PAIRED FLOW MATCHING: REPRESENTATION LEARNING WITHOUT A RECONSTRUCTION BOTTLENECK

Anonymous authors
Paper under double-blind review

ABSTRACT

Variational autoencoders (VAEs) are a standard tool for representation learning, but their samples often have lower fidelity than diffusion and flow models that generate directly in observation space. Latent diffusion and latent flow matching reduce computation by generating in a VAE latent space, yet they too lose sample quality. The shared limitation is the encode-and-decode design: a compact representation should discard presentation details, but the decoder needs those details to reconstruct the observation. The representation must therefore either retain unwanted detail or incur reconstruction error. To resolve this conflict, we propose Paired Flow Matching (PairedFM), which uses representations to guide transport between full observations. From ordered pairs drawn from a single data distribution, PairedFM jointly learns a representation and a family of flows in observation space, indexed by the pair's representations. At generation, a source observation supplies the details and a target representation determines where to move. PairedFM also comes with theoretical guarantees: We characterize when its representation is identifiably and the resulting transport achieves exact recovery of the truth. Across empirical studies, we find that PairedFM learns informative representations while maintaining sample quality where VAE and latent flow matching baselines degrade.

1 INTRODUCTION

Learning representations and generating observations address two aspects of the same problem: describing the latent variables underlying the data and reproducing the observation distribution. Variational autoencoders connect these tasks by learning an encoder together with a probabilistic decoder (Kingma & Welling, 2014). Structured flow autoencoders and structured coupling also jointly learn representations and flow-based generative models (Xu et al., 2026; Sumba et al., 2026). In latent diffusion and latent flow matching, the representation instead comes from a pretrained autoencoder, and a model learned in its representation space generates representations that are decoded into observations (Rombach et al., 2022; Dao et al., 2023). Even if the latent generative model reproduces the encoded distribution exactly, generation can remain inaccurate if decoding those representations does not reproduce the observation distribution. We study how learned representations can guide generation without a separate decoder that reconstructs complete observations from representations alone.

We consider observations $Y = f^\star(X) \in \mathbb{R}^D$ with distribution P, where $X \sim P_X$ is an unobserved latent variable supported on $\mathcal{X} \subset \mathbb{R}^d$. The observation map $f^\star$ is continuous and injective on $\mathcal{X}$, and the latent dimension d may be smaller than D. Both $f^\star$ and P_X are unknown. Given independent observations from P, we seek an encoder h whose representation $h(Y)$ retains the information needed to recover X, together with a generative model that reproduces P.

We propose *paired flow matching* (PairedFM), which builds on flow matching (Lipman et al., 2023) to learn an encoder jointly with a conditional velocity field in observation space. Source and target observations Y_0, Y_1 drawn independently from P supply the representations $h(Y_0)$ and $h(Y_1)$ that condition this field. For each pair, the aim is to produce the associated target Y_1, using the complete source observation Y_0 as the initial state and the target only through its representation $h(Y_1)$. During generation, the observation-space state follows the learned velocity while both representations remain fixed.

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

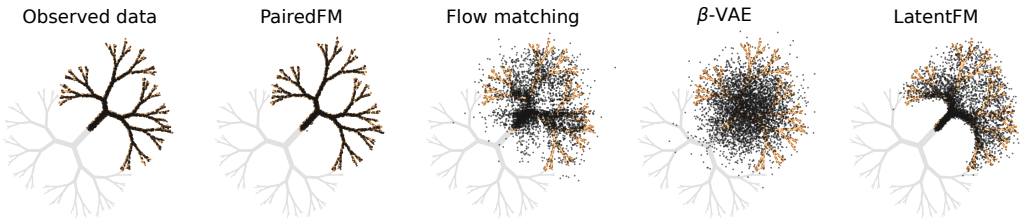


Figure 1: Generation on a branching fractal distribution. From left to right: observed data and samples generated by PairedFM, flow matching, β-VAE, and LatentFM. Each black point represents one observation or generated sample in $\mathbb{R}^2$. The orange and light-gray branches provide a common geometric reference across panels.

Our analysis asks what accurate velocity prediction requires the representation to retain. For fixed regularity bounds on the velocity field, accurate predictions on fresh independent observation pairs imply that Y can be recovered from $h(Y)$ with small mean-square error. This guarantee extends to X when $f^\star$ keeps observation distances at least proportional to the corresponding latent distances. The same prediction accuracy controls the mean-square difference between the generated output and its target observation, linking the information retained by the representation to the accuracy of generation. Since targets follow P, vanishing target error also makes the generated distribution approach P.

A branching fractal distribution makes the generation problem concrete (Figure 1). Each observation is a single point in $\mathbb{R}^2$, but probability mass is concentrated along thin branches that split at several scales. In the comparison shown, PairedFM samples closely follow these branches, whereas samples from the other algorithms are more dispersed around and between them.

Contributions.

1. **Joint representation learning and generation.** We introduce PairedFM, which jointly learns an encoder and an observation-space flow from independently sampled observation pairs through a single velocity-prediction objective. The learned source and target representations condition generation, while the full source observation initializes the flow. Training requires no reconstruction loss (Section 2).
2. **Representation and generation guarantees.** We show that the PairedFM population loss bounds recovery error from the learned representation and error in the generated output under regularity assumptions (Theorems 1 and 3). Additional regularity relates representation distances to latent distances and implies that zero-loss encoders agree up to changes of coordinates that are Lipschitz in both directions (Theorem 2 and Corollary 1).
3. **Recovery and generation with grouped pairing.** We extend PairedFM to within-group pairs while retaining the same velocity-prediction objective. We establish sufficient conditions for recovering the variables that vary within groups and generating observations that combine the target's varying variables with the source's shared variables. This allows a target representation to be used for generation across different source groups (Theorem 4).
4. **Empirical representation and generation quality.** We evaluate PairedFM on branching fractal point clouds and IFN-β stimulation in peripheral blood mononuclear cells (PBMCs) (Kang et al., 2018), and Grouped PairedFM on Shapes3D (Burgess & Kim, 2018). The results show accurate observation recovery and generation on fractals, improved stimulation-response prediction, and object-hue transfer across scenes (Section 4).

1.1 Related work

Generative representation learning. Diffusion algorithms learn representations through denoising or by conditioning generation on learned semantic representations (Preechakul et al., 2022; Mittal et al., 2023). Latent flow matching learns a flow in the representation space of a pretrained autoencoder and decodes generated representations into observations (Dao et al., 2023). Structured flow

autoencoders combine latent graphical models with conditional flows as observation models (Xu et al., 2026). Structured coupling instead places structured latent variables and additional noise in the source distribution transported by the flow (Sumba et al., 2026). Variational rectified flow matching introduces latent variables to represent multiple possible velocities at the same position and time (Guo & Schwing, 2025). PairedFM jointly learns an encoder and a velocity field from pairs of observations. The source and target representations condition a flow in observation space, and generation starts from the full source observation. Training uses velocity prediction without a separate reconstruction loss.

Pairing and transport. Conditional and multisample flow matching study how source–target couplings affect flow training and transport paths (Tong et al., 2024; Pooladian et al., 2023). Diversified flow matching learns a common translation across several pairs of conditional distributions (Shrestha & Fu, 2025). Counterfactual cocycles study transport maps for which successive changes in observed conditions agree with a direct change (Dance & Bloem-Reddy, 2025). In PairedFM, both the source and target are observations from the data distribution, and their learned representations condition the transport. The base model uses independently sampled pairs; Grouped PairedFM uses within-group pairs with the same velocity-prediction objective.

Representation recovery. Identification of nonlinear latent factors generally requires restrictions on the model or available supervision (Hyvärinen & Pajunen, 1999; Locatello et al., 2019). Auxiliary-variable and multiview algorithms obtain identification under assumptions on conditional latent distributions or observation views (Khemakhem et al., 2020; Gresele et al., 2020). Weakly supervised pairs can support recovery of latent factors (Shu et al., 2020; Locatello et al., 2020), while augmentation-based results recover content shared across views (von Kügelgen et al., 2021). Related diffusion theory studies approximate disentanglement with a known style function and editing guarantees with multiple views (Wang et al., 2025). Our analysis relates the PairedFM population loss to representation recovery and generation error under regularity assumptions. For Grouped PairedFM, we give additional sufficient conditions for recovering the variables that vary within groups and using their representations for generation across groups.

