# WHAT MAKES A LATENT SPACE SHARED? DIRECT AND RECOVERABLE COMPATIBILITY IN MOTION VAEs

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

A shared latent space is often assumed to make representations from different schemas directly interchangeable, yet sharing parameters, dimensionality, or a prior does not determine how corresponding instances are expressed in latent coordinates. We show that this assumption conflates recoverable correspondence, direct compatibility, and decoder reuse. In a controlled multi-schema VAE on HumanML3D, even identical inputs yield only $5.9$–$17.7\%$ raw Recall@1 and $0.221$–$0.484$ paired-token cosine, despite correspondence being almost perfectly recoverable after linear calibration. Cycle consistency strengthens recoverability but still leaves direct retrieval and token agreement near zero. In contrast, only $64$ paired Rotations clips raise raw Recall@1 to $88.4$–$90.3\%$ and token cosine to approximately $0.82$, making most post-hoc alignment unnecessary, although crossed decoding remains substantially worse than native reconstruction. We further find that reducing measured marginal discrepancy does not imply schema invariance. These results show that a shared backbone or prior does not ensure directly compatible latent coordinates, while limited paired correspondence can substantially improve direct compatibility in the tested Rotations setting.

## 1 INTRODUCTION

Human motion representation is a foundational problem in motion research, providing the interface between physical movement and computational models. The same movement can be represented through different skeletons, joint sets, coordinate systems, rotation parameterizations, temporal resolutions, and preprocessing pipelines (Mahmood et al., 2019), yet learning algorithms operate on these encodings rather than on the physical movement itself. Existing approaches often address this heterogeneity by imposing a common representation or parameterization (Guo et al., 2022; Loper et al., 2015), or by supplying structural information such as skeletal topology, kinematic structure, or known cross-skeleton correspondences (Aberman et al., 2020; Lee et al., 2023). However, different schemas can still describe the same underlying movement. If their latent representations are intended to be shared, it matters whether correspondence is merely recoverable through a schema-specific transformation or is already expressed in compatible coordinates. We therefore ask what forms of compatibility are induced by a shared learning objective, from recoverable correspondence to direct latent agreement and decoder reuse.

We study this question in a deliberately minimal multi-schema variational autoencoder, where each schema retains its native parameterization while schema-specific interfaces connect it to a shared encoder–decoder backbone regularized by a common Gaussian prior. Rather than assuming that this construction yields a shared representation, we examine which forms of compatibility its reconstruction, prior, and cross-schema objectives support. We construct controlled views of the same HumanML3D motions (Guo et al., 2022), so that correspondence is known exactly while schema variation can be isolated. Full-gallery retrieval directly in the learned representation tests clip-level compatibility under a fixed pooling and cosine rule, while Procrustes, ridge regression, and canonical correlation analysis measure correspondence recoverable after calibration on training pairs with the representation model frozen. For same-rate views, we additionally compare synchronized posterior-mean latent tokens before temporal pooling, directly testing whether paired frames occupy compatible coordinates. Cross-schema decoding then tests the stronger requirement that latent tokens produced by one interface can be reused by another interface's decoder. Because post-hoc recoverability also



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

depends on the comparison family and pair budget, we interpret calibrated retrieval as a diagnostic of accessible correspondence rather than a standalone measure of representation quality.

The results show that these forms of compatibility can differ substantially. Across three training seeds, two interfaces receiving identical inputs within one shared VAE achieve only $5.9$–$17.7\%$ raw Recall@1, while their paired-token cosine similarity is only $0.221$–$0.484$, compared with $0.015$–$0.019$ for mismatched pairs; nevertheless, CCA fitted on $20{,}942$ training pairs reaches $99.64$–$99.86\%$. Cycle consistency strengthens recoverable correspondence further, exceeding $99\%$ CCA retrieval while synchronized token similarity remains near zero, showing that round-trip consistency can induce a highly recoverable relation without selecting the same direct coordinate frame. Explicit paired correspondence has a different effect: with only $64$ paired Rotations clips, raw retrieval reaches $88.4$–$90.3\%$ and paired-token cosine rises to approximately $0.82$, yet crossed decoding remains substantially worse than native reconstruction. Marginal objectives provide another distinct signal: reducing measured latent discrepancy does not by itself establish paired compatibility or schema invariance. Together, these findings distinguish marginal similarity, recoverable correspondence, direct clip- and token-level compatibility, and decoder compatibility, showing that a shared backbone or prior is not sufficient evidence of an interoperable latent representation. Our experiments use controlled deterministic or lossily derived views of the same HumanML3D motions, while independently collected, noisy, or asynchronous motion domains remain outside the scope of this study.

## 2 Related Work

**Heterogeneous motion representations.** AMASS (Mahmood et al., 2019) fixes a common schema by converting heterogeneous motion-capture archives to the SMPL+H body model (Loper et al., 2015; Romero et al., 2017), while HumanML3D (Guo et al., 2022) defines a canonical root-relative feature layout with continuous 6D joint rotations, later adopted by (Tevet et al., 2023; Zhang et al., 2023; Guo et al., 2024). Representation choices also arise within learned motion bottlenecks, where controlled comparisons of discrete tokenizers show that quantization design can materially affect reconstruction quality, semantic preservation, and token efficiency even when the encoder–decoder architecture and training protocol are fixed (de Oliveira et al., 2026). When schemas remain distinct, prior work instead introduces information that relates them through skeletal topology, semantic body-part correspondences, or paired motions (Aberman et al., 2020; Lee et al., 2023; Abdul-Massih et al., 2017; Hu et al., 2023; Yamane et al., 2010); these approaches either impose a common parameterization or provide explicit structure between representations, whereas we ask what forms of compatibility emerge when schema-specific interfaces are trained through a shared latent backbone and prior.

**Consistency objectives.** When paired examples are unavailable, cross-domain structure is often encouraged through cycle consistency, which became a standard signal for unpaired translation (Zhu et al., 2017) and has also been combined with shared latent representations (Liu et al., 2017). In motion, Neural Kinematic Networks (Villegas et al., 2018) combine cycle consistency with adversarial retargeting, while Skeleton-Aware Networks (Aberman et al., 2020) additionally constrain the latent code of a retargeted motion, but a successful round trip does not by itself identify the intended intermediate correspondence. Exact cycle-consistent solutions can remain ambiguous up to distribution-preserving automorphisms (Moriakov et al., 2020), and cycle models may preserve source information without recovering the intended mapping (Chu et al., 2017), motivating our distinction between constraints that make cross-schema relations recoverable and those that actually produce directly compatible representations.

**Representation alignment and identifiability.** Distributional alignment by MMD (Gretton et al., 2012; Long et al., 2015) or adversarial objectives (Ganin & Lempitsky, 2015) does not specify which individual samples correspond, and domain invariance alone need not guarantee transfer (Zhao et al., 2019). Latent identifiability depends on structural assumptions (Locatello et al., 2019; Khemakhem et al., 2020), such as diagonal posteriors, which can favor axis-aligned solutions (Rolinek et al., 2019), or paired observations, which can identify additional shared structure (Von Kügelgen et al., 2021; Daunhawer et al., 2023). Representation comparison is likewise relative to the transformation and downstream operation being allowed, since CCA, CKA, and fitted linear maps impose different invariances (Raghu et al., 2017; Morcos et al., 2018; Kornblith et al., 2019; Roeder et al., 2021), while anchor-based methods explicitly construct relations between otherwise distinct latent spaces

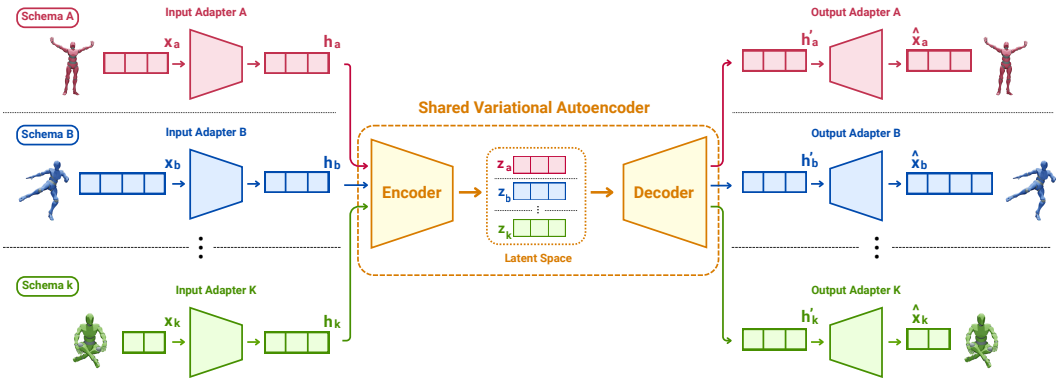



Figure 1: Multi-schema variational autoencoder. Each motion view retains its feature representation and connects to a shared backbone through schema-specific input and output adapters. All schemas share the variational autoencoder and use the same latent dimensionality.

(Moschella et al., 2023; Maiorca et al., 2023). Recent work shows that geometry-preserving and retrieval-oriented alignment can favor different solutions (Achara et al., 2026) and that functional stitching can be directional (Almudévar & Ortega, 2026). We study these distinctions within a jointly trained multi-schema model, separating distributional similarity, recoverable correspondence, coordinate agreement, and decoder compatibility.

## 3 Learning Setting

We investigate whether jointly learning motion representations across schemas produces compatible latent representations using a minimal multi-schema variational autoencoder (VAE) (Kingma & Welling, 2014). Each schema retains its native representation while schema-specific interfaces connect it to a shared latent backbone, allowing us to vary the cross-schema training signal without changing the basic model construction.

### 3.1 Multi-Schema Motion Representation

Let $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_N\}$ denote a collection of motion datasets or controlled views. A sequence $x^{(i)} \in \mathbb{R}^{T_i \times d_i}$ contains $T_i$ frames with $d_i$ features per frame, and inputs retain their native feature layout and are standardized within each schema.

Schema-specific adapters $A_i^{\text{in}}$ and $A_i^{\text{out}}$ connect each representation to a shared encoder $E$ and decoder $D$, as illustrated in Figure 1. The input adapter maps each frame to a common hidden width, while the output adapter returns the decoded features to schema $i$,

$$x^{(i)} \xrightarrow{A_i^{\text{in}}} h^{(i)} \xrightarrow{E,\ \text{sampling}} z^{(i)} \xrightarrow{D} h'^{(i)} \xrightarrow{A_i^{\text{out}}} \hat{x}^{(i)}. \tag{1}$$

We write $F_i = E \circ A_i^{\text{in}}$ and $G_i = A_i^{\text{out}} \circ D$ for the effective encoder and decoder. In the main model, the adapters and shared encoder–decoder blocks are framewise multilayer perceptrons without temporal mixing; complete architectural details are given in Appendix A.2.

