a_k, b_k]$, bounded by the directions at these two endpoints.

Center of the covered arc (no explicit wrapping). *Let*

$$\theta_L := \theta_{\ell^*}, \qquad \theta_R := \theta_{\ell^*+1} \quad \text{(with } \theta_{n+1} := \theta_1 + 2\pi\text{)}.$$

Let $L := 2\pi - \delta_{\ell^*}$ be the angular length of the covered arc. Starting from θ_R, its midpoint is reached after a counterclockwise angular displacement of $L/2$. Define the reference angle by

$$\alpha := \operatorname{atan2}\big(\sin(\theta_R + L/2), \cos(\theta_R + L/2)\big). \tag{42}$$

Finally, define the 1-D feature

$$\widehat{f}^{(i)}(x) \;:=\; \operatorname{atan2}\big(\sin(\theta(x) - \alpha),\; \cos(\theta(x) - \alpha)\big) \;\in (-\pi, \pi]. \tag{43}$$

Conclusion (affine angular readout). *For every $x = \varphi(z)$ in the considered support of c_k,*

$$\widehat{f}^{(i)}(x) \;=\; c_0^{(i)} \left(z^{(i)} - \frac{a_k + b_k}{2} \right).$$

This is an exact, continuous affine readout centered at the midpoint $(a_k + b_k)/2$ of the latent support. It is strictly increasing when $c_0^{(i)} > 0$ and strictly decreasing when $c_0^{(i)} < 0$.

Proof. For any connected pair $x_0 = \varphi(z_0) \sim_{\mathcal{A}} x = \varphi(z)$, applying the relative-angle formula to their normalized feature blocks gives

$$\theta(x) - \theta(x_0) \;\equiv\; \theta_i^{\mathrm{pr}}\big(\tilde{f}(x); \tilde{f}(x_0)\big) \pmod{2\pi}.$$

This rotation relation gives, for every pair in this support:

$$\theta(x) - \theta(x_0) \;\equiv\; c_0^{(i)} \left(z^{(i)} - z_0^{(i)}\right) \pmod{2\pi}.$$

Set $\beta^{(i)} := \theta(x_0) - c_0^{(i)} z^{(i)}(x_0) \in \mathbb{R}/2\pi\mathbb{Z}$ to obtain Equation (41).

(1) Largest gap ⇒ covered arc and its endpoints. By construction the n gaps $\{\delta_\ell\}$ partition the circle and sum to 2π. Removing the selected largest gap δ_{ℓ^*} leaves a shortest circular arc that covers the sample used to determine the reference direction; its endpoints are precisely the two unit directions with principal angles θ_R and θ_L adjacent across δ_{ℓ^*}. By the endpoint wrap-gap assumption, these endpoints are the samples at the latent endpoints $z^{(i)} = a_k$ and $z^{(i)} = b_k$. The complementary arc therefore corresponds to the entire considered support and has length $L = 2\pi - \delta_{\ell^*} \in (0, 2\pi)$.

(2) Midpoint of the covered arc. Starting at θ_R and moving counterclockwise through the covered arc, its midpoint is reached after an angular displacement of $L/2$. The midpoint angle is therefore $\theta_R + L/2$ modulo 2π, and Equation (42) returns its principal value.

(3) Signed offsets about α are single-branch differences. Use a continuous lift $\phi(x)$ of the covered data arc, with midpoint ϕ_{mid} and length $L < 2\pi$. Then $\alpha \equiv \phi_{\text{mid}} \pmod{2\pi}$ and $|\phi(x) - \phi_{\text{mid}}| \le L/2 < \pi$. Therefore

$$\widehat{f}^{(i)}(x) = \phi(x) - \phi_{\text{mid}}.$$

(4) Affine readout centered at the support midpoint. The modular-affine law and the endpoint wrap-gap assumption allow the same lift to be written as $\phi(x) = c_0^{(i)} z^{(i)} + \beta_0$ with one real constant β_0 on the covered arc. Its endpoint angles have signed difference $c_0^{(i)}(b_k - a_k)$ and absolute difference L; thus the reasoning also covers a negative slope. Their midpoint is

$$\phi_{\text{mid}} = c_0^{(i)} \frac{a_k + b_k}{2} + \beta_0.$$

Subtracting this midpoint gives

$$\widehat{f}^{(i)}(x) = c_0^{(i)} \left(z^{(i)} - \frac{a_k + b_k}{2} \right).$$