2 Paired flow matching

PairedFM separates the role of a representation from the computation used to generate an observation. The representation should distinguish data points and reflect which latent states are similar or substantially different. Generation uses a full source observation and a flow conditioned on the source and target representations. A single velocity-prediction objective trains the encoder and flow, without a separate decoder or reconstruction loss.

2.1 Problem setup and generative flow

Let $Y_1, \dots, Y_n$ be independent observations from a distribution P on $\mathbb{R}^D$, with support $\mathcal{Y}$. We consider observations of the form $Y = f^\star(X)$, where $X \sim P_X$ is supported on $\mathcal{X} \subset \mathbb{R}^d$ and $f^\star$ is continuous and injective on $\mathcal{X}$. The latent dimension d may be smaller than D. Both P_X and $f^\star$ are unknown. Our goal is to learn a representation that retains enough information to distinguish data points and reveal their latent structure, together with a flow that generates observations from P. The algorithm uses only the observed samples and forms pairs by independent sampling from the dataset.

PairedFM jointly learns an encoder and a velocity field from the classes $\mathcal{H} = \{h_\theta : \theta \in \Theta\}$ and $\mathcal{V} = \{v_\phi : \phi \in \Phi\}$. The encoders map $\mathbb{R}^D$ to $\mathbb{R}^k$, where k is the chosen representation dimension. Each velocity field takes a state in $\mathbb{R}^D$, a time in $[0, 1]$, and source and target representations in $\mathbb{R}^k$, and returns a velocity in $\mathbb{R}^D$. All maps are measurable.

The flow aims to produce a target observation from a source observation. Given a source $y_0 \in \mathcal{Y}$ and a target $y_1 \in \mathcal{Y}$, their representations condition the flow

$$\frac{dZ_t}{dt} = v_\phi(Z_t, t \mid h_\theta(y_0), h_\theta(y_1)), \qquad Z_0 = y_0.$$

Both representations remain fixed while the state Z_t evolves in observation space, and the output is Z_1. The full source observation supplies the initial state, while the target enters only through its

representation. The representation guides generation without being the sole input from which an observation must be reconstructed.

For a fixed encoder, each representation value indexes a conditional distribution of observations. This distribution describes the variation among observations assigned that representation. Under independent pairing, conditioning on the source and target representations gives independent draws from the corresponding conditional distributions. PairedFM models transport between these conditional distributions in observation space. The representations specify the source and target distributions, while the complete source observation initializes the transport.

Independent source and target draws from P induce a distribution over the output Z_1. Exact recovery of the target makes this output distribution equal to P. Grouped PairedFM uses the same flow with pairs selected using group information (Section 2.4).

2.2 Training

To train this flow, we use a path with a known velocity for each pair. Draw a source Y_0 and a target Y_1 independently from P. Their linear interpolation and its displacement are

$$Y_t = (1-t)Y_0 + tY_1, \qquad \frac{dY_t}{dt} = Y_1 - Y_0.$$

The key observation is that the displacements required for transport constrain what the representations must retain. For a fixed source and two different targets, the required displacements differ even when the interpolated states are close to the source. Spatial regularity limits how strongly the velocity can respond to these small differences in state, so accurate prediction requires information from the target representation. Section 3.2 formalizes this connection under the stated regularity assumptions.

This learning signal also shapes representation distances. If the velocity has bounded sensitivity to the target representation, similar target representations give similar velocity predictions at the same state, time, and source representation. The encoder and observation-map bounds keep nearby latent points nearby in representation space. Theorem 2 combines these constraints to relate representation distances to latent distances, up to multiplicative constants and an error controlled by the population loss. The same loss controls generation error (Theorem 3), linking the geometry learned by the encoder to the accuracy of the generated observations.

We train the encoder and velocity field together to predict this displacement from Y_t, time, and the source and target representations. For $h \in \mathcal{H}$ and $v \in \mathcal{V}$, the population loss is

$$R(h, v) = \mathbb{E}_{(Y_0,Y_1)\sim P\otimes P} \int_0^1 \left\| v(Y_t, t \mid h(Y_0), h(Y_1)) - (Y_1 - Y_0) \right\|^2 dt. \tag{1}$$

Gradients pass through both representations, so the same displacement loss trains the encoder and the observation-space velocity without a separate reconstruction objective.

Algorithm 1 applies stochastic gradient updates to the empirical counterpart of (1) by sampling pairs and times. Both observations are drawn independently from the dataset, with replacement. Conditional on the data and current parameters, the minibatch loss $\widehat{R}$ is unbiased for the loss under this empirical product distribution and uniform time sampling. Let $\widehat{h} = h_{\widehat{\theta}}$ and $\widehat{v} = v_{\widehat{\phi}}$ denote the resulting networks.

2.3 Generation

After training, we apply the flow in Section 2.1 with the learned encoder $\widehat{h}$ and velocity $\widehat{v}$. For independent source and target draws from P, Theorem 3 gives conditions for existence and uniqueness for almost every pair and bounds the error relative to the target.

Algorithm 2 uses explicit Euler integration to compute the output. Other ODE solvers can use the same trained functions. The generation guarantee concerns the continuous-time flow.

2.4 Grouped PairedFM

When group labels or known relationships are available, they can guide pair selection. Grouped PairedFM replaces $P \otimes P$ in (1) by a chosen joint observation law Π on $\mathcal{Y} \times \mathcal{Y}$, giving the population

Algorithm 1 PairedFM training

Input Observations $Y_1, \dots, Y_n$; initialized encoder h_θ and velocity v_ϕ; batch size b; iterations S; step sizes $\eta_1, \dots, \eta_S$.

for $s = 1, \dots, S$ **do**
  Draw i_r, j_r independently and uniformly from $\{1, \dots, n\}$, $r = 1, \dots, b$.
  Draw independent $t_r \sim \mathrm{Unif}[0,1]$, $r = 1, \dots, b$.
  $y_{r,t_r} \leftarrow (1 - t_r) Y_{i_r} + t_r Y_{j_r}$, $r = 1, \dots, b$.
  $\widehat{R} \leftarrow \frac{1}{b} \sum_{r=1}^{b} \| v_\phi(y_{r,t_r}, t_r \mid h_\theta(Y_{i_r}), h_\theta(Y_{j_r})) - (Y_{j_r} - Y_{i_r}) \|^2$.
  $(\theta, \phi) \leftarrow (\theta, \phi) - \eta_s \nabla_{(\theta,\phi)} \widehat{R}$.
end for
$(\widehat{\theta}, \widehat{\phi}) \leftarrow (\theta, \phi)$.
return $\widehat{h} = h_{\widehat{\theta}}$ and $\widehat{v} = v_{\widehat{\phi}}$.

Algorithm 2 PairedFM generation

Input Encoder $\widehat{h}$; velocity $\widehat{v}$; source y_0 and target y_1; $N \geq 1$ integration steps.

Compute $\widehat{h}(y_0)$ and $\widehat{h}(y_1)$ once; hold them fixed during integration.
$z_0 \leftarrow y_0$.
for $r = 0, \dots, N-1$ **do**
  $z_{r+1} \leftarrow z_r + \frac{1}{N} \widehat{v}(z_r, r/N \mid \widehat{h}(y_0), \widehat{h}(y_1))$.
end for
return z_N.

loss $R_\Pi(h, v) = \mathbb{E}_{(Y_0, Y_1) \sim \Pi} \int_0^1 \| v(Y_t, t \mid h(Y_0), h(Y_1)) - (Y_1 - Y_0) \|^2 \, dt$. For example, one can sample two observations within a group, or one observation from each of two specified groups. The selected comparisons and their sampling frequencies determine Π. Only the pair sampler in Algorithm 1 changes; the encoder, velocity field, displacement loss, and generation algorithm remain the same. Group labels are used for sampling and are not additional network inputs. The theory first studies independent pairing, then gives recovery and generation guarantees for a within-group model under additional assumptions.