For each frame, the shared encoder parameterizes a diagonal-Gaussian posterior with mean $\mu_i(x_t^{(i)})$ and variance $\sigma_i^2(x_t^{(i)})$,

$$z_t^{(i)} = \mu_i(x_t^{(i)}) + \sigma_i(x_t^{(i)}) \odot \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, I_{d_z}). \tag{2}$$

Stacking these vectors gives $z^{(i)} \in \mathbb{R}^{T_i \times d_z}$ with one latent token per frame. Training objectives use sampled latents, whereas deterministic representation comparisons use the posterior-mean tokens $m_t^{(i)} = \mu_i(x_t^{(i)})$.

### 3.2 TRAINING OBJECTIVES

All conditions share a common floor of native reconstruction and KL regularization toward a standard Gaussian prior $p$. They differ only in the additional cross-schema information supplied during training, as summarized in Table 1.

Table 1: Training conditions and their additional cross-schema signals. The discriminator row gives the VAE-side objective, while the domain classifier minimizes $\mathcal{L}_{\mathrm{dom}}$.

| Condition | VAE objective | Additional signal |
|---|---|---|
| Floor | $\mathcal{L}_{\mathrm{floor}}$ | None |
| Cycle | $\mathcal{L}_{\mathrm{floor}} + \lambda_z \mathcal{L}_{z\mathrm{cyc}} + \lambda_x \mathcal{L}_{x\mathrm{cyc}}$ | Round trips |
| MMD | $\mathcal{L}_{\mathrm{floor}} + \lambda_m \mathcal{L}_{\mathrm{MMD}}$ | Latent marginals |
| Discriminator | $\mathcal{L}_{\mathrm{floor}} - \lambda_d \mathcal{L}_{\mathrm{dom}}$ | Schema identity |
| Anchor | $\mathcal{L}_{\mathrm{floor}} + \lambda_a \mathcal{L}_{\mathrm{anchor}}$ | Motion pairs |

**Reconstruction and prior regularization.** The common floor is

$$\mathcal{L}_{\mathrm{floor}} = \mathcal{L}_{\mathrm{rec}} + \beta \mathcal{L}_{\mathrm{KL}}, \tag{3}$$

with

$$\mathcal{L}_{\mathrm{rec}} = \mathbb{E}\left[\frac{\|x^{(i)} - G_i(z^{(i)})\|_F^2}{T_i d_i}\right], \qquad \mathcal{L}_{\mathrm{KL}} = \mathbb{E}\left[\frac{D_{\mathrm{KL}}\Big(q_i(z \mid x^{(i)}) \,\|\, p(z)\Big)}{T_i d_z}\right]. \tag{4}$$

The floor jointly trains the adapters and shared parameters without using pair identities, so neither term specifies how the latent coordinate frames of two interfaces should relate.

**Cross-schema cycle consistency.** For $j \neq i$, a sampled source code is decoded through schema $j$ and re-encoded,

$$\tilde{x}^{(i\to j)} = G_j(z^{(i)}), \qquad \tilde{z}^{(i\to j)} \sim q_j(z \mid \tilde{x}^{(i\to j)}), \tag{5}$$

and Cycle penalizes both latent and motion round-trip error,

$$\mathcal{L}_{z\mathrm{cyc}} = \mathbb{E}\left[\frac{\|\,\mathrm{sg}(z^{(i)}) - \tilde{z}^{(i\to j)}\|_F^2}{T_i d_z}\right], \qquad \mathcal{L}_{x\mathrm{cyc}} = \mathbb{E}\left[\frac{\|x^{(i)} - G_i(\tilde{z}^{(i\to j)})\|_F^2}{T_i d_i}\right]. \tag{6}$$

The intermediate sequence has no paired reconstruction target, so the objective constrains a cross-interface round trip without identifying which target-schema motion should correspond to the source.

**Marginal discrepancy regularization.** MMD, adversarial regularization, and anchors operate on pooled sampled latents $r^{(i)} = [\mathrm{mean}_t\, z_t^{(i)}; \mathrm{std}_t\, z_t^{(i)}] \in \mathbb{R}^{2d_z}$. Let $\nu_i$ denote the distribution of $r^{(i)}$ under schema $i$. Squared maximum mean discrepancy (Gretton et al., 2012) compares these marginals through

$$\mathcal{L}_{\mathrm{MMD}} = \mathbb{E}_{i\neq j}\left[\mathbb{E}k(u,u') + \mathbb{E}k(v,v') - 2\mathbb{E}k(u,v)\right], \tag{7}$$

where $u, u' \sim \nu_i$ and $v, v' \sim \nu_j$. Thus MMD uses only distributional information and no motion pair identities; the kernel construction is specified in Appendix A.3.

**Adversarial schema regularization.** A small MLP domain classifier $H_\phi$ predicts the source schema from pooled latents,

$$\mathcal{L}_{\mathrm{dom}} = -\mathbb{E}_{i,x^{(i)},z^{(i)}}\left[\log H_\phi(i \mid r^{(i)})\right]. \tag{8}$$

The classifier minimizes this loss, while gradient reversal applies $-\lambda_d \mathcal{L}_{\mathrm{dom}}$ to the representation model (Ganin et al., 2016), discouraging schema-predictive information without supplying motion correspondences.

**Paired anchors.** Anchor is the only primary condition supplied with explicit correspondence between motion instances. For paired windows, symmetric InfoNCE (Oord et al., 2018) encourages the representation of each motion to score above motions with different clip identities,

$$\ell_{i\to j} = -\frac{1}{B}\sum_{k=1}^{B}\log\frac{\exp\Big(s_a(r_k^{(i)}, r_k^{(j)})/\tau\Big)}{\sum_{\ell\in\mathcal{C}_k}\exp\Big(s_a(r_k^{(i)}, r_\ell^{(j)})/\tau\Big)}, \qquad \mathcal{L}_{\text{anchor}} = \frac{\ell_{i\to j}+\ell_{j\to i}}{2}, \tag{9}$$

where $s_a$ is cosine similarity and $\tau$ is the temperature. Candidate construction, pair sampling, optimization settings, and the paired-attraction ablation are detailed in Appendix A.3.

# 4 Characterizing Cross-Schema Latent Compatibility

## 4.1 Direct and Recoverable Correspondence

The architecture places latent vectors from every schema in the same ambient space, but this does not imply that corresponding motions are directly compatible. We distinguish correspondence that is already accessible in the learned representations from correspondence recovered only after fitting a transformation on separate motion pairs. At the clip level, *direct retrieval compatibility* means that corresponding motions can be identified under a fixed representation and similarity rule without cross-schema fitting, whereas *recoverable correspondence* refers to comparability obtained only after post-hoc calibration. For synchronized views, we separately use *direct token agreement* to describe numerical similarity between paired posterior-mean latent tokens before temporal pooling.

Marginal similarity and decoder compatibility provide two further, distinct notions of sharing. Similar latent distributions do not determine which individual motions correspond, while direct or recoverable correspondence does not imply that one interface's latent tokens can be consumed by another interface's decoder at native quality. We therefore treat marginal similarity, recoverable correspondence, direct clip- and token-level compatibility, and decoder compatibility as separate operational properties rather than as equivalent consequences of a nominally shared latent space.

## 4.2 Post-Hoc Diagnostics of Correspondence

After training, we freeze the representations and evaluate cross-schema correspondence by full-gallery retrieval, asking whether each motion identifies its known counterpart among all target clips. Raw retrieval uses the fixed clip representation and cosine similarity without fitting a cross-schema map and therefore measures direct clip-level retrieval compatibility. We then fit Procrustes, ridge regression, and canonical correlation analysis (CCA) on separate calibration pairs. Procrustes permits only a centered orthogonal transformation of the pooled clip representations, ridge learns a general linear map from schema B to A, and CCA projects both schemas into a learned common comparison space. Together they form a diagnostic ladder from direct comparison to increasingly flexible forms of recoverable correspondence.

These measurements characterize correspondence only under the comparison being tested rather than equality of the complete latent spaces. In particular, an orthogonal relation between pooled mean–standard deviation descriptors does not imply an orthogonal relation between framewise latent tokens, since componentwise temporal standard deviation is not equivariant to a general token rotation. Calibrated retrieval likewise depends on both the map family and the number of correspondence pairs used to fit it, so high mapped retrieval should be interpreted as correspondence accessible to that calibration protocol rather than as a standalone measure of representation quality. Direct token comparisons and decoder reuse are evaluated separately, with the full protocols specified in Section 5.2.

## 4.3 Coordinate Ambiguity and Training Signals

The floor objective does not uniquely determine relative latent coordinates across schemas. For a signed permutation $P_i$ applied to every latent frame,

$$(\mu_i, \Sigma_i, G_i) \mapsto (P_i\mu_i,\ P_i\Sigma_i P_i^\top,\ G_i P_i^{-1}), \qquad \Sigma_i = \operatorname{diag}(\sigma_i^2), \tag{10}$$

preserves native reconstruction through the compensating decoder transformation. Signed permutations also preserve the diagonal-Gaussian posterior family and unit Gaussian prior, and therefore

Table 2: Controlled views of the same HumanML3D motions. Schema A is native HumanML3D except in Disjoint.

| View pair | Schema A | Schema B | Controlled difference |
|---|---|---|---|
| Identical | native, 263-d | identical, 263-d | none |
| Reordered | native, 263-d | permuted, 263-d | joint/block order |
| Positions | native, 263-d | positions, 71-d | feature selection |
| Rotations | native, 263-d | rotations, 134-d | feature selection |
| Sparse | native, 263-d | sparse, 41-d | joint set and frame subsampling |
| Disjoint | positions, 63-d | rotations, 126-d | disjoint feature blocks |

leave KL regularization unchanged. Thus, at the level of the effective encoder and decoder functions, each schema can admit a different signed permutation without changing the floor objective. Realizing these transformations within the shared architecture requires the closure assumption in Appendix H; our empirical results do not rely on that assumption.

What matters for cross-schema compatibility is the *relative* coordinate frame. A common signed permutation changes the global frame while preserving pooled cosine comparisons, whereas different permutations across interfaces can leave native reconstruction unchanged while separating corresponding latent tokens and destroying direct comparability. The additional objectives constrain this ambiguity in different ways: cycles impose cross-schema round trips without specifying the intended intermediate motion, MMD and adversarial alignment act on latent marginals, and anchors supply explicit motion correspondences. We therefore test whether these signals produce directly compatible clip and token representations or instead make correspondence recoverable only after calibration.

## 5 Experimental Design

We compare the training conditions in Table 1 on paired motion views and evaluate distinct forms of cross-schema compatibility, testing direct clip and token agreement, correspondence recoverable after calibration, latent marginal structure, and cross-schema decoder reuse.