Continuity follows since no sample sits on the cut, and Step 3 used signed principal differences. □

Corollary 3 (Single-period cyclic component determines the slope (clusterwise phase))**.** *Fix a style plane j and an $\mathcal{A}$-connected cluster c_k whose j-th variation component ranges over a full continuous period $p > 0$. Assume nonzero feature blocks satisfying the linear rotation relation in Equation* (36)*, with a common slope $c_0^{(j)} \neq 0$, for all within-cluster comparisons using coordinates in $I = [a, a + p)$. Assume the feature direction varies continuously through a full period and returns to its starting direction at the end. Suppose further that*

$$|c_0^{(j)}| p < 4\pi.$$

Then

$$c_0^{(j)} p \in \{-2\pi, 2\pi\}.$$

Consequently, distinct values in I have distinct feature directions. There exist constants $r_{j,k} > 0$ and $A_{j,k} \in \mathrm{SO}(2)$ such that

$$\tilde{f}_{\text{sty}}^{(j)}(\varphi(z)) = r_{j,k} A_{j,k} \begin{pmatrix} \cos\big(c_0^{(j)} z^{(j)}\big) \\ \sin\big(c_0^{(j)} z^{(j)}\big) \end{pmatrix} \tag{44}$$

for every point in this support.

Proof. The rotation relation preserves norms, giving a constant radius $r_{j,k} > 0$. Continuity across the period boundary gives $R(c_0^{(j)} p) b = b$ for a nonzero reference block b. Hence $c_0^{(j)} p = 2\pi n$ for some integer n. The nonzero slope excludes $n = 0$, and the bound $|c_0^{(j)}| p < 4\pi$ excludes $|n| \ge 2$. Thus $n = \pm 1$. For two coordinates $z, z' \in I$, equal feature directions imply $c_0^{(j)}(z - z') \in 2\pi\mathbb{Z}$. Since $|c_0^{(j)}(z - z')| < 2\pi$, this forces $z = z'$. Fixing a reference coordinate $z_0^{(j)}$ and its block phase ϕ_0, set $A_{j,k} = R(\phi_0 - c_0^{(j)} z_0^{(j)})$. The rotation relation then gives the stated form. □

The argument uses continuity and closure over a full period. Injectivity on finitely many sampled cyclic values alone does not rule out multiple turns.

Remark 9 (Global alignment after post-processing)**.** ***Non-cyclic case.*** *Under the conditions of Corollary 2, a common latent coordinate, signed slope $c_0^{(i)}$, and support midpoint $(a_i + b_i)/2$ give*

$$\widehat{f}^{(i)}(x) = c_0^{(i)}\Big(z^{(i)} - \tfrac{a_i+b_i}{2}\Big)$$

across clusters. The scalar readouts therefore agree at matching latent values. Different support midpoints introduce cluster-dependent offsets.

Cyclic case. *We retain two-dimensional blocks to represent the full cycle continuously. Under the conditions of Corollary 3, with a common latent coordinate and signed slope across clusters, phase correction gives*

$$A_{j,k}^{-1}\tilde{f}_{\text{sty}}^{(j)}(\varphi(z)) = r_{j,k}\begin{pmatrix}\cos\big(c_0^{(j)}z^{(j)}\big)\\ \sin\big(c_0^{(j)}z^{(j)}\big)\end{pmatrix}.$$

The corrected feature directions agree at matching cyclic values. The full feature blocks also agree when their radii are equal, as under CRED's zero radius loss, where $r_{j,k} = \omega$.