3 Representation and Generation Guarantees

We first relate the PairedFM population loss to representation recovery and generation error, then study the stronger factor recovery available under grouped pairing. Proofs are in Appendices B and C.

3.1 Model and Regularity

Let $\mathcal{X} = \mathrm{supp}(X) \subset \mathbb{R}^d$ be compact, and consider

$$Y = f^\star(X), \qquad \mathcal{Y} = f^\star(\mathcal{X}) \subset \mathbb{R}^D, \qquad d < D, \tag{2}$$

where $f^\star : \mathcal{X} \to \mathcal{Y}$ is continuous and injective. Then $\mathcal{Y} = \mathrm{supp}(Y)$, and its inverse $h^\star : \mathcal{Y} \to \mathcal{X}$ satisfies $h^\star(Y) = X$. This inverse is defined only on the observation support: the model need not generate every point in $\mathbb{R}^D$. The latent law is otherwise unrestricted. Injective observation maps also appear in identifiable latent-variable models (Khemakhem et al., 2020).

We condition on the training data and randomness, treating the learned pair $(\widehat{h}, \widehat{v}) \in \mathcal{H} \times \mathcal{V}$ as fixed. The population loss (1) is evaluated on fresh independent observations. No encoder is assumed injective.

Assumption 1 (Observation-map regularity)**.** For constants $0 < \underline{\mathsf{L}}_f \leq \overline{\mathsf{L}}_f < \infty$,

$$\underline{\mathsf{L}}_f \, \|x - x'\| \leq \|f^\star(x) - f^\star(x')\| \leq \overline{\mathsf{L}}_f \, \|x - x'\| \qquad (x, x' \in \mathcal{X}).$$

Assumption 2 (Search-class regularity)**.** Every $h \in \mathcal{H}$ satisfies, for a common $\mathsf{L}_{\mathcal{H}} > 0$,

$$\|h(y) - h(y')\| \le \mathsf{L}_{\mathcal{H}} \|y - y'\| \qquad (y, y' \in \mathcal{Y}). \tag{3}$$

A common nonnegative measurable function $\mathsf{L}_{\mathcal{V}}$ with $\int_0^1 \mathsf{L}_{\mathcal{V}}(t)\, dt < \infty$ satisfies

$$\|v(z, t \mid c_0, c_1) - v(z', t \mid c_0, c_1)\| \le \mathsf{L}_{\mathcal{V}}(t) \|z - z'\| \tag{4}$$

for every $(h, v) \in \mathcal{H} \times \mathcal{V}$, almost every t, all $z, z' \in \mathbb{R}^D$, and all $c_0, c_1 \in h(\mathcal{Y})$.

Assumption 1 imposes quantitative distance bounds on the observation map $f^\star$. The lower bound controls the sensitivity of the inverse $h^\star$, allowing observation recovery guarantees to extend to latent variables. The upper bound limits the expansion of latent distances under $f^\star$. For observation maps represented by neural networks, related inverse Lipschitz bounds are studied by Puthawala et al. (2022).

Assumption 2 controls the regularity of the velocity and encoder. The spatial bound (4) limits the velocity's sensitivity to the interpolated observation when the source and target representations are fixed. Together with integrability of $\mathsf{L}_{\mathcal{V}}$, the spatial bound allows small population loss to imply that the representations retain information about the observations. The encoder bound (3) provides additional control over distances between learned representations and is not needed for the population-loss information recovery bound.

We impose velocity regularity on $\mathbb{R}^D$ because interpolations and generated paths may leave $\mathcal{Y}$. Spatial Lipschitz conditions are standard in the literature (Chen et al., 2018; Stéphanovitch, 2026). An integrable time-dependent bound permits sensitivity to vary along the flow while providing the stability estimates used in the recovery analysis (Teschl, 2012, Section 2.3).

3.2 Representation recovery

We measure recovery by the mean-square error of the conditional mean given the representation. Define

$$\beta[\mathsf{L}_{\mathcal{V}}] = \int_0^1 \big[(1 - t\mathsf{L}_{\mathcal{V}}(t))_+^2 + (1 - (1-t)\mathsf{L}_{\mathcal{V}}(t))_+^2\big] dt, \quad u_+ = \max\{u, 0\}.$$

Integrability of $\mathsf{L}_{\mathcal{V}}$ ensures $\beta[\mathsf{L}_{\mathcal{V}}] > 0$.

Theorem 1 (Representation recovery from the population loss)**.** *For any observation law with finite second moment, suppose $\widehat{v}$ satisfies (4) for all $c_0, c_1 \in \widehat{h}(\mathrm{supp}(Y))$, with integrable $\mathsf{L}_{\mathcal{V}}$, and $R(\widehat{h}, \widehat{v}) < \infty$. Then*

$$\mathbb{E}\left\| Y - \mathbb{E}[Y \mid \widehat{h}(Y)] \right\|^2 \le \frac{R(\widehat{h}, \widehat{v})}{\beta[\mathsf{L}_{\mathcal{V}}]}. \tag{5}$$

Under (2) and the lower bound in Assumption 1,

$$\mathbb{E}\left\| X - \mathbb{E}[X \mid \widehat{h}(Y)] \right\|^2 \le \frac{R(\widehat{h}, \widehat{v})}{\underline{\mathsf{L}}_f^2 \beta[\mathsf{L}_{\mathcal{V}}]}.$$

Exact recovery at zero loss. At zero population loss, (5) gives $\mathbb{E}[Y \mid \widehat{h}(Y)] = Y$ almost surely, so the learned representation determines the observation. For $Y = f^\star(X)$ in (2), zero loss also gives $\mathbb{E}[X \mid \widehat{h}(Y)] = X$ almost surely. Exact latent recovery does not require the lower distance bound in Assumption 1 or endpoint continuity of the velocity (Appendix B.1).

Theorem 1 connects the PairedFM population loss to representation quality. For a fixed integrable $\mathsf{L}_{\mathcal{V}}$, small population loss implies that Y can be predicted from $\widehat{h}(Y)$ with small mean-square error. When $Y = f^\star(X)$ and $f^\star$ satisfies the lower distance bound in Assumption 1, the latent variables X can also be predicted accurately, with the error controlled by $\underline{\mathsf{L}}_f$.

If the regularity bounds vary, vanishing observation and latent recovery errors follow from $R/\beta[\mathsf{L}_{\mathcal{V}}] \to 0$ and $R/(\underline{\mathsf{L}}_f^2 \beta[\mathsf{L}_{\mathcal{V}}]) \to 0$, respectively. Appendix B.2 bounds the probability that separated observations receive nearby representations.

3.3 IDENTIFICATION WITHIN THE SEARCH CLASS

Theorem 1 connects PairedFM's velocity-prediction objective to recovery of Y from the learned representation $\widehat{h}(Y)$. When observations are compared through their representations, we also need to understand how representation distances relate to differences between latent variables. For $Y = f^\star(X)$ and $Y' = f^\star(X')$, we therefore study how the population loss controls the relation between $\left\|\widehat{h}(Y) - \widehat{h}(Y')\right\|$ and $\|X - X'\|$.