### 5.1 Controlled Schema Comparisons

All comparisons use controlled views derived from the same HumanML3D (Guo et al., 2022) motions, providing exact instance correspondence while varying the representation presented to each interface. Schema A is native 263-dimensional HumanML3D except in *Disjoint*, while schema B varies as shown in Table 2. Training uses $64$-frame source windows, leaving $20{,}942$ of the $23{,}384$ training clips eligible, whereas retrieval evaluates all $4{,}384$ test clips in full. These controlled views isolate representation differences within one corpus and do not model independently collected, noisy, or asynchronous data sources.

Each view pair defines a separate experiment with one shared VAE and two schema-specific adapter pairs. Pair identities are available for evaluation and calibration but enter representation learning only in the paired conditions. The ladder ranges from identical and reordered inputs to feature-selected, temporally subsampled, and disjoint views; in particular, *Disjoint* compares positional and rotational blocks without shared input channels. Exact feature definitions are given in Appendix A.1.

### 5.2 Direct and Recoverable Correspondence

For clip-level evaluation, each sequence is represented by the temporal mean and standard deviation of its posterior-mean tokens, $r^{(i)} = [\mathrm{mean}_t\, m_t^{(i)}; \mathrm{std}_t\, m_t^{(i)}]$. Schema B queries are ranked by cosine similarity against all $4{,}384$ schema A test clips, with Recall@1 as the primary exact-counterpart metric. Raw retrieval uses these representations directly and therefore measures direct clip-level retrieval compatibility; additional ranking and mirror-error analyses are reported in Appendix B.1.

For same-rate paired views, we additionally compare synchronized posterior-mean tokens before temporal pooling using paired-token cosine similarity and a symmetric relative distance. Measurements are averaged first within each clip and then equally across clips. No cross-schema transformation or

normalization is fitted, and a fixed mismatched-identity control provides a reference for agreement without the true pairing. Metric definitions and further token statistics are reported in Appendix B.2.

Recoverable correspondence is measured by fitting centered orthogonal Procrustes, ridge regression, or canonical correlation analysis (CCA) to frozen representations using separate training correspondences, followed by the same full-gallery retrieval. Procrustes permits only an orthogonal transformation of the pooled descriptors, ridge learns a general linear map from schema B to A, and CCA projects both into a common comparison space. CCA uses a fixed maximum of 16 components, reduced only when required by the fitting sample or representation dimension, and its numerical settings are not selected using test retrieval.

All primary maps and preprocessing statistics are fitted exclusively on training data and applied to the untouched test split. Unless a smaller budget is specified, calibration uses all $20{,}942$ eligible training pairs. Because flexible calibration can also expose identity information in random representations, trained Floor and untrained random encoders are compared under matched $256$-pair and full-pair fitting budgets. Estimator details are given in Appendix B.3, with complete budget and random-feature controls in Appendix C.3.

### 5.3 Training Comparisons and Diagnostics

Objective comparisons use *Rotations* unless stated otherwise. Unpaired conditions use unsynchronized windows without pair identities or semantic labels. The main framewise MLP VAE uses $d_z = 64$ and $\beta = 0.3$, with primary results over three different seeds. The limited-pair Anchor study uses nested budgets of $16$, $64$, $256$, and $1{,}024$ unique training clips and evaluates the final $200{,}000$-step checkpoint to avoid introducing additional paired validation information. At each budget, post-hoc maps fitted to frozen Floor models use the same identities and aligned windows, matching the pair budget and representation granularity but not repeated exposure or training compute. Disjoint is evaluated with Floor, Cycle, and full-pair Anchor, but not with the limited-pair dose study.

Marginal structure is evaluated separately from paired correspondence using a held-out logistic schema probe and a fixed multi-bandwidth RBF MMD kernel whose reference bandwidth is estimated from training representations and held constant across models. Cross-model MMD comparisons use this fixed kernel, while the adaptive training statistic is defined in Appendix A.3. Together, the probe and MMD test complementary aspects of marginal structure, and a low kernel discrepancy is not treated as evidence of schema indistinguishability.

Decoder compatibility is also evaluated separately from clip retrieval and token agreement. For selected Identical and Rotations checkpoints, posterior means are decoded either through the native schema A path or from schema B through the schema A decoder on the same $512$ test clips. Local MPJPE is measured over the $21$ non-root RIC joints, excluding root trajectory and orientation. This is a directed $B \to A$ decoder-reuse test, while training-fitted token maps diagnose whether poor direct reuse is linearly recoverable.

## 6 Results

We report full-gallery $B \to A$ Recall@1 in percent, with random-ranking chance $0.023\%$. Primary two-interface calibrations use training correspondences only and, unless stated otherwise, are fitted to all $20{,}942$ eligible training clips. Ranges give minima and maxima over training seeds $42$, $1$, and $2$, rather than confidence intervals.

### 6.1 Shared Training Does Not Ensure Direct Clip or Token Compatibility

The separation between direct and recoverable correspondence appears even when both interfaces receive identical observations. Under Floor, *Identical* reaches only $5.91\%$–$17.72\%$ raw Recall@1, while CCA reaches $99.64\%$–$99.86\%$ (Table 3). This gap is not introduced only by temporal pooling. Synchronized posterior-mean tokens from the same frames have cosine similarity of $0.221$–$0.484$, compared with $0.015$–$0.019$ under mismatched identities, and relative distance $0.989$–$1.227$ versus $1.388$–$1.392$ for the mismatch control. Thus the two interfaces preserve strongly recoverable instance information while assigning corresponding observations to substantially different latent coordinates.

Table 3: Floor retrieval across three training seeds per view pair. Each experiment has two interfaces sharing one backbone. Maps are fitted separately to all 20,942 eligible training pairs. Entries are Recall@1 percentages; Disjoint uses positional schema A.

| View pair | Raw | + Procrustes | + Ridge | + CCA-16 |
|---|---|---|---|---|
| Identical | 5.91–17.72 | 96.30–98.36 | 97.88–99.11 | 99.64–99.86 |
| Reordered | 0.30–36.82 | 94.37–98.61 | 97.54–99.04 | 99.77–99.91 |
| Positions | 0.11–0.94 | 58.65–61.95 | 60.74–63.59 | 95.78–96.10 |
| Rotations | 0.16–1.64 | 78.85–82.66 | 82.89–87.66 | 95.78–96.30 |
| Sparse | 0.21–0.36 | 44.07–47.35 | 45.28–47.38 | 90.90–91.95 |
| Disjoint | 0.14–0.30 | 19.32–21.26 | 28.60–31.32 | 55.66–57.62 |

Table 4: Unpaired objectives on Rotations across three training seeds. Entries are Recall@1 percentages, with each calibration fitted separately to all eligible training pairs.

| Condition | Raw | + Procrustes | + Ridge | + CCA-16 |
|---|---|---|---|---|
| Floor | 0.16–1.64 | 78.85–82.66 | 82.89–87.66 | 95.78–96.30 |
| Cycle | 0.09–0.16 | 94.25–94.89 | 95.80–96.37 | 99.16–99.25 |
| MMD | 0.23–2.85 | 80.18–83.07 | 84.24–87.18 | 97.08–97.67 |
| Discriminator | 12.55–44.14 | 82.66–85.26 | 85.77–88.59 | 96.58–97.42 |

The same pattern holds when the schemas differ. Rotations Floor gives only $0.16\%$–$1.64\%$ raw retrieval and paired-token cosine $0.114$–$0.233$, despite $95.78\%$–$96.30\%$ retrieval after CCA. Reordered also exhibits substantial seed variation in direct compatibility; the seed reaching $36.82\%$ raw retrieval is accompanied by substantially higher token agreement, around $0.71$. Across the evaluated checkpoints, raw Recall@1 and paired-token cosine track one another closely ($\rho_{\mathrm{Spearman}} \approx 0.98$), providing descriptive evidence that poor raw retrieval reflects direct token mismatch rather than being created solely by the pooling rule.

Mapped retrieval nevertheless depends strongly on the calibration protocol. With $256$ Rotations training pairs, CCA recovers approximately $67\%$ Recall@1 from trained Floor representations but only about $7\%$ from untrained random encoders. With all $20{,}942$ pairs, the scores rise to approximately $96\%$ and $74\%$, respectively, so the gap narrows as the amount of paired calibration increases. Procrustes remains more selective at the full budget, reaching approximately $82\%$ for Floor versus $37\%$ for random representations. Large-budget CCA therefore demonstrates correspondence accessible to a flexible fitted comparison, but its absolute score is not a standalone measure of learned representation quality. Full calibration-budget, random-feature, rank, and normalization controls are reported in Appendix C.3.

### 6.2 Unpaired Constraints Improve Different Properties

Cycle provides the clearest separation between recoverable and direct compatibility. On Rotations it raises Procrustes retrieval to $94.25\%$–$94.89\%$ and CCA to $99.16\%$–$99.25\%$, while raw retrieval remains only $0.09\%$–$0.16\%$ (Table 4). The same discrepancy appears before pooling: paired-token cosine ranges from $-0.025$ to $0.026$, close to the mismatched reference. Cycle therefore regularizes a highly recoverable relation between the interfaces without selecting the same direct coordinate frame.

Marginal objectives affect a different property. Using one fixed evaluation kernel across models, MMD with $\lambda_m = 1$ reduces $\mathrm{MMD}^2$ from $0.0130$–$0.0158$ under Floor to $0.0061$–$0.0098$, while raw retrieval remains $0.23\%$–$2.85\%$. The held-out schema probe nevertheless retains $99.95\%$–$99.99\%$ balanced accuracy, with AUC approximately one. The experiment therefore establishes that reducing a measured marginal discrepancy does not imply either schema invariance or paired compatibility; it does not claim that the learned marginals have become indistinguishable.

The discriminator provides the complementary case. Raw retrieval rises to $12.55\%$–$44.14\%$, yet the independent schema probe remains at $99.94\%$–$100\%$ balanced accuracy. Direct counterpart retrieval can therefore improve substantially while schema identity remains readily recoverable. Cycle shows the same qualitative direct-versus-recoverable separation on Disjoint, where no input channels are

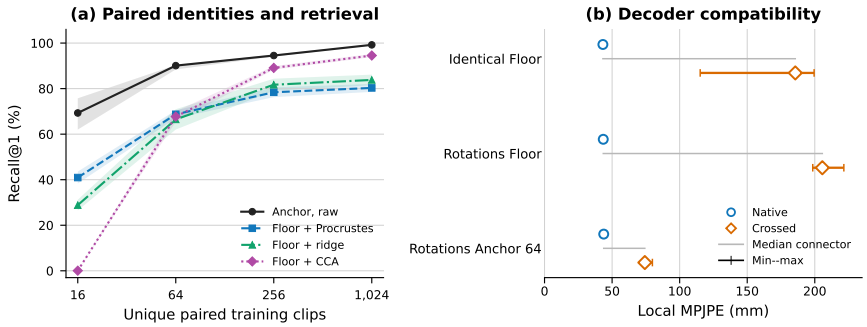



Figure 2: Paired supervision and decoder reuse. (a) Rotations Recall@1 versus paired clip identities. (b) Native and crossed local MPJPE on the same 512 clips at matched checkpoints. Lines and markers summarize three runs as described in the text.

shared, with detailed marginal diagnostics in Appendix D.2 and objective and architecture robustness analyses in Appendix F.