D.7 Proof of Proposition 4

Proposition 9 (Linearity extension to composite augmentations)**.** *Let $\mathcal{A}_{\text{single}} = \{a_i\}_{i=1}^{d_{\text{sty}}}$ be the augmentation families satisfying the assumptions of Corollary 1 in their respective coordinates. Let $a_{\text{as}} \in \mathcal{A} \setminus \mathcal{A}_{\text{single}}$ be a composite augmentation. For every considered pair $(t_{\text{as}}, z_{\text{as}}) \in \mathcal{T}_{\text{as}} \times \mathcal{Z}_{\text{train}}$, assume that the composite call satisfies perfect equivariance and admits a finite chain satisfying the conditions below. Write $x_{\text{as}} := \varphi(z_{\text{as}})$ and $v := \Delta_{a_{\text{as}}}(t_{\text{as}}, z_{\text{as}})$, with the displacement expressed in the chosen linearizing coordinates. The chain is*

$$z_0 := z_{\text{as}},\ z_1, \dots, z_{n-1} \in \mathcal{Z}_{\text{train}}, \quad z_n := \varphi^{-1}\big(a_{\text{as}}(t_{\text{as}}, x_{\text{as}})\big)$$

*(z_n need not lie in $\mathcal{Z}_{\text{train}}$). Every call $u_j^{\pm}$ below applies a subset of $\mathcal{A}_{\text{single}}$ and is covered by the equivariance condition **E1** of Lemma 2; omitted families have zero displacement.*

1. *For each interior edge (z_j, z_{j+1}) with $j = 0, \dots, n-2$, there exist augmentations $u_j^+(t_j^+)$, $u_j^-(t_j^-)$ and an intermediate point $z_j^\star \in \mathcal{Z}$ with*

   $$\varphi^{-1}\big(u_j^+(t_j^+, \varphi(z_j))\big) = \varphi^{-1}\big(u_j^-(t_j^-, \varphi(z_{j+1}))\big) =: z_j^\star,$$

   and corresponding $\tilde{q}$-displacements $m_j^{\pm} \in \mathbb{R}^{d_{\text{sty}}}$ defined coordinatewise by $m_j^{\pm,(i)} := \tilde{q}_{a_i}(t_j^{\pm,(i)}) \in [-M_i, M_i]$ for each applied a_i, and $m_j^{\pm,(i)} = 0$ otherwise.

2. *For the terminal edge (z_{n-1}, z_n), there exists an augmentation $u_{n-1}^+(t_{n-1}^+)$ satisfying*

   $$\varphi(z_n) = u_{n-1}^+(t_{n-1}^+, \varphi(z_{n-1})).$$

   Equivariance is imposed on this call from $z_{n-1} \in \mathcal{Z}_{\text{train}}$, so no augmentation starting at z_n is needed. Define $m_{n-1}^{+,(i)} := \tilde{q}_{a_i}(t_{n-1}^{+,(i)}) \in [-M_i, M_i]$ for each applied a_i, and $m_{n-1}^{+,(i)} = 0$ otherwise.

Require the accumulated displacement of the chain to equal the displacement v of the composite call:

$$v = \sum_{j=0}^{n-2}\big(m_j^+ - m_j^-\big) + m_{n-1}^+.$$

For cyclic components, this equality is required before reduction modulo the period. Under these conditions, the learned angle vector $g_{\text{as}} : \mathcal{T}_{\text{as}} \times \mathcal{X} \to [-\pi, \pi]^{d_{\text{sty}}}$ satisfies

$$g_{\text{as}}^{(i)}(t_{\text{as}}, x_{\text{as}}) \equiv c_i\, v^{(i)} \pmod{2\pi} \quad \textit{for every } i. \tag{45}$$

If $|c_i v^{(i)}| < \pi$ for every i, then

$$g_{\mathrm{as}}(t_{\mathrm{as}}, x_{\mathrm{as}}) = c_0 \odot v. \tag{46}$$

Equation (45) establishes weak disentanglement (Definition 6, conditions 1–2) for each affected component under a_{as} on the considered pairs.