Theorem 2 (Approximate identification)**.** *Suppose Assumptions 1 and 2 hold and $R(\widehat{h}, \widehat{v}) < \infty$. Suppose also that a common $\mathsf{L}_{\mathcal{V},h} > 0$ satisfies*

$$\|v(z, t \mid c_0, c_1) - v(z, t \mid c_0, c_1')\| \le \mathsf{L}_{\mathcal{V},h} \|c_1 - c_1'\| \tag{6}$$

for every $(h, v) \in \mathcal{H} \times \mathcal{V}$, almost every $t \in (0,1)$, all $z \in \mathbb{R}^D$, and all $c_0, c_1, c_1' \in h(\mathcal{Y})$. For $a \in (0,1)$, define $I_a = \{t \in (0,1) : t\mathsf{L}_{\mathcal{V}}(t) \le 1 - a\}$, and let $|I_a|$ denote its length. Integrability of $\mathsf{L}_{\mathcal{V}}$ ensures that $|I_a| > 0$. For every $a \in (0,1)$, independent copies X, X', and $\eta \in (0,1)$, with probability at least $1 - \eta$,

$$\frac{a\underline{\mathsf{L}}_f}{\mathsf{L}_{\mathcal{V},h}} \|X - X'\| - \frac{1}{\mathsf{L}_{\mathcal{V},h}} \sqrt{\frac{2R(\widehat{h}, \widehat{v})}{|I_a|\eta}} \le \left\|\widehat{h}(f^\star(X)) - \widehat{h}(f^\star(X'))\right\| \le \mathsf{L}_{\mathcal{H}}\overline{\mathsf{L}}_f \|X - X'\|. \tag{7}$$

The upper bound in (7) ensures that nearby latent values X, X' receive nearby representations $\widehat{h}(Y), \widehat{h}(Y')$. The lower bound ensures that the representations remain separated whenever the latent distance is large enough to make that bound positive. The upper bound holds for every pair; the lower bound holds with probability at least $1 - \eta$ over independent draws X, X'.

For fixed regularity bounds, a, and η, the additive error decreases as $\sqrt{R(\widehat{h}, \widehat{v})}$. Learning a pair with smaller population loss therefore allows representation distances to distinguish progressively smaller latent differences. The choice of a balances the two terms in the lower bound: increasing a strengthens the coefficient of $\|X - X'\|$, but reduces $|I_a|$, which can increase the additive error. Appendix B.2 provides the proof.

When $R(\widehat{h}, \widehat{v}) = 0$, the learned representations separate every pair of distinct latent values in $\mathcal{X}$. The resulting distance bounds also give an invertible correspondence between the representation sets of any two zero-loss encoders.

Corollary 1 (Exact identification)**.** *Under the assumptions of Theorem 2, suppose $R(\widehat{h}, \widehat{v}) = 0$. Then $x \mapsto \widehat{h}(f^\star(x))$ is a bijection from $\mathcal{X}$ onto $\widehat{h}(\mathcal{Y})$, and*

$$\frac{\underline{\mathsf{L}}_f}{\mathsf{L}_{\mathcal{V},h}} \|x - x'\| \le \left\|\widehat{h}(f^\star(x)) - \widehat{h}(f^\star(x'))\right\| \le \mathsf{L}_{\mathcal{H}}\overline{\mathsf{L}}_f \|x - x'\| \qquad (x, x' \in \mathcal{X}). \tag{8}$$

For any other $(h, v) \in \mathcal{H} \times \mathcal{V}$ with $R(h, v) = 0$, the correspondence $\widehat{h}(y) \longmapsto h(y), y \in \mathcal{Y}$, defines a bijection from $\widehat{h}(\mathcal{Y})$ onto $h(\mathcal{Y})$. Its inverse sends $h(y)$ to $\widehat{h}(y)$. Both maps are $\mathsf{L}_{\mathcal{H}}\mathsf{L}_{\mathcal{V},h}$-Lipschitz.

Corollary 1 shows that different zero-loss encoders retain the same information about Y. Their representations for the same observation determine one another, and their representation distances for the same pair of observations differ by at most the stated multiplicative factors. Identification therefore allows different numerical coordinates while controlling the information and distances they represent.

Together with Theorem 1, these results establish recovery and representation-distance guarantees through PairedFM's velocity-prediction objective and the stated regularity assumptions, without adding a reconstruction loss. Related work studies approximate agreement between learned representations up to rigid transformations (Nelson et al., 2026).

3.4 RECOVERY DURING GENERATION

Theorem 1 controls prediction of Y from its learned representation $\widehat{h}(Y)$. We now study whether the learned velocity produces an accurate output using the source and target representations. For

independent observations Y_0, Y_1, let $Y_t = (1-t)Y_0 + tY_1$. Starting from Y_0, we integrate the learned velocity with both source and target representations fixed and compare the generated output with Y_1:

$$\frac{dZ_t}{dt} = \widehat{v}(Z_t, t \mid \widehat{h}(Y_0), \widehat{h}(Y_1)), \qquad Z_0 = Y_0. \tag{9}$$

Theorem 3 (Generation error)**.** *Suppose $\widehat{v}$ is measurable, the spatial bound (4) in Assumption 2 holds, and $R(\widehat{h}, \widehat{v}) < \infty$. Define*

$$\kappa[\mathsf{L}_\mathcal{V}] = \int_0^1 \exp\left(2\int_s^1 \mathsf{L}_\mathcal{V}(r)\,dr\right) ds \le \exp\left(2\int_0^1 \mathsf{L}_\mathcal{V}(r)\,dr\right) < \infty.$$

Equation (9) has a unique absolutely continuous solution for almost every sampled pair, and its output Z_1 satisfies

$$\mathbb{E}\,\|Z_1 - Y_1\|^2 \le \kappa[\mathsf{L}_\mathcal{V}] R(\widehat{h}, \widehat{v}). \tag{10}$$

At zero loss, $Z_t = Y_t$ for every $t \in [0,1]$, almost surely.

The same conclusions hold for any observation pair law Π with finite second moments, replacing R by the population loss R_Π evaluated under Π, provided $R_\Pi(\widehat{h}, \widehat{v}) < \infty$ and the spatial bound (4) holds on its representation pairs.

Theorem 3 bounds the error of the generated output Z_1 relative to the target observation Y_1 whose representation conditions the flow. For fixed $\mathsf{L}_\mathcal{V}$, this error vanishes with the population loss. The PairedFM velocity-prediction objective therefore controls both recovery from the learned representation and recovery using the learned velocity. The factor $\kappa[\mathsf{L}_\mathcal{V}]$ accounts for the amplification of velocity errors along the generated trajectory. The bound concerns the continuous-time solution; numerical integration introduces an additional error.

3.5 Grouped PairedFM: recovery and generation

The preceding results concern recovery of the full latent variable X and generation of the target observation Y_1. For grouped pairing, write $X = (U, G)$, where G is shared within a group and U varies between its observations. We seek a representation that recovers U consistently across groups, so that generation can combine the target's U with the source's G.

We consider $U \perp G$, with $\mathcal{U} = \mathrm{supp}(U)$, $\mathcal{G} = \mathrm{supp}(G)$, and compact $\mathcal{X} = \mathcal{U} \times \mathcal{G}$. Let $h^\star_U$ denote the corresponding component of $h^\star$, so that $h^\star_U(Y) = U$. The paired observations are

$$Y_0 = f^\star(U_0, G), \quad Y_1 = f^\star(U_1, G), \quad U_0, U_1 \overset{\text{iid}}{\sim} P_U, \quad (U_0, U_1) \perp G. \tag{11}$$

Only the pairing law changes; R_Π denotes the population loss under (11).

Theorem 4 (Recovery and generation under grouped pairing)**.** *Under (11), suppose (4) holds on its representations, $R_\Pi(\widehat{h}, \widehat{v}) < \infty$, and $\widehat{h}(Y_0) \perp \widehat{h}(Y_1)$. For almost every g, suppose $u \mapsto \widehat{h}(f^\star(u, g))$ belongs to a class $\mathcal{A}$ of measurable maps $\mathcal{U} \to \mathbb{R}^k$ satisfying*

$$a, b \in \mathcal{A}, \quad a(U) \overset{d}{=} b(U) \quad \Longrightarrow \quad a(U) = b(U) \text{ almost surely.} \tag{12}$$

Then

$$\widehat{h}(f^\star(u, g)) = \widehat{h}(f^\star(u, g')) \quad \text{for } P_U \otimes P_G \otimes P_G\text{-almost every } (u, g, g'). \tag{13}$$

For generation, draw a source observation $Y_a = f^\star(U_a, G_a)$ and a target observation $Y_b = f^\star(U_b, G_b)$ independently. Their groups may differ. Let Z_t solve (9) from $Z_0 = Y_a$, conditioned on the source and target representations $\widehat{h}(Y_a), \widehat{h}(Y_b)$. Then

$$\mathbb{E}\,\|Z_1 - f^\star(U_b, G_a)\|^2 \le \kappa[\mathsf{L}_\mathcal{V}] R_\Pi(\widehat{h}, \widehat{v}). \tag{14}$$

At zero loss, $\mathbb{E}[U \mid \widehat{h}(Y)] = h^\star_U(Y) = U$ almost surely; for almost every g, $u \mapsto \widehat{h}(f^\star(u, g))$ is injective on a set of full P_U-measure; and $Z_1 = f^\star(U_b, G_a)$ almost surely.