### 6.3 Paired Correspondence Improves Direct Compatibility but Not Native-Quality Decoder Reuse

Explicit motion correspondences have a markedly different effect. Figure 2(a) shows the Rotations Anchor dose response, with raw retrieval reaching $62.00\%$–$75.75\%$ from 16 paired training clips, $88.41\%$–$90.33\%$ from 64, and $94.07\%$–$94.98\%$ from 256. Post-hoc maps use the same identity budgets and aligned-window granularity, although repeated pair exposure and optimization compute are not matched, so this is not an equal-compute efficiency comparison.

The improvement also appears before temporal pooling. Across the three Anchor-64 runs, paired-token cosine reaches $0.824$–$0.826$, compared with $0.114$–$0.233$ for Rotations Floor, showing that explicit correspondence improves the synchronized latent coordinates rather than only the pooled retrieval descriptor. A paired-MSE ablation yields the same qualitative direct-alignment effect (Appendix E).

The limited-pair result is specific to Rotations. With full supervision, Disjoint Anchor reaches $99.59\%$–$99.73\%$ raw retrieval without shared input channels, while full-pair Rotations combines perfect retrieval with very small measured MMD and an almost perfectly accurate schema probe, further separating direct compatibility from schema invariance (Appendix C.2; Appendix D.2).

**Decoder compatibility.** Strong direct compatibility does not ensure native-quality decoder reuse in the tested $B \to A$ setting. Despite the strong Anchor-64 alignment, crossed MPJPE remains $73.87$–$79.83$ mm versus $43.38$–$44.45$ mm natively (Figure 2(b)). A fitted linear token map substantially improves Floor and Cycle but provides little additional gain for Anchor-64, indicating that its remaining decoder gap is not captured by a simple global linear correction. Complete results are reported in Appendix G.1.

## 7 Conclusion

We studied what it means for motion schemas to share a latent representation when schema-specific interfaces connect them to the same VAE backbone and prior. Our results separate several properties that are often conflated: correspondence can be highly recoverable after calibration despite weak direct clip and token agreement, while explicit motion pairs substantially improve direct compatibility without ensuring native-quality cross-decoder reuse. Marginal discrepancy and schema invariance likewise need not track either form of correspondence. These conclusions are limited to controlled HumanML3D views, with low-pair alignment established for Rotations and calibrated recoverability dependent on the comparison family and correspondence budget. More broadly, whether a latent space is "shared" should be tied to the operation it is expected to support rather than inferred from shared architecture, prior, or marginal statistics alone.

### AI USE STATEMENT

Generative AI tools were used for language editing and proofreading, including rephrasing text for clarity, and as coding assistants for implementation tasks. All AI-assisted text and code were reviewed by the authors, who take full responsibility for the final content of the paper.

## REFERENCES

Michel Abdul-Massih, Innfarn Yoo, and Bedrich Benes. Motion style retargeting to characters with different morphologies. In *Computer Graphics Forum*, volume 36, pp. 86–99. Wiley Online Library, 2017.

Kfir Aberman, Peizhuo Li, Dani Lischinski, Olga Sorkine-Hornung, Daniel Cohen-Or, and Baoquan Chen. Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics (ToG)*, 39(4):62–1, 2020.

Akshit Achara, Tatiana Gaintseva, Mateo Mahaut, Pritish Chakraborty, Viktor Stenby Johansson, Melih Barsbey, Emanuele Rodolà, and Donato Crisostomi. Multi-way representation alignment. *arXiv preprint arXiv:2602.06205*, 2026.

Antonio Almudévar and Alfonso Ortega. Bridging functional and representational similarity via usable information. *arXiv preprint arXiv:2601.21568*, 2026.

Casey Chu, Andrey Zhmoginov, and Mark Sandler. CycleGAN, a master of steganography. *arXiv preprint arXiv:1712.02950*, 2017.

Imant Daunhawer, Alice Bizeul, Emanuele Palumbo, Alexander Marx, and Julia E. Vogt. Identifiability results for multimodal contrastive learning. In *International Conference on Learning Representations*, 2023.

Kaiki Camilo de Oliveira, Iago Alves Brito, Lucas Arruda Fernandes, Walcy Rios, and Arlindo Rodrigues Galvão Filho. Benchmarking motion tokenizers: A controlled study of human motion quantization. In *ECCV26 Workshop: Human Motion-Informed World Models and Socially Intelligent Action*, 2026. URL https://openreview.net/forum?id=J6GFuk0FQR.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59):1–35, 2016.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The journal of machine learning research*, 13:723–773, 2012.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3D human motions from text. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5142–5151. IEEE, 2022.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. MoMask: Generative masked modeling of 3D human motions. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1900–1910. IEEE, 2024.

Lei Hu, Zihao Zhang, Chongyang Zhong, Boyuan Jiang, and Shihong Xia. Pose-aware attention network for flexible motion retargeting by body part. *IEEE Transactions on Visualization and Computer Graphics*, 30(8):4792–4808, 2023.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International conference on artificial intelligence and statistics*, pp. 2207–2217. PMLR, 2020.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019.

Sunmin Lee, Taeho Kang, Jungnam Park, Jehee Lee, and Jungdam Won. SAME: Skeleton-agnostic motion embedding for character animation. In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11, 2023.

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, 30, 2017.

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.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pp. 97–105. PMLR, 2015.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6), November 2015. ISSN 0730-0301. doi: 10.1145/2816795.2818013. URL https://doi.org/10.1145/2816795.2818013.

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5442–5451, 2019.

Valentino Maiorca, Luca Moschella, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. Latent space translation via semantic alignment. *Advances in Neural Information Processing Systems*, 36:55394–55414, 2023.

Ari Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in neural information processing systems*, 31, 2018.

Nikita Moriakov, Jonas Adler, and Jonas Teuwen. Kernel of CycleGAN as a principal homogeneous space. In *International Conference on Learning Representations*, 2020.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *International Conference on Learning Representations*, 2023.

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

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Geoffrey Roeder, Luke Metz, and Durk Kingma. On linear identifiability of learned representations. In *International Conference on Machine Learning*, pp. 9030–9039. PMLR, 2021.

Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational autoencoders pursue PCA directions (by accident). In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12398–12407. IEEE, 2019.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG)*, 36(6):245:1–245:17, 2017. doi: 10.1145/3130800.3130883.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit Haim Bermano. Human motion diffusion model. In *International Conference on Learning Representations*, 2023.

Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. Neural kinematic networks for unsupervised motion retargetting. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 8639–8648. IEEE, 2018.

Julius Von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. *Advances in neural information processing systems*, 34:16451–16467, 2021.

Katsu Yamane, Yuka Ariki, and Jessica Hodgins. Animating non-humanoid characters with human motion data. In *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pp. 169–178, 2010.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14730–14740, 2023.

Han Zhao, Remi Tachet Des Combes, Kun Zhang, and Geoffrey Gordon. On learning invariant representations for domain adaptation. In *International conference on machine learning*, pp. 7523–7532. PMLR, 2019.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 2242–2251. Ieee, 2017.

# A Data and Implementation

## A.1 Splits and Controlled Views

All views are derived from HumanML3D source clips. The official training split contains $23{,}384$ clips, of which $20{,}942$ have at least $64$ source frames and are eligible for training. The same requirement retains $1{,}300$ of $1{,}460$ validation clips. Evaluation includes all $4{,}384$ official test clips with at least two source frames; the training-window requirement is not applied to the test gallery. Original and mirrored clips remain within their official splits, and the test set contains $2{,}192$ original–mirror families.

The native HumanML3D feature vector contains root features at indices $[0, 4)$, root-relative joint positions at $[4, 67)$, continuous 6D rotations at $[67, 193)$, local joint velocities at $[193, 259)$, and foot contacts at $[259, 263)$. Intervals use zero-based indexing and exclude their upper endpoint. The positional and rotational blocks describe $21$ non-root joints, whereas the velocity block contains $22$ joints.

Table 5: Schema definitions. ROOT, POS, ROT, VEL, and CONTACT denote the native HumanML3D feature blocks. Every row defines a separate two-interface experiment.

| Comparison | Schema A | Schema B | B width |
|---|---|---|---|
| Identical | native | Native features in original order | 263 |
| Reordered | native | Permuted joints and feature blocks | 263 |
| Positions | native | ROOT, POS, CONTACT | 71 |
| Rotations | native | ROOT, ROT, CONTACT | 134 |
| Sparse | native | ROOT, 11 non-root positions, CONTACT; every second frame | 41 |
| Disjoint | POS, 63-d | ROT without root or contact features | 126 |

Reordered uses NumPy's PCG64 generator with seed 1235 independently of the training seed. Joint feature groups are permuted intact, while ROOT and CONTACT preserve their internal order. Sparse retains non-root positional joints $\{0, 2, 4, \ldots, 20\}$ and every second source frame. Root increments and contacts are copied at those frames, so Sparse is a subsampled observation rather than a representation recomputed at the lower frame rate.

Inputs use the HumanML3D per-feature means and standard deviations, transformed according to the same channel selection and permutation as the data, with standard deviations bounded below by $10^{-6}$. All 134 Rotations channels occur in native schema A. Disjoint removes this channel overlap by comparing POS and ROT alone, so the Rotations and Disjoint comparisons should not be interpreted as isolating only the contribution of root or contact features.

### A.2 Architecture and Optimization

The main architecture contains separate input and output adapters for each schema and one shared encoder and decoder. Table 6 gives the exact framewise MLP construction. A residual block maps

$$h \mapsto h + W_2 \,\mathrm{ReLU}(W_1 h + b_1) + b_2 \tag{11}$$

without normalization or dropout. The encoder has separate linear heads for posterior mean and log-variance, yielding one latent token per input frame.