Proof. Write $x_j := \varphi(z_j)$.

Interior edges. For each $j = 0, \dots, n-2$, perfect equivariance applied along the two routes to $z_j^\star$ gives

$$\mathcal{R}(c_0 \odot m_j^+)\, \tilde{f}(x_j) = \tilde{f}(x_j^\star) = \mathcal{R}(c_0 \odot m_j^-)\, \tilde{f}(x_{j+1}),$$

hence

$$\tilde{f}(x_{j+1}) = \mathcal{R}\big(c_0 \odot (m_j^+ - m_j^-)\big)\, \tilde{f}(x_j).$$

Terminal edge. One-sided equivariance from x_{n-1} to x_n yields

$$\tilde{f}(x_n) = \mathcal{R}(c_0 \odot m_{n-1}^+)\, \tilde{f}(x_{n-1}).$$

Composition. Composing all n steps and using the fact that rotations on distinct planes commute (and rotations within the same plane add their angles),

$$\tilde{f}(x_n) = \mathcal{R}\left(c_0 \odot \Big(\sum_{j=0}^{n-2} (m_j^+ - m_j^-) + m_{n-1}^+ \Big) \right) \tilde{f}(x_0) = \mathcal{R}(c_0 \odot v)\, \tilde{f}(x_0).$$

Perfect equivariance of the considered composite call gives $\tilde{f}(x_n) = \mathcal{R}(g_{\mathrm{as}}(t_{\mathrm{as}}, x_{\mathrm{as}}))\, \tilde{f}(x_0)$. By **E3**, every style block of $\tilde{f}(x_0)$ is nonzero. Comparing the two rotations on each block therefore gives

$$g_{\mathrm{as}}^{(i)}(t_{\mathrm{as}}, x_{\mathrm{as}}) \equiv c_i v^{(i)} \pmod{2\pi} \quad \text{for every } i,$$

yielding Equation (45). If $|c_i v^{(i)}| < \pi$ for every i, the congruence and $g_{\mathrm{as}}^{(i)}(t_{\mathrm{as}}, x_{\mathrm{as}}) \in [-\pi, \pi]$ imply equality as real numbers, giving Equation (46).

Disentanglement consequence. By Equation (45), the rotation in each affected plane admits the common real-valued angle representation $c_i v^{(i)}$. Since $c_i \neq 0$, this representation is strictly monotonic in the corresponding displacement $v^{(i)}$. Zero displacement yields the identity rotation. This establishes conditions 1–2 of Definition 6 on the covered calls. □

D.8 Proof of Proposition 5

Composite augmentation with a pure anchor. We consider a composite augmentation that induces a new variation component together with variations covered by existing single-axis augmentations. At an anchor configuration, it affects only the new component. For example, with x- and y-translation as the existing augmentations, image-centered rotation changes both the position and orientation of an off-center object, but only its orientation when the object is centered in the image. We use a *gated composite head* that allows rotations only in the new plane at the anchor and in all style planes away from it.