Theorem 4 shows that the learned representation need not retain every latent variable appearing in the generated observation. At $R_\Pi(\widehat{h}, \widehat{v}) = 0$, the target representation determines U_b, while generation starts from the full source observation Y_a and recovers $f^\star(U_b, G_a)$. Thus omitting G from the representation is compatible with reproducing the source's G_a in the output. For fixed $\mathsf{L}_\mathcal{V}$, the bound (14) shows that the mean-square error relative to this output vanishes with the population loss. The proof is given in Appendix C.

4 EXPERIMENTS

Branching fractals. We compare PairedFM with FM (Lipman et al., 2023), β-VAE (Higgins et al., 2017), and LatentFM (Dao et al., 2023) on the distribution in Figure 1. PairedFM receives held-out sources and target codes; baselines sample from noise or priors. PairedFM leads all four generation metrics, with branch-mass TV about one-third of the lowest baseline value (Table 1). It also leads all four observation-recovery and distance metrics under common frozen-code evaluation (Table 2). Both tables report means $\pm$ sample standard deviations; details are in Appendix A.2.

Method	$SW_1 \downarrow$	RBF $MMD^2 \downarrow$	Branch-mass TV$\downarrow$	Outside 99% density$\downarrow$
PairedFM	**0.0081 $\pm$ 0.0017**	**-0.000057 $\pm$ 0.000111**	**0.0438 $\pm$ 0.0030**	**0.0188 $\pm$ 0.0048**
FM	0.0248 $\pm$ 0.0050	0.000530 $\pm$ 0.000371	0.1423 $\pm$ 0.0034	0.5986 $\pm$ 0.0270
β-VAE	0.0571 $\pm$ 0.0022	0.004707 $\pm$ 0.000306	0.1547 $\pm$ 0.0028	0.8044 $\pm$ 0.0031
LatentFM	0.0249 $\pm$ 0.0147	0.001545 $\pm$ 0.002396	0.1328 $\pm$ 0.0631	0.5196 $\pm$ 0.0485

Table 1: Generation fidelity on the branching fractal distribution.

Code	8-NN Y MSE$\downarrow$	8-NN Y MAE$\downarrow$	8-NN Y $R^2 \uparrow$	Distance $\rho \uparrow$
PairedFM	**0.0000275 $\pm$ 0.0000025**	**0.00312 $\pm$ 0.00010**	**0.999754 $\pm$ 0.000024**	**0.9957 $\pm$ 0.0010**
β-VAE	0.0034028 $\pm$ 0.0029023	0.01699 $\pm$ 0.00925	0.969781 $\pm$ 0.025650	0.7676 $\pm$ 0.0902
LatentFM	0.0007542 $\pm$ 0.0012285	0.00872 $\pm$ 0.00862	0.993316 $\pm$ 0.010881	0.6453 $\pm$ 0.1659

Table 2: Observation recovery and distance agreement on fractals.

IFN-β stimulation in PBMCs. We evaluate stimulation-response transfer in peripheral blood mononuclear cells (PBMCs) (Kang et al., 2018), using 24,673 cells and 5,000 selected genes. For each of seven target cell types, its stimulated cells are excluded from training. PairedFM has the highest mean response R^2_Δ for every target, averaging 0.802, and the lowest response MAE and cell-distribution MMD^2 (Table 3). GeneSpaceShift has higher response-sign agreement. The LatentFM comparison uses an autoencoder-code edit; its flow prior is unused.

Common frozen-code probes assess condition and cell-type prediction and linear recovery of observed expression (Table 4). PairedFM has the highest mean classification accuracies with a 16-dimensional code; the comparators use 32 dimensions. LatentFM's autoencoder has slightly lower decoder MSE (0.0914 versus 0.0941). Thus PairedFM supports both response transfer and informative representations without a reconstruction objective. Both tables equally average seven targets. The within-cohort split, preprocessing, target-wise results, distribution diagnostics, and figures are detailed in Appendix A.3.

Method	$R^2_\Delta \uparrow$	MAE$\downarrow$	$MMD^2 \downarrow$	Sign$\uparrow$
PairedFM	**0.802**	**0.0226**	**0.0199**	0.748
β-VAE	0.696	0.0262	0.0714	0.728
LatentFM	0.750	0.0304	0.0567	0.688
GeneSpaceShift	0.594	0.0255	0.0378	**0.776**

Table 3: PBMC response prediction.

	Condition acc.$\uparrow$		Cell type$\uparrow$	MSE$\downarrow$
Encoder	Linear	MLP	Linear	Decoder
PairedFM	**0.9873**	**0.9819**	**0.9506**	0.0941
β-VAE	0.9852	0.9767	0.9333	0.0992
LatentFM	0.9865	0.9801	0.9464	**0.0914**

Table 4: Frozen PBMC code probes.

Additional numerical results, including Grouped PairedFM on Shapes3D (Burgess & Kim, 2018), are provided in Appendix A.

5 DISCUSSION

We introduced PairedFM, which jointly learns representations and observation-space flows through velocity prediction, without a reconstruction objective. Under regularity assumptions, we showed that its population loss controls representation recovery and generation error, with additional factor-transfer guarantees for grouped pairing. We leave optimization and generalization bounds connecting finite-sample neural-network training to these population guarantees for future work.

AI USE STATEMENT

During the preparation of this manuscript, generative AI tools were utilized solely to assist with structural organization, language editing, and LaTeX typesetting. The authors reviewed and edited all content thoroughly and take full responsibility for the integrity, validity, and conclusions of the work.

REPRODUCIBILITY STATEMENT

All theoretical statements are proved in Appendices B and C. Algorithms 1 and 2 in Section 2 specify training and generation. Implementation and evaluation details are provided in Appendix A. The PBMC and Shapes3D datasets are publicly available and cited in Section 4 and Appendix A.

REFERENCES

Chris Burgess and Hyunjik Kim. 3D Shapes Dataset. https://github.com/google-deepmind/3d-shapes, 2018.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html.

Hugh Dance and Benjamin Bloem-Reddy. Counterfactual cocycles: A framework for robust and coherent counterfactual transports. *arXiv preprint arXiv:2405.13844v4*, 2025. URL https://arxiv.org/abs/2405.13844v4.

Quan Dao, Hao Phung, Binh Nguyen, and Anh Tran. Flow matching in latent space. *arXiv preprint arXiv:2307.08698*, 2023. URL https://arxiv.org/abs/2307.08698.

Luigi Gresele, Paul K. Rubenstein, Arash Mehrjou, Francesco Locatello, and Bernhard Schölkopf. The incomplete Rosetta Stone problem: Identifiability results for multi-view nonlinear ICA. In *Proceedings of the 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 217–227. PMLR, 2020. URL https://proceedings.mlr.press/v115/gresele20a.html.

Pengsheng Guo and Alex Schwing. Variational rectified flow matching. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 20921–20940. PMLR, 2025. URL https://proceedings.mlr.press/v267/guo25i.html.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Sy2fzU9gl.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999.

Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M. Lanata, Rachel E. Gate, Sara Mostafavi, Alexander Marson, Noah Zaitlen, Lindsey A. Criswell, and Chun Jimmie Ye. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology*, 36(1):89–94, 2018. doi: 10.1038/nbt.4042.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 2020. URL https://proceedings.mlr.press/v108/khemakhem20a.html.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014. arXiv:1312.6114.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

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 *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

Francesco Locatello, Ben Poole, Gunnar Raetsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 6348–6359. PMLR, 2020. URL https://proceedings.mlr.press/v119/locatello20a.html.

Sarthak Mittal, Korbinian Abstreiter, Stefan Bauer, Bernhard Schölkopf, and Arash Mehrjou. Diffusion based representation learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24963–24982. PMLR, 2023. URL https://proceedings.mlr.press/v202/mittal23a.html.