Table 6: Main MLP architecture. The default hidden and latent widths are 256 and 64.

| Module | Layers |
|---|---|
| Input adapter | Linear $d_i \to 256$, ReLU, Linear $256 \to 256$ |
| Shared encoder | Two width-256 residual blocks, followed by parallel mean and log-variance heads $256 \to 64$ |
| Shared decoder | Linear $64 \to 256$, followed by two width-256 residual blocks |
| Output adapter | Linear $256 \to 256$, ReLU, Linear $256 \to d_i$ |
| Domain classifier | Linear $128 \to 128$, LeakyReLU, Linear $128 \to 128$, LeakyReLU, Linear $128 \to 2$ |

Posterior noise is sampled independently across frames. Reconstruction and auxiliary objectives use sampled latents, KL is computed analytically from the posterior parameters, and deterministic representation comparisons use posterior means. The normalized diagonal-Gaussian KL is

$$\mathcal{L}_{\mathrm{KL}}(x) = \frac{1}{2Td_z} \sum_{t=1}^{T} \sum_{k=1}^{d_z} \left(\mu_{tk}^2 + \sigma_{tk}^2 - 1 - \log \sigma_{tk}^2\right). \tag{12}$$

Reconstruction is averaged over frames and features before averaging sequences within the mixed batch.

All runs use 200,000 optimization steps and Adam with learning rate $10^{-4}$, weight decay $10^{-4}$, and global gradient-norm clipping at 1.0. Each step draws a mixed batch of 64 unsynchronized schema windows, with 64 source frames per window or 32 retained frames for Sparse. Anchor training additionally draws a batch of 64 corresponding window pairs. The reference model uses $d_z = 64$ and $\beta = 0.3$.

The KL coefficient is increased linearly over the first 10,000 steps and auxiliary objective coefficients over the first 20,000. Default settings are $\lambda_z = \lambda_x = 1$, $\lambda_m = 1$, reversed domain-gradient weight 0.1, anchor weight 1, and temperature $\tau = 0.07$.

The temporal-convolutional network (TCN) robustness model replaces only the shared residual blocks with residual Conv1D blocks using kernel size 5, stride 1, and dilation 1. Adapters and posterior heads remain framewise. Its encoder and decoder each have a 17-token receptive field, yielding a 33-frame end-to-end receptive field. For Native/Rotations, the TCN has approximately 3.14M parameters compared with 1.04M for the MLP, so this robustness comparison changes both temporal context and parameter count.

### A.3 Objective Implementation and Sampling

Section 3.2 defines the training objectives; here we specify their sampling and implementation details.

**Cycle.** Cycle training averages both interface directions. A source posterior sample is decoded through the other interface and re-encoded using a fresh posterior sample. Stop-gradient applies only

to the reference code in $\mathcal{L}_{z\mathrm{cyc}}$, while the source code entering the cross-schema decoder remains differentiable. Both cycle penalties are per-element means. Sparse cycles retain the source token count and therefore do not perform an explicit frame-rate conversion.

**MMD.** Training-time MMD uses the pooled sampled representation defined in Section 3.2 and averages five radial-basis kernels,

$$k(u,v) = \frac{1}{5} \sum_{a \in \{0.25,0.5,1,2,4\}} \exp\left( -\frac{\|u-v\|_2^2}{2(a\delta)^2} \right), \tag{13}$$

where $\delta$ is the median off-diagonal pairwise distance in the combined minibatch, detached from the gradient and bounded below by $10^{-6}$. Within-schema diagonal terms are omitted. Since this bandwidth is recomputed for every minibatch, the training objective does not correspond to one fixed population kernel; the fixed-kernel statistic used for cross-model evaluation is described separately in Appendix B.4.

**Discriminator.** The domain classifier uses LeakyReLU with negative slope $0.2$. It minimizes $\mathcal{L}_{\mathrm{dom}}$, while gradient reversal sends the negative classifier gradient, weighted by $\lambda_d$, to the encoder and input adapters. The shared decoder and output adapters receive no direct gradient from the domain objective.

**Anchor.** Paired clips are sampled with replacement and the two windows in each pair use the same source interval. If $c_k$ is the identity associated with pair $k$, the InfoNCE candidate set is

$$\mathcal{C}_k = \{k\} \cup \{\ell \mid c_\ell \neq c_k\}, \tag{14}$$

so repeated occurrences of the same clip identity are not treated as negatives. There is no negative queue. The Anchor objective updates both input adapters and the shared encoder through sampled latent representations.

The paired-MSE ablation replaces the contrastive objective with direct squared-error attraction between the same pooled paired representations, leaving the paired-window sampling and representation granularity unchanged.

### A.4 Pair Budgets and Checkpoint Policy

The primary limited-pair study uses nested prefixes of a fixed random permutation of the training clips generated with subset seed 777. The 16-pair set is therefore contained in the 64-, 256-, and 1,024-pair sets. A budget denotes the number of unique paired clip identities, not the number of positive windows seen during optimization. Training seeds $42$, $1$, and $2$ vary optimization independently of the subset seed.

Validation-selected Anchor checkpoints would introduce paired validation examples outside the stated limited training budget. Consequently, limited-budget Anchor results use the fixed final checkpoint at step $200{,}000$, avoiding additional correspondence information during checkpoint selection. All-pairs Anchor references use validation-selected checkpoints, while Floor, Cycle, MMD, and Discriminator runs use their validation-selected checkpoints unless stated otherwise.

## B Evaluation and Calibration

### B.1 Clip Representations and Retrieval

For a clip $x$, let $m_t = [\mu_i(x)]_t$ denote its posterior-mean token at frame $t$. The full-clip representation concatenates temporal mean and population standard deviation,

$$u = \frac{1}{T}\sum_{t=1}^{T} m_t, \qquad v_k = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(m_{tk}-u_k)^2}, \qquad r = [u;v] \in \mathbb{R}^{2d_z}. \tag{15}$$

The temporal standard deviation measures variation of posterior means across frames and should not be interpreted as posterior uncertainty. Full clips are encoded directly without sliding-window blending.

Each schema B test clip queries all $4{,}384$ schema A clips using cosine similarity, with its exact source counterpart as the positive. Recall@1 is the primary metric, while Recall@5, Recall@10, mean reciprocal rank, and median rank are retained for ancillary analyses,

$$\mathrm{R@}\,K = \frac{1}{Q}\sum_{q=1}^{Q}\mathbf{1}[\pi_q \le K], \qquad \mathrm{MRR} = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{\pi_q}. \tag{16}$$

Uniform random ranking gives Recall@1 $100/4384 = 0.02281\%$. Mirrored versions of the same source motion count as errors under the primary exact-counterpart criterion and are analyzed separately in Appendix C.4.

### B.2 Direct Token Agreement

For same-rate views, synchronized posterior-mean tokens are compared without pooling, fitted transformations, centering, or cross-schema normalization. For $a = m_t^{(A)}$ and $b = m_t^{(B)}$, paired-token cosine is the primary measure, accompanied by

$$d_{\mathrm{rel}}(a,b) = \frac{\|a-b\|_2}{\sqrt{\frac{1}{2}\left(\|a\|_2^2 + \|b\|_2^2\right) + 10^{-8}}}, \tag{17}$$

and token RMSE as an additional diagnostic. Measurements are first averaged over synchronized frames within each clip and then equally across clips, preventing longer sequences from receiving greater weight.

To provide an interpretable null, we construct one fixed mismatched pairing of the $4{,}384$ test identities using a derangement generated with NumPy seed 20260926. Any fixed points are removed by swapping identities, and the same pairing is used for every checkpoint. If the mismatched target clip is shorter than the source, source frame $t$ is paired with frame $\min(t, T_{\mathrm{target}} - 1)$. The null is used only as an interpretive reference and never for fitting, checkpoint selection, or hyperparameter choice.

### B.3 Training-Only Calibration

Procrustes, ridge, and CCA are fitted after representation learning with the encoder frozen. Let $A$ and $B$ contain corresponding clip representations as rows and let $A_c, B_c$ denote their centered forms. Orthogonal Procrustes fits

$$Q = \arg\min_{Q^\top Q = I} \|B_c Q - A_c\|_F^2, \tag{18}$$

allowing reflections but no scale transformation. Ridge instead learns a general linear map,

$$W = (G + \alpha_{\mathrm{clip}} I)^{-1} B_c^\top A_c, \qquad G = B_c^\top B_c, \qquad \alpha_{\mathrm{clip}} = 10^{-3}\frac{\mathrm{tr}(G)}{p_B}. \tag{19}$$

No additional feature standardization is applied for ridge.

CCA standardizes each representation using the fitting-set mean and sample standard deviation, regularizes each within-schema covariance by $10^{-6} I$, and floors eigenvalues at $10^{-6}$ during whitening. An SVD of the whitened cross-covariance gives the two canonical projections, retaining

$$k = \min(16, m - 2, p_B, p_A) \tag{20}$$

components for $m$ fitting observations. Canonical coordinates are compared using cosine similarity without weighting by their canonical correlations, and no pseudoinverse or additional numerical-rank threshold is used.

The component cap and numerical regularization are fixed across the reported experiments and are not selected using test retrieval. The available records do not establish the historical rationale for choosing 16 components, so we do not describe this value as selected a priori.

Unless a reduced budget is explicitly reported, calibration uses all $20{,}942$ eligible training correspondences. All preprocessing statistics and fitted maps are estimated exclusively from training data and applied unchanged to the test split. Reduced-budget analyses use fixed training subsets; test-fold

calibration occurs only in explicitly identified ancillary sanity checks and is never mixed with the primary protocol.

For pair-budget comparisons with Anchor, post-hoc maps use the same nested seed-777 clip identities as Anchor training. Window-based calibration uses eight aligned 64-frame windows per identity, matching the identity budget and observation granularity, although repeated pair exposure and optimization computation are not matched. Complete trained-versus-random and pair-budget results are reported in Appendix C.3.

### B.4 Marginal Diagnostics

The held-out schema probe is independent of the domain classifier used during representation learning. Within each fold, a StandardScaler is fit before logistic regression with L2 regularization and $C = 1$. Five unshuffled GroupKFold splits keep original and mirrored versions of each motion in the same fold. Balanced accuracy and AUC are averaged over folds, and no probe hyperparameter is selected from the resulting test scores.

Cross-model MMD comparisons use a fixed evaluation kernel rather than the adaptive bandwidth used by the training loss. Its reference bandwidth,

$$\delta_{\mathrm{ref}} = 4.484861, \tag{21}$$

is the median off-diagonal Euclidean distance among $2{,}000$ schema A and $2{,}000$ schema B training representations from the Rotations Floor seed-42 reference, sampled with generator seed 0. The RBF bandwidth multipliers $\{0.25, 0.5, 1, 2, 4\}$ are then held fixed for every evaluated model.

Evaluation uses the same disjoint test halves across checkpoints, obtained from a seed-0 permutation of the test set, and computes the MMD U-statistic without clipping negative estimates. The resulting statistic therefore uses a common kernel across models, while the adaptive-bandwidth statistic associated with the training objective is examined separately as a consistency check.