Proposition 10 (Disentanglement of composite augmentations with a pure anchor)**.** *Let $\mathcal{A}_{\mathrm{single}} = \{a_i(t_i)\}_{i=1}^{d_{\mathrm{sty}}-1}$ be single-axis augmentation families with $\tilde{q}$-displacements $\Delta_{a_i}(t_i, z) = \tilde{q}_{a_i}(t_i) e_i$ and $|\tilde{q}_{a_i}(t_i)| \leq M_i$. Let $a_{\mathrm{cmp}}(t) \in \mathcal{A} \setminus \mathcal{A}_{\mathrm{single}}$ be a composite augmentation affecting a new variation component indexed by d_{sty}. Assume:*

- ***(A1) Pure at an anchor fiber.*** *There exist $J \subseteq \{1, \dots, d_{\mathrm{sty}} - 1\}$ and $\bar{z}_J$ such that, on the fiber $\mathsf{F}_J(\bar{z}_J) := \{z : \pi_J(z) = \bar{z}_J\}$,*

$$\Delta_{a_{\mathrm{cmp}}}(t, z) = \tilde{q}_{a_{\mathrm{cmp}}}(t)\, e_{d_{\mathrm{sty}}}, \qquad |\tilde{q}_{a_{\mathrm{cmp}}}(t)| \leq M_{d_{\mathrm{sty}}}.$$

(A2) Chain coverage off the anchor. *For each considered off-anchor pair (t, z), the composite endpoint $\varphi^{-1}\big(a_{\mathrm{cmp}}(t, \varphi(z))\big)$ is connected to z by a finite chain with the shared-endpoint and terminal-edge structure of Proposition 9. Each call either applies a subset of $\mathcal{A}_{\mathrm{single}}$ covered by **E1**, or applies a_{cmp} alone from a source in $\mathsf{F}_J(\bar{z}_J)$. All sources lie in $\mathcal{Z}_{\mathrm{train}}$, all calls satisfy the stated perfect-equivariance constraints, and all source feature blocks are nonzero. For an anchor call with parameter τ, assumption **(A1)** gives its displacement as $m = \tilde{q}_{a_{\mathrm{cmp}}}(\tau) e_{d_{\mathrm{sty}}}$. Let $v(t, z) := \Delta_{a_{\mathrm{cmp}}}(t, z)$ be the displacement of the composite call. Require the chain to satisfy*

$$v(t,z) = \sum_{j=0}^{n-2} (m_j^+ - m_j^-) + m_{n-1}^+ \in \mathbb{R}^{d_{\mathrm{sty}}}, \qquad m_j^{\pm,(i)} \in [-M_i, M_i],$$

with equality before reduction modulo the period for cyclic components.

Gated composite head. *We define a selector $S : \mathbb{R}^{|J|} \to \{0,1\}^{d_{\mathrm{sty}}}$ that opens only the new channel at the anchor and all channels off it:*

$$S(\pi_J(z)) = \begin{cases} e_{d_{\mathrm{sty}}}, & \pi_J(z) = \bar{z}_J, \\ \mathbf{1}_{d_{\mathrm{sty}}}, & \pi_J(z) \neq \bar{z}_J, \end{cases}$$

and constrain the composite head $g_{\mathrm{cmp}} : \mathbb{R}^{|J|} \times \mathcal{T}_{a_{\mathrm{cmp}}} \to [-\pi, \pi]^{d_{\mathrm{sty}}}$ to satisfy

$$g_{\mathrm{cmp}}(\pi_J(z), t) \in S(\pi_J(z)) \odot [-\pi, \pi]^{d_{\mathrm{sty}}}.$$

Conclusions. *Under perfect equivariance (Lemma 2), nontriviality of g_{cmp} on nonzero displacement, and the linearity given by Corollary 1 for both single-axis and pure-anchor cases:*

(C1) Disentanglement of the new component at the anchor. *There exists $c_{d_{\mathrm{sty}}} \neq 0$ such that for all $z \in \mathsf{F}_J(\bar{z}_J)$,*

$$g_{\mathrm{cmp}}^{(d_{\mathrm{sty}})}(\bar{z}_J, t) = c_{d_{\mathrm{sty}}} \tilde{q}_{a_{\mathrm{cmp}}}(t), \qquad g_{\mathrm{cmp}}^{(j)}(\bar{z}_J, t) = 0 \quad \forall\, j \neq d_{\mathrm{sty}}.$$

In particular, f satisfies weak disentanglement (Definition 6) with respect to the new component d_{sty} under a_{cmp} on the covered calls with sources in the anchor fiber.