Walter Nelson, Marco Fumero, Theofanis Karaletsos, and Francesco Locatello. Statistical and structural identifiability in representation learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Wa3cfE3Iay.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T. Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28100–28127. PMLR, 2023. URL https://proceedings.mlr.press/v202/pooladian23a.html.

Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10619–10629, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Preechakul_Diffusion_Autoencoders_Toward_a_Meaningful_and_Decodable_Representation_CVPR_2022_paper.html.

Michael Puthawala, Konik Kothari, Matti Lassas, Ivan Dokmanić, and Maarten de Hoop. Globally injective ReLU networks. *Journal of Machine Learning Research*, 23(105):1–55, 2022. URL https://jmlr.org/papers/v23/21-0282.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, 2022.

Sagar Shrestha and Xiao Fu. Diversified flow matching with translation identifiability. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 55360–55378. PMLR, 2025. URL https://proceedings.mlr.press/v267/shrestha25a.html.

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://arxiv.org/abs/1910.09772.

Arthur Stéphanovitch. Lipschitz regularity in flow matching and diffusion models: sharp sampling rates and functional inequalities, 2026. URL https://arxiv.org/abs/2604.06065.

Xavier Sumba, Carles Balsells-Rodas, and Yingzhen Li. Structured coupling for flow matching. *arXiv preprint arXiv:2605.07676*, 2026. URL https://arxiv.org/abs/2605.07676.

Gerald Teschl. *Ordinary Differential Equations and Dynamical Systems*, volume 140 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012. ISBN 978-0-8218-8328-0. URL https://www.mat.univie.ac.at/~gerald/ftp/book-ode/.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2302.00482.

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. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16451–16467, 2021. URL https://arxiv.org/abs/2106.04619.

Liming Wang, Muhammad Jehanzeb Mirza, Yishu Gong, Yuan Gong, Jiaqi Zhang, Brian Tracey, Katerina Placek, Marco Vilela, and Jim Glass. Can diffusion models disentangle? a theoretical perspective. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/00bd71667563acfea40499ad76229894-Abstract-Conference.html.

Yidan Xu, Yixin Wang, and Long Nguyen. Structured flow autoencoders: Learning structured probabilistic representations with flow matching. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/f6fbaf131fa71b036d7b8c49379f3af4-Paper-Conference.pdf.

Supplementary Materials for "Paired Flow Matching: Representation Learning without a Reconstruction Bottleneck"

A Additional numerical results

A.1 Grouped PairedFM on Shapes3D

Task and implementation. Shapes3D (Burgess & Kim, 2018) contains 480,000 64×64 RGB images covering six factors. For hue transfer, U is object hue (ten values) and G comprises floor hue, wall hue, scale, shape, and orientation. Grouped pairs share G and draw their two hues independently. We reserve 5% of scenes from Grouped training; the baseline encoders were trained on the complete grid. Factor labels select pairs but are not model inputs. At generation, the source image initializes the flow, the target image supplies the desired hue through its code, and the factor grid gives the exact source-scene/target-hue image.

The Grouped model uses a ten-dimensional code, 6,000 updates, batch size 256, Adam learning rate 10^{-4}, and EMA decay 0.999. We integrate its image-space flow with 50 steps for generation. The ten-dimensional β-VAE ($\beta = 4$) and LatentFM encoders each have 15,000 autoencoder updates. LatentFM has a further 8,000 latent-flow updates. The LatentFM encoder supplies the code comparisons below.

Recovery and scene information. Grouped PairedFM recovers object hue much more reliably than either comparator (Table 5). NMI/ARI use 3,000 images per run; the probes fit on 1,600 images and test on 1,600 from disjoint scenes. Floor and wall hue remain decodable above chance from the Grouped code, so the evidence supports reduced scene information, not exact invariance. Figure 2 colors the same evaluation codes by changing and shared factors.

Encoder	Hue NMI↑	Hue ARI↑	Linear hue↑	1-NN hue↑	Floor linear↓	Wall linear↓
Grouped PairedFM	**.9971**	**.9975**	**1.000**	**1.000**	**.301**	**.334**
β-VAE	.1404	.0592	.889	.537	.975	.968
LatentFM	.1772	.0629	.880	.531	.948	.994

Table 5: Object-hue representation diagnostics. All metrics are means over all runs.

Common coordinates across scenes. Grouped PairedFM's cross-scene code distance at fixed hue is smaller than a one-step hue distance; both comparator ratios exceed one (Table 6). Its paired-code dCor is close to the shuffled-target value. These finite-sample diagnostics examine common hue coordinates across scenes.

Encoder	Cross-scene/one-step↓	Hue-step ρ ↑	Paired dCor↓	Shuffled dCor mean
Grouped PairedFM	**.249**	**.867**	**.165**	.183
β-VAE	5.296	.596	.885	.422
LatentFM	8.685	.629	.976	.259

Table 6: One seed native-code geometry. The ratio is median same-hue, different-scene distance (600 pairs) divided by median one-hue-step, same-scene distance (120 pairs). Spearman ρ uses 600 within-scene hue-difference pairs. Distance correlation uses 128 within-scene source–target code pairs; the last column is the mean over 100 target-code permutations.

Conditional generation. Across three Grouped PairedFM runs, transfer SSIM is 0.9710 ± 0.0037 and pixel MSE is 35.84 ± 10.15 (mean $\pm$ sample SD). Each run uses 128 held-out source–target transfers, including unchanged-hue cases. Figure 3 shows hue-transfer examples.

The separate object-intervention variant changes hue, shape, and scale while retaining the source scene. Its 46,000-update checkpoint attains SSIM 0.8815 and pixel MSE 465.5 on 96 cases; some outputs retain an incorrect shape or scale (Figure 4).

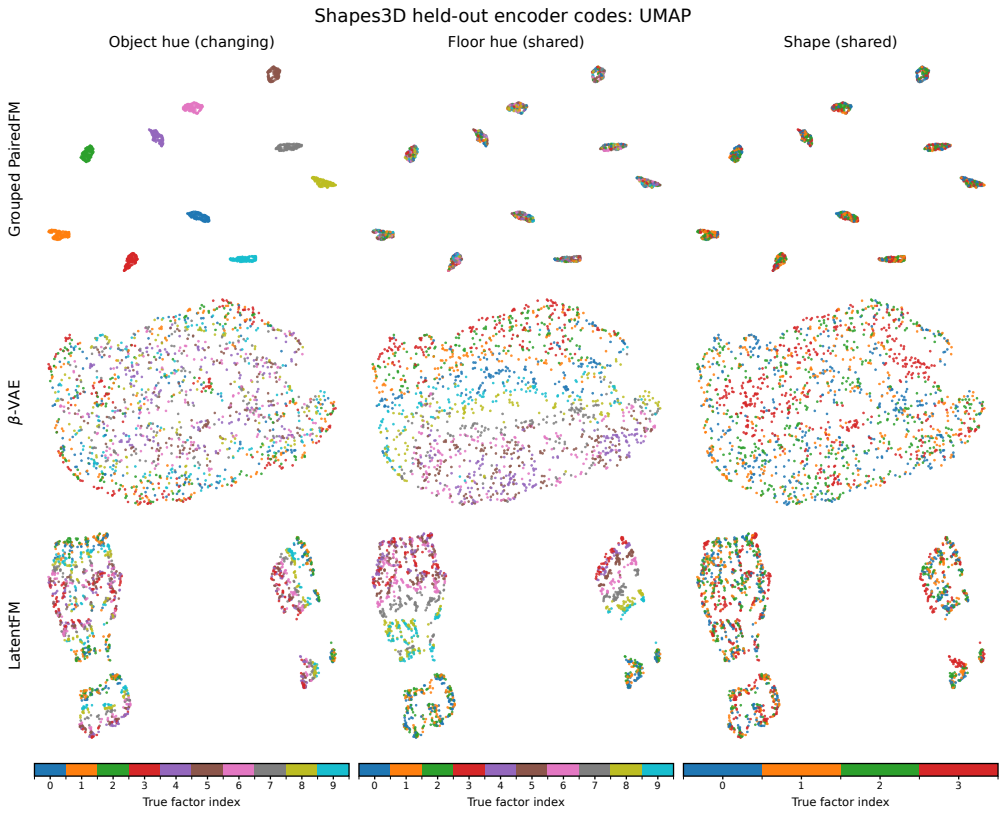


Figure 2: UMAP of the same 1,600 Shapes3D evaluation images for Grouped PairedFM, β-VAE, and LatentFM. Within each method row the coordinates stay fixed while point colors show object hue, floor hue, or shape. Each encoder has a separately fitted projection.