### B.5 Decoder and Token-Map Diagnostics

Decoder compatibility is evaluated as a directed $B \to A$ operation. Native decoding uses

$$G_A\big(\mu_A(x^A)\big)\,, \tag{22}$$

while crossed decoding feeds schema B posterior means directly to the schema A decoder,

$$G_A\big(\mu_B(x^B)\big)\,. \tag{23}$$

Predictions and paired schema A targets are denormalized using schema A statistics. Local MPJPE is computed directly over the $21$ non-root RIC position channels, excluding root translation and orientation and without forward kinematics, rigid alignment, or global trajectory recovery. Every reported condition uses the same $512$ test clips, selected once with NumPy generator seed 0.

A separate token-ridge diagnostic tests whether poor direct decoder reuse can be reduced by a fitted linear relation between latent tokens. The map is estimated from $200{,}000$ aligned training-frame pairs sampled from $6{,}250$ training clips using centered ridge regression with $\alpha_{\mathrm{token}} = 1$. Before decoding, mapped schema B tokens are transformed as

$$\widetilde{\mu}_B = (\mu_B - \bar{\mu}_B)W + \bar{\mu}_A. \tag{24}$$

This diagnostic uses substantially more paired identities than the 64-pair Anchor experiment and is therefore intended to test functional recoverability rather than provide a matched-budget comparison.

### B.6 Reporting and Reproducibility

Across training seeds performance bounds are reported as descriptive minima and maxima rather than parametric confidence intervals. Single-seed ablations are explicitly identified, and evaluations across multiple evaluation windows or interfaces derived from a single trained checkpoint are treated as dependent evaluations rather than independent training replications.

Implementation details, exported metrics, and saved configurations were validated against source files and checkpoint metadata during post-experiment auditing. Environment verification and analysis

were executed using PyTorch 2.10.0+cu128, NumPy 2.2.6, SciPy 1.17.0, and scikit-learn 1.8.0. For dimensionality reduction, Canonical Correlation Analysis (CCA) was fixed at 16 components to balance variance coverage with computational efficiency across runs. Finally, to maintain conservative attribution, observed correspondences are reported without isolating shared backbone effects, as an independently trained Native-versus-Rotations baseline was outside the scope of this study.

# C Complete Retrieval and Calibration Results

## C.1 Per-Seed Controlled Comparisons

Tables 7 and 8 give the individual training seeds underlying the ranges reported in the main paper. Every mapped result in these tables uses all $20{,}942$ eligible training correspondences and the representation model remains frozen during calibration.

Table 7: Floor retrieval for every controlled view and training seed. Entries are full-gallery $B \to A$ Recall@1 percentages. Each experiment contains two interfaces sharing one backbone; Disjoint uses the positional representation as schema A.

| View pair | Seed | Raw | Proc. | Ridge | CCA |
|---|---|---|---|---|---|
| Identical | 42 | 10.65 | 98.36 | 99.11 | 99.86 |
| Identical | 1 | 5.91 | 96.30 | 97.88 | 99.64 |
| Identical | 2 | 17.72 | 98.02 | 98.77 | 99.86 |
| Reordered | 42 | 0.30 | 94.37 | 97.54 | 99.79 |
| Reordered | 1 | 1.46 | 98.15 | 98.77 | 99.77 |
| Reordered | 2 | 36.82 | 98.61 | 99.04 | 99.91 |
| Positions | 42 | 0.94 | 58.65 | 60.74 | 95.85 |
| Positions | 1 | 0.11 | 61.95 | 63.59 | 96.10 |
| Positions | 2 | 0.16 | 60.06 | 61.13 | 95.78 |
| Rotations | 42 | 0.32 | 82.66 | 87.66 | 96.30 |
| Rotations | 1 | 1.64 | 78.85 | 82.89 | 96.24 |
| Rotations | 2 | 0.16 | 82.30 | 84.92 | 95.78 |
| Sparse | 42 | 0.30 | 44.07 | 45.28 | 90.90 |
| Sparse | 1 | 0.21 | 47.35 | 47.38 | 91.95 |
| Sparse | 2 | 0.36 | 46.17 | 45.76 | 91.40 |
| Disjoint | 42 | 0.14 | 21.26 | 31.32 | 57.62 |
| Disjoint | 1 | 0.21 | 19.73 | 28.60 | 57.19 |
| Disjoint | 2 | 0.30 | 19.32 | 29.77 | 55.66 |

## C.2 Additional Controlled Views

Table 9 completes the Disjoint and Sparse comparisons used to test whether the direct-versus-recoverable separation persists under more restrictive observations. Disjoint removes shared input channels, while Sparse reduces both the observed joint set and temporal resolution. Disjoint Anchor uses all available training correspondences and therefore does not extend the limited-pair Rotations result to a low-overlap setting.

Cycle substantially increases calibrated retrieval on both views while leaving raw retrieval weak. In particular, its improvement on Disjoint does not require the two interfaces to share input channels. Full-pair Anchor instead produces strong direct retrieval on Disjoint, showing that explicit correspondences can identify a compatible relation even for these non-overlapping feature blocks when paired supervision is abundant.

## C.3 Calibration Budgets and Random Features

Mapped retrieval depends jointly on the frozen representation, the calibration family, and the number of paired examples available to fit it. To quantify this dependence, we compare trained Floor

Table 8: Rotations objective comparisons for individual training seeds. Maps use all eligible training correspondences. MMD uses $\lambda_m = 1$, and Anchor uses all training pairs. Entries are Recall@1 percentages.

| Condition | Seed | Raw | Proc. | Ridge | CCA |
|---|---|---|---|---|---|
| Cycle | 42 | 0.11 | 94.37 | 95.80 | 99.16 |
| Cycle | 1 | 0.16 | 94.89 | 96.37 | 99.20 |
| Cycle | 2 | 0.09 | 94.25 | 95.94 | 99.25 |
| MMD | 42 | 0.23 | 83.07 | 87.18 | 97.08 |
| MMD | 1 | 0.87 | 82.07 | 86.09 | 97.22 |
| MMD | 2 | 2.85 | 80.18 | 84.24 | 97.67 |
| Discriminator | 42 | 44.14 | 85.26 | 87.89 | 96.81 |
| Discriminator | 1 | 12.55 | 82.66 | 85.77 | 96.58 |
| Discriminator | 2 | 17.56 | 82.92 | 88.59 | 97.42 |
| Anchor, all pairs | 42 | 100.00 | 100.00 | 100.00 | 100.00 |
| Anchor, all pairs | 1 | 100.00 | 100.00 | 100.00 | 100.00 |
| Anchor, all pairs | 2 | 100.00 | 100.00 | 100.00 | 100.00 |

representations with untrained encoders of the same architecture under matched calibration budgets. The trained ranges cover seeds $42$, $1$, and $2$, while random encoders use initialization seeds $0$, $1$, and $2$; these seeds are independent and are not treated as paired replications.

Table 9: Additional controlled comparisons. Entries are Recall@1 percentages and ranges are minima and maxima across three training seeds. All calibration maps use the full set of eligible training correspondences.

| View | Condition | Raw | Proc. | Ridge | CCA |
|---|---|---|---|---|---|
| Disjoint | Floor | 0.14–0.30 | 19.32–21.26 | 28.60–31.32 | 55.66–57.62 |
| Disjoint | Cycle | 0.14–0.32 | 50.05–55.02 | 60.47–63.05 | 77.08–78.49 |
| Disjoint | Anchor, all | 99.59–99.73 | 99.61–99.75 | 99.70–99.82 | 99.16–99.57 |
| Sparse | Floor | 0.21–0.36 | 44.07–47.35 | 45.28–47.38 | 90.90–91.95 |
| Sparse | Cycle | 0.05–0.18 | 61.06–61.45 | 72.49–73.70 | 96.08–96.94 |

Table 10: Training-only calibration as a function of correspondence budget. Values report median Recall@1 followed by min–max across three trained or random seeds. $\Delta$R@1 is the difference between trained and random medians. The $3{,}507$-pair condition is an intermediate sensitivity analysis rather than part of the primary comparison.

| View | Map | Pairs | Trained Floor | Random encoder | $\Delta$R@1 |
|---|---|---|---|---|---|
| Rotations | Proc. | 256 | 78.49 (75.57–78.70) | 24.04 (22.31–26.69) | +54.45 |
| Rotations | Proc. | 3,507 | 82.19 (78.44–82.78) | 34.83 (31.84–37.09) | +47.35 |
| Rotations | Proc. | 20,942 | 82.30 (78.85–82.66) | 36.68 (33.10–38.18) | +45.62 |
| Rotations | CCA | 256 | 67.24 (67.04–69.53) | 6.84 (6.64–7.82) | +60.40 |
| Rotations | CCA | 3,507 | 95.76 (95.39–96.05) | 69.48 (68.84–71.12) | +26.28 |
| Rotations | CCA | 20,942 | 96.24 (95.78–96.30) | 73.63 (72.15–73.79) | +22.60 |
| Disjoint | Proc. | 256 | 14.80 (13.96–15.31) | 4.43 (2.01–5.09) | +10.38 |
| Disjoint | Proc. | 3,507 | 19.32 (18.75–20.10) | 8.69 (5.47–9.76) | +10.63 |
| Disjoint | Proc. | 20,942 | 19.73 (19.32–21.26) | 9.97 (6.41–10.45) | +9.76 |
| Disjoint | CCA | 256 | 11.95 (11.59–12.11) | 1.62 (1.48–1.82) | +10.33 |
| Disjoint | CCA | 3,507 | 54.88 (53.76–56.52) | 27.87 (27.10–29.43) | +27.01 |
| Disjoint | CCA | 20,942 | 57.19 (55.66–57.62) | 31.87 (30.52–33.51) | +25.32 |

The Rotations comparison illustrates why calibration budget is important for interpretation. With only $256$ training correspondences, CCA retrieval differs by approximately $60$ percentage points

between trained Floor and random representations. With all 20,942 pairs, the random baseline itself reaches approximately 74%, reducing the corresponding gap to about 23 points. Procrustes remains substantially more selective to learned structure at the full fitting budget.

Random representations therefore retain clip-identifying information that sufficiently flexible paired calibration can exploit, particularly with many fitting examples. High mapped retrieval should consequently be interpreted as correspondence accessible to a specified map family and pair budget rather than as an intrinsic measure of representation quality. At the same time, the large trained-versus-random gaps, especially under restricted calibration, show that the learned representations contain substantially more accessible cross-schema structure than random initialization alone.

### C.4 Retrieval Rank Distributions and Sanity Checks

Recall@1 is the primary retrieval metric in the main paper, but the direct-versus-calibrated separation extends across the full rank distribution. For the Rotations Floor seed-42 reference, raw retrieval places the true counterpart at median rank 1,140, with Recall@5 1.03%, Recall@10 1.82%, and MRR 0.011. Procrustes and CCA both reduce the median rank to 1, with MRR 0.889 and 0.980, respectively. The calibration effect is therefore not specific to the top-1 decision threshold.