(C2) Axis decomposition off the anchor. *For the off-anchor pairs (t, z) covered by **(A2)**,*

$$g_{\mathrm{cmp}}(\pi_J(z), t) \equiv c_0 \odot v(t, z) \pmod{2\pi \text{ per axis}},$$

*where $c_0 = (c_1, \ldots, c_{d_{\mathrm{sty}}})^\top$ combines the slopes of the original augmentations and the new slope established in **(C1)**. This congruence establishes weak disentanglement for each affected component $i \in \{1, \ldots, d_{\mathrm{sty}}\}$ on the covered calls. If $|c_i v^{(i)}(t, z)| < \pi$ for every i, then*

$$g_{\mathrm{cmp}}(\pi_J(z), t) = c_0 \odot v(t, z).$$

Proof. (C1) At the anchor. On $\mathsf{F}_J(\bar{z}_J)$, assumption (A1) gives $\Delta_{a_{\mathrm{cmp}}}(t, z) = \tilde{q}_{a_{\mathrm{cmp}}}(t)\, e_{d_{\mathrm{sty}}}$. Since the selector closes all channels except d_{sty}, $g_{\mathrm{cmp}}^{(j)}(\bar{z}_J, t) = 0$ for $j \neq d_{\mathrm{sty}}$. Applying the single-axis linearity (Corollary 1) on the d_{sty}-th plane,

$$g_{\mathrm{cmp}}^{(d_{\mathrm{sty}})}(\bar{z}_J, t) = c_{d_{\mathrm{sty}}}\, \tilde{q}_{a_{\mathrm{cmp}}}(t),$$

with $c_{d_{\mathrm{sty}}} \neq 0$ by the nontriviality assumption. Hence the predicted angle is strictly monotonic in the induced displacement $\tilde{q}_{a_{\mathrm{cmp}}}(t)$, verifying weak disentanglement (Definition 6) with respect to component d_{sty} on these covered anchor calls.

(C2) Off the anchor. For the off-anchor pairs (t, z) covered by **(A2)**, the original augmentation calls have the rotations established by Corollary 1 and **E1**, while the anchor calls have those established by **(C1)**.

Thus, each call in **(A2)** induces $\mathcal{R}(c_0 \odot m)$. Repeating the shared-endpoint chain argument from the proof of Proposition 9 and comparing with perfect equivariance of the composite call yields, modulo 2π per axis,

$$g_{\mathrm{cmp}}(\pi_J(z), t) \equiv c_0 \odot v(t, z).$$

For each affected component, this congruence gives the common real-valued angle representation $c_i v^{(i)}(t, z)$. Since $c_i \neq 0$, it is strictly monotonic in the corresponding displacement in the linearizing coordinates. This establishes weak disentanglement on the covered calls, including for the new component. If $|c_i v^{(i)}(t, z)| < \pi$ for every i, the congruence and the head's output range imply

$$g_{\mathrm{cmp}}(\pi_J(z), t) = c_0 \odot v(t, z). \qquad \square$$

Corollary 4 (Strong disentanglement under composite augmentations)**.** *Under the assumptions of Proposition 9 or Proposition 10, fix a cluster C_k on the considered support. Assume its intra-cluster connectivity is witnessed by calls for which the corresponding rotation–displacement relations have been established. Use cluster coordinates H_k, with nonzero feature blocks, and let $c_i \neq 0$ be the common slope for component i.*

For a non-cyclic component taking values in $[a_{k,i}, b_{k,i}]$, the same no-wrap condition

$$|c_i|(b_{k,i} - a_{k,i}) < 2\pi$$

yields strong disentanglement. In particular, a suitable fixed reference direction gives the angular readout

$$\widehat{f}^{(i)}(x) = c_i \left(s^{(i)} - \frac{a_{k,i} + b_{k,i}}{2} \right),$$

which depends only on component i and is strictly monotonic on its considered support.