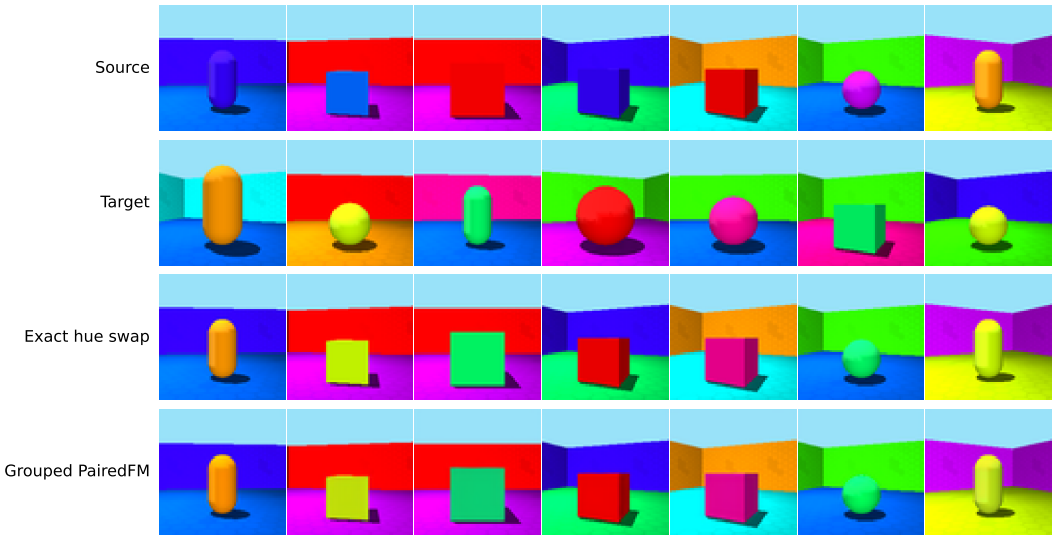


Figure 3: Grouped PairedFM object-hue transfer on seven Shapes3D examples. Each column shows the source image, a target image specifying the desired hue, the exact source-scene/target-hue image, and the generated image. The cases are the first seven hue changes in the saved evaluation order.

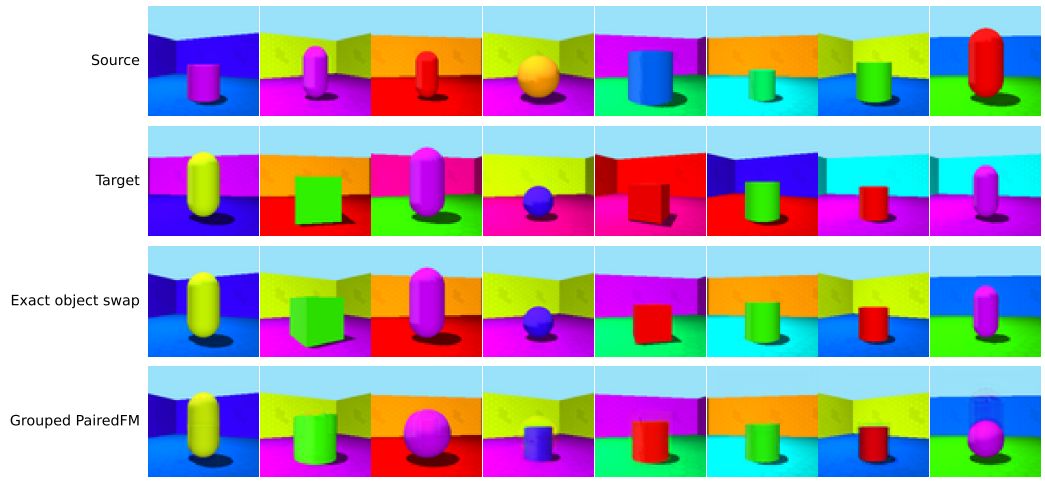


Figure 4: Object intervention from the separate Grouped PairedFM checkpoint. The first eight saved cases show source, target image, exact object-factor transfer, and generated image. The target image specifies object hue, shape, and scale.

A.2 Fractal evaluation

Protocol. Tables 1 and 2 report means and sample standard deviations. Generation uses 4,096 samples per method. PairedFM receives held-out observed sources and target codes; FM, β-VAE, and LatentFM sample from noise or priors. SW_1 is sliced Wasserstein-1 distance, and branch-mass TV compares branch probability masses. Branch and density scores use the synthetic geometry only for evaluation. The RBF bandwidth is fixed from the reference within each seed; unbiased MMD^2 can be negative.

The common 8-NN decoder fits on 3,072 held-out observation/code pairs and tests on another 1,024, with identical splits across encoders. Its errors average the two observed coordinates. Distance ρ is the Spearman correlation between observation and code distances over 50,000 observation pairs per seed. These diagnostics concern observed Y, not the unobserved latent X.

Fine-scale coverage. Table 7 tests whether samples cover the thin branches while remaining close to observed points. PairedFM's coverage and precision are near those of an independent draw from the reference distribution under this finite-sample test.

Method	Coverage↑	Precision↑
PairedFM	$\mathbf{0.808 \pm 0.016}$	$\mathbf{0.805 \pm 0.013}$
FM	0.489 ± 0.011	0.321 ± 0.034
β-VAE	0.331 ± 0.010	0.138 ± 0.007
LatentFM	0.515 ± 0.060	0.399 ± 0.041
Independent reference draw	0.813 ± 0.009	0.816 ± 0.015

Table 7: Nearest-neighbor coverage and precision (mean $\pm$ sample SD). The radius is twice the median distance from an independent reference draw to its nearest observed reference point. Coverage is the fraction of observed reference points near a generated point; precision reverses these roles. The final row is a finite-sample reference.

Held-out transport and code geometry. On 4,096 fresh independent source–target pairs per seed, PairedFM's mean squared velocity residual is $(3.20 \pm 0.48) \times 10^{-5}$ and its endpoint MSE is $(2.21 \pm 0.20) \times 10^{-5}$, measured as squared Euclidean norms (mean $\pm$ sample SD). With 64 integration steps, endpoint error divided by the target's squared Euclidean variance is $(9.74 \pm 0.93) \times 10^{-5}$; leaving the source unchanged gives 2.00 ± 0.03. These are held-out diagnostics of the learned transport.

Figure 5 displays corresponding held-out observations and codes. Figure 6 shows the spread of code distances at a given observed distance, beyond the single rank-correlation score in Table 2.

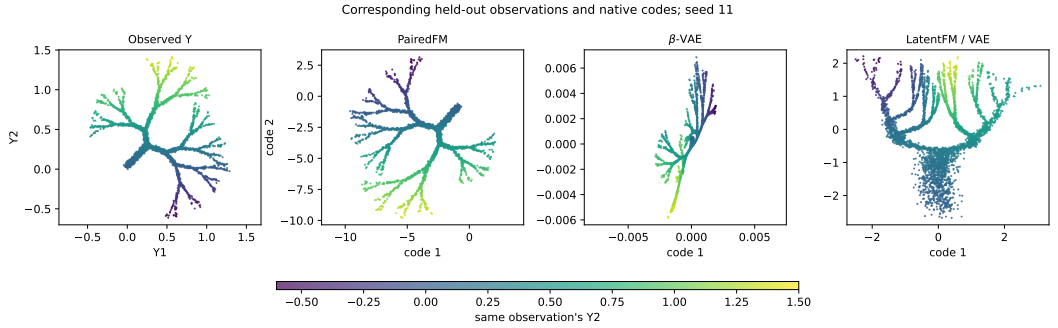


Figure 5: The same held-out fractal observations and their PairedFM, β-VAE, and LatentFM codes. Color denotes each observation's second coordinate in every panel. Code axes retain each encoder's native scale and are not aligned across panels.