Simple marginal corrections do not explain the Identical result. Centering and standardization with schema-specific training statistics leave every Identical Floor seed below 20% Recall@1. Reordered is more sensitive to scale in one run, where standardization raises the seed-2 score from 36.82% to 57.69%, but this remains well below its paired calibration results.

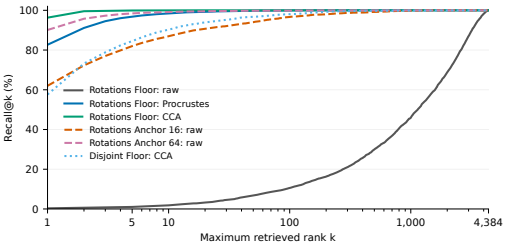



Figure 3: Full-gallery retrieval rank curves for representative conditions.

Several additional controls bound alternative explanations for retrieval. A reference supplied directly with known raw-channel correspondence retrieves 100% of counterparts from root features, 99.8% from contacts, and 100% from root plus contacts, showing that a small known feature subset can carry strong clip identity. These controls are given the feature correspondence explicitly and therefore do not constitute learned cross-schema alignment. Duration-only retrieval reaches approximately 4.4% Recall@1.

A separate shuffled-pair calibration control gives 0.00% Recall@1 under Procrustes and 0.02% under CCA, confirming that the calibrated results depend on the true correspondence structure rather than only on the fitting procedure. This control uses test-fold calibration and is reported only as a transductive sanity check, not as part of the primary training-only protocol.

Mirror substitutions account for some residual errors after calibration but do not explain the poor raw retrieval. In the Rotations Floor seed-42 reference, CCA retrieves the mirrored clip in 48 of its 162 errors, whereas raw retrieval retrieves a mirror in only 45 of 4,370 errors. Mirror ambiguity therefore becomes noticeable once most exact counterparts are already recovered, rather than serving as the primary cause of weak direct compatibility.

## D Direct Token and Marginal Diagnostics

### D.1 Synchronized Token Agreement

Table 11 reports the direct token diagnostics introduced in Section 5.2. No map or cross-schema normalization is fitted. The values are min–max across the three training seeds and use clip-equal aggregation.

The Identical control establishes that weak raw retrieval is accompanied by direct token mismatch even when both interfaces receive the same observations. Rotations Floor shows the same behavior under a feature-selection view. Cycle is especially informative: corresponding tokens remain close to the mismatched reference even though calibrated clip retrieval is almost perfect. Anchor-64 instead

Table 11: Direct synchronized-token agreement over three seeds. The fixed mismatched-identity reference spans $-0.057$–$0.019$ for cosine and $1.388$–$1.447$ for relative $L_2$ across conditions.

| Condition | Raw R@1 ↑ | Cosine ↑ | Rel. $L_2$ ↓ |
|---|---|---|---|
| Identical Floor | 5.91–17.72 | 0.221–0.484 | 0.989–1.227 |
| Reordered Floor | 0.30–36.82 | 0.084–0.706 | 0.706–1.338 |
| Rotations Floor | 0.16–1.64 | 0.114–0.233 | 1.224–1.319 |
| Rotations Cycle | 0.09–0.16 | $-0.025$–0.026 | 1.391–1.427 |
| Rotations Anchor-64 | 88.41–90.33 | 0.824–0.826 | 0.562–0.566 |

raises token cosine to approximately $0.82$, showing that pair identities alter the direct token relation rather than merely the pooled descriptor.

Across these 15 evaluated checkpoints, paired-token cosine and raw clip Recall@1 have a descriptive Spearman correlation of approximately $0.98$; over the nine Floor checkpoints alone it is approximately $0.97$. These correlations are descriptive rather than independent replication tests, but they provide no evidence that the low raw retrieval of the unpaired models is created solely by temporal pooling.

### D.2 Marginal Discrepancy and Schema Separability

Table 12 compares raw paired retrieval with the held-out schema probe and fixed-kernel MMD, using the same evaluation kernel across models.

Table 12: Marginal diagnostics on Rotations. Lower probe accuracy indicates weaker schema separability, while lower MMD indicates smaller measured marginal discrepancy. Fixed MMD$^2$ values are shown in units of $10^{-3}$.

| Condition | Raw R@1 ↑ | Probe acc. ↓ | Fixed MMD$^2$ ↓ |
|---|---|---|---|
| Floor | 0.16–1.64 | 99.91–99.97 | 13.04–15.78 |
| Cycle | 0.09–0.16 | 99.93–99.98 | 50.76–63.88 |
| MMD ($\lambda_m = 1$) | 0.23–2.85 | 99.95–99.99 | 6.14–9.81 |
| Discriminator | 12.55–44.14 | 99.94–100.00 | 16.40–21.54 |
| Anchor (all pairs) | 100.00 | 99.93–100.00 | 0.36–0.48 |

Single-seed MMD sensitivity runs with $\lambda_m = 0.1$ and $0.3$ give fixed MMD$^2$ values of $12.88$ and $11.29 \times 10^{-3}$, respectively, without improving direct retrieval. Adaptive-kernel MMD shows the same qualitative ordering (Table values omitted for clarity).

Using the common fixed kernel, MMD with $\lambda_m = 1$ reduces the measured discrepancy by roughly 40–55% relative to Floor, while Cycle increases it substantially. The adaptive and fixed-kernel values give the same qualitative ordering, so the observed reduction is not caused only by using a different adaptive bandwidth for each model.

No evaluated checkpoint becomes schema-indistinguishable. Balanced accuracy remains at least $99.9\%$ with AUC at least $0.9999$, including the strongest MMD condition and all-pairs Anchor. The empirical conclusion is therefore limited to the statement that reducing a measured marginal discrepancy does not establish paired correspondence or direct retrieval compatibility. The experiments do not provide a trained example of fully matched marginals with incorrect pair correspondence.

All-pairs Anchor gives the smallest measured MMD and perfect raw retrieval while remaining almost perfectly schema-classifiable. Direct counterpart compatibility and schema invariance are therefore distinct in the tested representations.

# E PAIRED-SUPERVISION CONTROLS

## E.1 CHECKPOINT SENSITIVITY

Table 13 compares validation-selected and fixed-final checkpoints for every nested Anchor budget. At the smallest budget the selected checkpoint is very early because its objective uses paired validation information. The fixed-final policy used in the main limited-budget analysis avoids that additional correspondence signal; it does not imply that final checkpoints are always more accurate.

Table 13: Nested Rotations Anchor budgets. Primary limited-budget results use the final 200,000-step checkpoint.

| Clip pairs | Seed | Selected step | Selected raw | Final raw |
|---|---|---|---|---|
| 16 | 42 | 2,000 | 18.23 | 62.00 |
| 16 | 1 | 2,000 | 25.82 | 69.30 |
| 16 | 2 | 2,000 | 25.18 | 75.75 |
| 64 | 42 | 4,000 | 85.13 | 90.10 |
| 64 | 1 | 4,000 | 86.20 | 88.41 |
| 64 | 2 | 4,000 | 85.70 | 90.33 |
| 256 | 42 | 6,000 | 99.04 | 94.53 |
| 256 | 1 | 6,000 | 98.91 | 94.07 |
| 256 | 2 | 6,000 | 99.20 | 94.98 |
| 1,024 | 42 | 20,000 | 99.98 | 99.22 |
| 1,024 | 1 | 12,000 | 100.00 | 99.02 |
| 1,024 | 2 | 14,000 | 99.89 | 99.32 |

## E.2 PAIR-BUDGET-MATCHED CALIBRATION

Table 14 gives the complete numerical comparison behind Figure 2(a). Calibration and Anchor use the same nested seed-777 identities. Full-clip fitting has one row per identity, while window fitting uses eight aligned rows per identity. This matches identity budget and representation granularity but not repeated pair exposure or optimization computation.

Table 14: Pair-budget comparison on the same nested identities. Map columns span three Floor seeds; Anchor spans three independently trained Anchor runs and uses final checkpoints. Ranges are min–max.

| Identities | Fit unit | Proc. | Ridge | CCA | Anchor raw |
|---|---|---|---|---|---|
| 16 | Clip | 10.36–15.56 | 15.40–16.97 | 4.56–5.82 | 62.00–75.75 |
| 16 | Window | 38.32–43.57 | 26.48–31.20 | 0.02–0.07 | 62.00–75.75 |
| 64 | Clip | 62.16–65.99 | 29.20–43.34 | 41.51–48.59 | 88.41–90.33 |
| 64 | Window | 66.99–70.46 | 62.02–70.96 | 65.42–70.05 | 88.41–90.33 |
| 256 | Clip | 75.48–79.20 | 77.69–83.90 | 62.18–71.88 | 94.07–94.98 |
| 256 | Window | 76.25–80.54 | 79.06–84.35 | 88.30–89.90 | 94.07–94.98 |
| 1,024 | Clip | 77.78–81.87 | 81.61–86.79 | 93.64–95.35 | 99.02–99.32 |
| 1,024 | Window | 78.56–82.05 | 81.48–86.15 | 93.84–95.23 | 99.02–99.32 |

The especially weak 16-identity window CCA fit should not be interpreted as a general failure of CCA. Its 128 fitting rows come from only 16 unique identities, and no conditioning diagnostics were retained for this run. No calibration family is selected using its test score.

## E.3 SUBSET AND PAIR-LOSS SENSITIVITY

Changing the selected paired identities introduces additional variation. At training seed 42, the three tested 64-identity subsets yield $85.68$–$90.10\%$ raw Recall@1, while the corresponding 256-identity subsets yield $94.32$–$95.26\%$. This sensitivity is comparable to or larger than optimization-seed variation at 64 pairs, so the limited-pair numbers should be interpreted as results for the tested nested subsets rather than a precise sample-complexity estimate.

Table 15: Pair-objective sensitivity on Rotations, training seed 42 and subset seed 777. All entries use final checkpoints.

| Clip pairs | Pair objective | Weight | Raw R@1 |
|---|---|---|---|
| 64 | InfoNCE | 0.1 | 90.44 |
| 64 | InfoNCE | 0.3 | 89.55 |
| 64 | InfoNCE | 1.0 | 90.10 |
| 256 | InfoNCE | 0.1 | 95.42 |
| 256 | InfoNCE | 0.3 | 94.41 |
| 256 | InfoNCE | 1.0 | 94.53 |
| 256 | Paired MSE | 1.0 | 97.92 |

The paired-MSE run is single-seed evidence. It shows only that the strong direct compatibility observed under paired supervision does not require an InfoNCE denominator in this tested condition; it does not establish that MSE is generally superior or more robust.

# F Objective and Architecture Robustness

## F.1 Cycle Components

Both cycle terms contribute to calibrated correspondence in the seed-42 Rotations ablation. Neither term alone nor their combination produces strong raw retrieval, while the combined objective gives the strongest mapped retrieval in this run.