For a cyclic component with intrinsic period p_i in the chosen style coordinates, the condition

$$c_i p_i = \pm 2\pi$$

yields strong disentanglement with component values identified modulo p_i.

Proof. Write $h_i(x) = [H_k f(x)]_i$. For two consecutive points x_j, x_{j+1} in a shared-endpoint chain, let d_j^+ and d_j^- denote the corresponding displacements in component i. Perfect equivariance and the established rotation–displacement relations give

$$R(c_i d_j^+) h_i(x_j) = R(c_i d_j^-) h_i(x_{j+1}),$$

and therefore

$$h_i(x_{j+1}) = R\big(c_i(d_j^+ - d_j^-)\big)\, h_i(x_j).$$

For a non-cyclic component, composing along the chain yields

$$h_i(x_B) = R\Big(c_i(s_B^{(i)} - s_A^{(i)})\Big)\, h_i(x_A).$$

The no-wrap condition then gives injectivity and the stated angular readout by the same argument as Proposition 8.

For a cyclic component, the accumulated displacement can differ from $s_B^{(i)} - s_A^{(i)}$ by an integer multiple of p_i. Since $c_i p_i = \pm 2\pi$, this difference leaves the feature rotation unchanged. The nonzero feature blocks thus have the same direction if and only if

$$s_B^{(i)} - s_A^{(i)} \in p_i \mathbb{Z}.$$

This establishes injectivity modulo the intrinsic period. $\square$

Remark 10 (Training with pure indicators and a fixed gating code)**.** *The anchored extension uses prior knowledge of the augmentation to specify configurations satisfying (A1). For example, an object centered in the image changes orientation but not position under image-centered rotation. Let $\mathcal{X}_{\text{pure}} \subset \mathcal{X}_{\text{train}}$ denote samples at such configurations. Identifying or constructing this subset requires information about the anchor condition, such as object position in this example, but not the full latent vector z. The exact guarantee assumes (A1); using near-anchor samples instead would be a practical approximation. For image-centered rotation, an object whose center is within distance δ of the image center changes position by at most 2δ. Assign the same fixed gating code to all samples in this subset:*

$$h(x) \equiv \mathbf{0} \in \mathbb{R}^{|J|} \quad \textit{for all } x \in \mathcal{X}_{\text{pure}} \qquad (\textit{the code is not learned}).$$

To use $h(x)$ in the selector, replace its anchor value $\bar{z}_J$ with $\mathbf{0}$ and require $h(x) \neq \mathbf{0}$ outside $\mathcal{X}_{\text{pure}}$. The gate then opens only channel d_{sty} on $\mathcal{X}_{\text{pure}}$ and all channels elsewhere. As a soft alternative to the explicit gate, one can penalize the other output channels on pure samples:

$$\mathcal{L}_{\text{gate}} := \mathbb{E}_{x \in \mathcal{X}_{\text{pure}},\, t} \sum_{i \neq d_{\text{sty}}} \big(g^{(i)}_{\text{cmp}}(h(x), t)\big)^2.$$

This penalty encourages those outputs to vanish. The proposition requires them to be zero on the considered pure calls. Under (A2), perfect equivariance makes each covered off-anchor rotation agree with the rotation determined by the single-axis and pure-anchor constraints along its comparison chain.

Role of the fixed code. *Fixing $h(x) \equiv \mathbf{0}$ on pure samples makes the anchor prediction depend only on t, so the same parameter produces the same predicted angle across these samples. This provides the sample-independent angle function used in the single-axis linearity analysis. Away from the anchor, the augmentation can also change the other feature blocks. If the corresponding rotations depend on the sample, $h(x)$ must retain the information needed to predict them.*