Table 16: Cycle-component ablation on Rotations, seed 42. Maps use all training correspondences.

| Condition | Raw | Proc. | Ridge | CCA |
|---|---|---|---|---|
| Floor | 0.32 | 82.66 | 87.66 | 96.30 |
| Latent cycle only | 0.14 | 88.62 | 92.18 | 98.08 |
| Motion cycle only | 0.09 | 83.90 | 87.14 | 97.49 |
| Both cycles | 0.11 | 94.37 | 95.80 | 99.16 |

This single-seed analysis establishes sensitivity to the two cycle terms but does not provide evidence for a statistically verified interaction or for a universal property of cycle-consistent objectives.

## F.2 Temporal Backbone

The TCN preserves the qualitative separation between direct and calibrated correspondence. Floor and Cycle remain weak under raw retrieval while calibrated correspondence becomes high, and all-pairs Anchor again produces direct compatibility.

Table 17: Backbone robustness on Rotations. Each range covers three training seeds. The TCN comparison is not parameter matched.

| Condition | Backbone | Raw | Proc. | Ridge | CCA |
|---|---|---|---|---|---|
| Floor | MLP | 0.16–1.64 | 78.85–82.66 | 82.89–87.66 | 95.78–96.30 |
| Floor | TCN | 0.16–1.98 | 93.77–94.48 | 95.03–95.21 | 99.50–99.82 |
| Cycle | MLP | 0.09–0.16 | 94.25–94.89 | 95.80–96.37 | 99.16–99.25 |
| Cycle | TCN | 0.07–0.30 | 97.31–98.04 | 98.08–98.59 | 99.82–99.91 |
| Anchor, all | MLP | 100.00 | 100.00 | 100.00 | 100.00 |
| Anchor, all | TCN | 100.00 | 100.00 | 100.00 | 99.98–100.00 |

The result shows that the direct-versus-recoverable distinction is not confined to the framewise MLP, although the TCN simultaneously changes temporal context and parameter count.

# G Decoder Compatibility

## G.1 Matched-Checkpoint Cross-Decoding

Table 18 summarizes matched-checkpoint native and crossed decoding for the conditions used in the token analysis. Every range covers the same three training seeds as the corresponding retrieval result.

Table 18: Matched-checkpoint retrieval and decoding. MPJPE is local 21-joint error in millimeters on the fixed 512-clip set.

| Condition | Raw R@1 | Native MPJPE | Crossed MPJPE |
|---|---|---|---|
| Identical Floor | 5.91–17.72 | 42.99–43.36 | 115.20–199.51 |
| Reordered Floor | 0.30–36.82 | 42.93–43.56 | 65.47–214.70 |
| Rotations Floor | 0.16–1.64 | 42.98–43.72 | 198.50–221.52 |
| Rotations Cycle | 0.09–0.16 | 33.08–33.35 | 433.50–509.66 |
| Rotations Anchor-64 | 88.41–90.33 | 43.38–44.45 | 73.87–79.83 |

Identical is particularly diagnostic because its decoding gap cannot be attributed to one schema observing less source information. Rotations additionally changes the observations, so its crossed-decoding gap should not be attributed exclusively to coordinate mismatch.

For reference, the seed-42 Anchor checkpoints at 256 and 1,024 paired identities give raw retrieval $94.53\%$ and $99.22\%$, crossed MPJPE $69.05$ and $60.30$ mm, and native MPJPE $44.51$ and $43.75$ mm, respectively. These higher-budget decoder results are single-seed references rather than replicated decoding curves.

## G.2 Token Calibration

A fitted token-level ridge map tests whether poor direct decoder reuse is recoverable through a simple linear transformation. Table 19 reports the existing seed-42 diagnostics.

Table 19: Token-calibrated decoding, seed 42. Ridge is fitted on 200,000 paired frames from 6,250 training clips. Entries are local MPJPE in millimeters.

| Condition | Native | Direct cross | Ridge cross |
|---|---|---|---|
| Floor | 43.72 | 205.59 | 92.08 |
| Cycle | 33.16 | 496.71 | 82.28 |
| Anchor, $n = 64$ | 43.38 | 74.23 | 72.48 |

Floor and Cycle improve substantially after the fitted token map, showing that useful decoder-level information can remain recoverable even when direct reuse is poor. Anchor-64 changes only from $74.23$ to $72.48$ mm, so its remaining decoder gap is not explained by a simple global linear correction of token coordinates. This analysis should not be interpreted as a pair-budget comparison because the token map uses many more training identities than Anchor-64.

# H Coordinate Reparameterization Analysis

## H.1 Reconstruction and Prior Invariance

Consider the effective posterior $q_i$ and native decoder $G_i$ of interface $i$. Let $P_i$ be a signed permutation applied to every latent frame and define

$$\mu_i' = P_i \mu_i, \qquad \Sigma_i' = P_i \Sigma_i P_i^\top, \qquad G_i'(z') = G_i(P_i^{-1} z'). \tag{25}$$

A signed permutation is orthogonal and maps every diagonal covariance matrix to another diagonal covariance matrix, so the transformed posterior remains in the same variational family.

Couple the original and transformed posterior samples by $z' = P_i z$. Then $G'_i(z') = G_i(z)$ for every sample, so native reconstruction is unchanged. For one latent token, KL to the unit Gaussian is

$$D_{\mathrm{KL}}\left(\mathcal{N}(\mu, \Sigma) \,\|\, \mathcal{N}(0, I)\right) = \frac{1}{2}\left(\|\mu\|_2^2 + \operatorname{tr}\Sigma - \log\det\Sigma - d_z\right). \tag{26}$$

Orthogonality preserves the mean norm, trace, and determinant, so the KL is unchanged as well. Independent signed permutations therefore preserve the Floor objective at the level of the effective posterior–decoder functions.

Signed permutations preserve the complete diagonal-Gaussian posterior family. Special posterior families can admit larger symmetry groups; for example, isotropic covariance is preserved by arbitrary orthogonal transformations. This argument does not claim that independently learned representations differ only by signed permutations.

### H.2 Shared-Architecture Qualification

Effective-function invariance is not automatically a parameter-space symmetry of every shared neural architecture. For independent $P_i$, an exact architectural claim requires the jointly representable family of effective posteriors and decoders to remain closed under Equation 25. Shared nonlinear parameters can restrict which combinations are realizable.

The invariance argument therefore identifies freedom in the objective under this closure condition. The empirical comparisons directly test what the implemented architecture learns and do not require every independent reparameterization to be realizable. A common coordinate change across all interfaces is also different from independent relative changes, since only the latter alter cross-interface comparisons.

### H.3 Stochastic Cycles and Correspondence

Cycle training resamples the re-encoded posterior. For a fixed source vector $z$, suppose the cross-interface posterior has mean $m_j(z)$ and covariance $V_j(z)$. Its expected latent residual is

$$\mathbb{E}\left[\|z - \widetilde{z}\|_2^2 \mid z\right] = \|z - m_j(z)\|_2^2 + \operatorname{tr} V_j(z). \tag{27}$$

An identity mean round trip therefore does not make the sampled cycle loss zero, and stop-gradient changes the propagated derivative rather than this forward expected value.

A simple construction nevertheless shows why stochastic round trips need not identify the intended cross-schema correspondence. Let two identical input schemas satisfy $x \sim \mathcal{N}(0, I_d)$, choose $0 < \eta < 1$, and set $a = \sqrt{1-\eta^2}$. For arbitrary signed permutations $P_i$, define

$$q_i(z \mid x) = \mathcal{N}(aP_i x, \eta^2 I_d), \qquad G_i(z) = a^{-1} P_i^\top z. \tag{28}$$

Each aggregate posterior is exactly $\mathcal{N}(0, I_d)$ independently of $P_i$. With independent $\epsilon, \epsilon' \sim \mathcal{N}(0, I_d)$,

$$z = aP_i x + \eta\epsilon, \qquad \widetilde{z} = aP_j G_j(z) + \eta\epsilon' = z + \eta\epsilon'. \tag{29}$$

The expected per-element reconstruction, latent-cycle, and motion-cycle losses are respectively $\eta^2/a^2$, $\eta^2$, and $2\eta^2/a^2$, independent of the relative permutations, and the per-input KL is independent of $P_i$. However, decoding the source posterior mean through interface $j$ gives

$$G_j(aP_i x) = P_j^\top P_i x, \tag{30}$$

which need not equal the paired target $x$. This is an analytical illustration of the remaining correspondence ambiguity rather than a claim that every trained nonlinear model realizes this family or that these solutions are global optima.

### H.4 Pooling and the Scope of Coordinate Comparisons

Temporal means commute with linear transformations, whereas componentwise temporal standard deviations generally do not. Under a general token transformation, the variance of one transformed coordinate depends on the full covariance between original coordinates, not only on their individual

standard deviations. Consequently, an orthogonal map between pooled mean–standard-deviation descriptors does not imply an orthogonal map between framewise latent tokens.

Signed permutations are an exception. For column-vector tokens and the pooling operator $\rho$,

$$\rho(Pz) = \begin{bmatrix} P & 0 \\ 0 & |P| \end{bmatrix} \rho(z), \tag{31}$$

where $|P|$ is the entrywise absolute value. The block transformation is orthogonal, so a common signed permutation preserves cosine comparisons between pooled descriptors, while independent interface transformations need not preserve their cross-interface comparisons.

The empirical claims follow the same hierarchy. Raw clip retrieval measures direct compatibility under the chosen pooling and similarity rule, while synchronized-token diagnostics measure numerical agreement before pooling. Calibrated retrieval measures correspondence accessible to a particular fitted transformation and correspondence budget. Cross-schema decoding imposes a stronger functional requirement. None of these measurements alone establishes equality of complete latent geometries, and the paper therefore reports them as distinct operational properties of a nominally shared representation.

# I Limitations and Future Work

Our study is deliberately controlled. All schemas are derived from the same HumanML3D motions, providing exact correspondence but not capturing the noise, missing observations, temporal misalignment, or distribution shift of independently collected datasets. The limited-pair alignment study is established only for Rotations; Disjoint is evaluated with full-pair supervision. Recoverable correspondence also depends on the calibration family and correspondence budget, and the decoder-compatibility analysis considers the directed $B \to A$ setting rather than bidirectional reuse. Finally, although the main findings persist with a temporal convolutional backbone, the architectures are not parameter-matched and the experiments do not isolate the causal role of backbone sharing in producing recoverable correspondence.

Future work should test whether these distinctions persist across independently collected motion datasets, heterogeneous skeletons, and asynchronous or noisy observations, and determine how little paired correspondence is required when schemas share no input channels. It would also be useful to study objectives that directly target decoder compatibility and to characterize when architectural constraints or additional structure can identify compatible latent coordinates without explicit motion pairs.