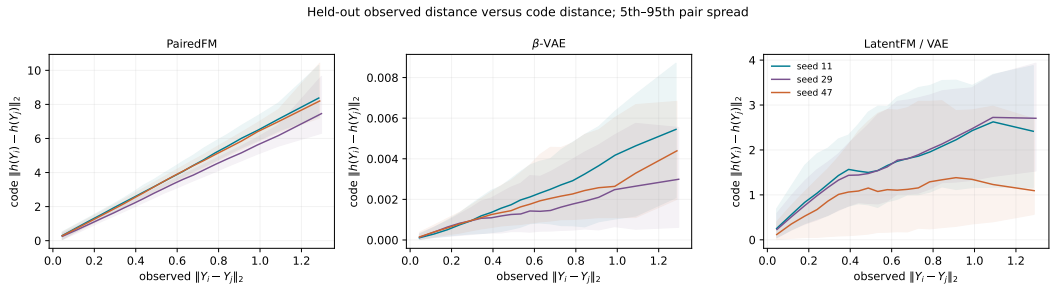


Figure 6: Held-out observed-Y distance versus native code distance. Lines show within-seed medians; bands show the fifth–95th percentile across observation pairs within each runs. Each encoder has its own code-distance scale.

A.3 IFN-β PBMC EVALUATION

Data and training. The source study profiled PBMCs from eight lupus patients after six hours of interferon-β stimulation or an untreated control condition (Kang et al., 2018) (GEO GSE96583). The processed matrix used here contains 24,673 cells and 5,000 selected highly variable genes. These are all genes evaluated in this run, rather than the complete transcriptome.

The seven targets are B cells, CD14$^+$ monocytes, CD4 and CD8 T cells, dendritic cells, FCGR3A$^+$ monocytes, and natural killer cells. Each target's stimulated cells are excluded from model training. Methods share source controls, non-target references, and target stimulated test cells within a seed. The 10% cell holdout is within patient and cell-type bins. Patients are shared between training and testing, and genes were selected on the full cohort before target exclusion; the experiment therefore measures within-cohort transfer.

PairedFM uses a 16-dimensional code and 10,000 updates. The β-VAE and LatentFM autoencoders use 32-dimensional codes and 10,000 updates. LatentFM also trains a latent flow for 6,000 updates, and its stimulation edit uses post hoc autoencoder-code arithmetic. GeneSpaceShift applies a stimulated-minus-control mean shift estimated from non-target cells directly in gene space.

Target responses. For selected gene g, let Δ_g be the observed target stimulated-minus-control mean and $\widehat{\Delta}_g$ its predicted shift. With $G = 5{,}000$ and $\overline{\Delta} = G^{-1}\sum_g \Delta_g$, we measure

$$R^2_\Delta = 1 - \frac{\sum_g (\widehat{\Delta}_g - \Delta_g)^2}{\sum_g (\Delta_g - \overline{\Delta})^2}, \qquad \mathrm{MAE}_\Delta = \frac{1}{G}\sum_g |\widehat{\Delta}_g - \Delta_g|.$$

Response-sign agreement uses genes with $|\Delta_g| \geq 0.01$. Unbiased RBF MMD^2 compares predicted and held-out stimulated cells in the full 5,000-gene space and can be negative in finite samples. Aggregate results average each target over three seeds and then give the seven targets equal weight. Table 8 retains the target-wise response scores.

Method	B	CD14	CD4	CD8	DC	FCGR3A	NK
PairedFM	0.864 ± 0.006	0.746 ± 0.006	0.712 ± 0.013	0.805 ± 0.038	0.807 ± 0.003	0.886 ± 0.014	0.793 ± 0.010
β-VAE	0.739 ± 0.053	0.632 ± 0.024	0.398 ± 0.122	0.795 ± 0.078	0.700 ± 0.069	0.835 ± 0.014	0.775 ± 0.008
LatentFM	0.824 ± 0.007	0.667 ± 0.014	0.611 ± 0.070	0.708 ± 0.024	0.776 ± 0.013	0.877 ± 0.002	0.789 ± 0.015
GeneSpaceShift	0.765 ± 0.000	0.483 ± 0.000	0.102 ± 0.000	0.594 ± 0.000	0.747 ± 0.000	0.767 ± 0.000	0.699 ± 0.000

Table 8: Response R^2_Δ by target cell type (mean $\pm$ sample SD).

PairedFM has the highest mean R^2_Δ for each target. The all-gene scatter in Figure 7 shows individual predicted responses and their departures from the identity line; the table alone summarizes these responses into one value per target.

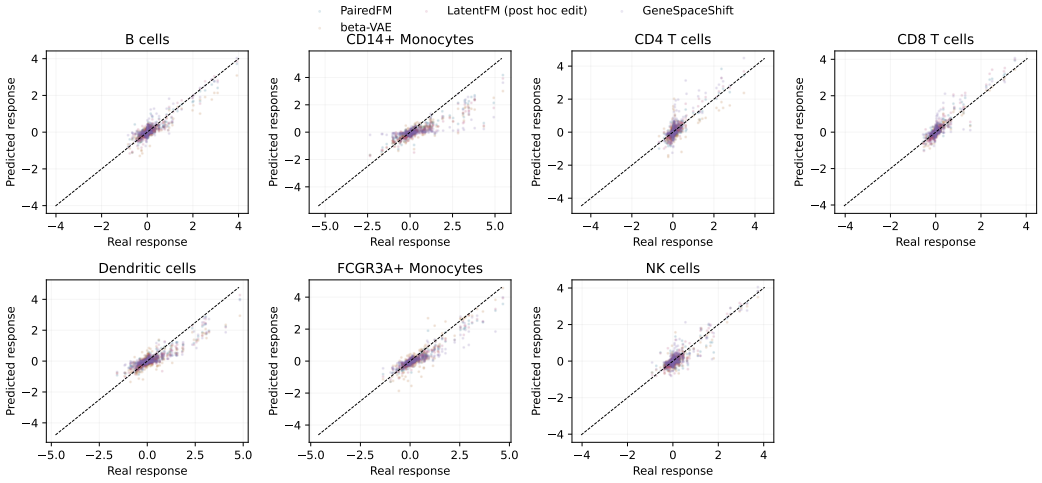


Figure 7: Predicted versus observed mean responses for all 5,000 selected genes in each target. The four methods share axes within each target panel; dashed lines indicate exact prediction. Tables summarize all seeds.

Codes and cell-to-cell variation. The frozen-code probes in Table 4 use common linear and MLP condition classifiers, a linear cell-type classifier, and a linear decoder of observed expression Y. They assess information available in each code, not recovery of an unobserved biological latent. The encoder dimensions differ between PairedFM and the comparators.

Table 9 asks a different question from the response table: whether generated cells reproduce variation within the stimulated population. Mean and variance r^2 are squared Pearson correlations across genes. The total-variance ratio compares predicted and observed stimulated-cell variance, with one as its reference value. Two-PC SW$_1$ uses a PCA basis fitted to real reference cells; MMD^2 uses the full selected-gene space.

Method	Mean $r^2 \uparrow$	Var. $r^2 \uparrow$	Var. MAE$\downarrow$	Total var. ratio$\rightarrow 1$	2-PC SW$_1 \downarrow$	MMD$^2 \downarrow$
PairedFM	0.952	0.749	0.0378	1.057	2.017	0.0199
β-VAE	0.927	0.326	0.0904	0.065	2.625	0.0714
LatentFM	0.942	0.498	0.0806	0.169	2.237	0.0567

Table 9: Target-cell distribution diagnostics. Each entry averages all the seed result within each target and then averages the seven targets equally.

PairedFM's total-variance ratio is 1.057, compared with 0.065 and 0.169 for the two autoencoder edits. Its full-space MMD^2 is also lower in the target average. Figure 8 displays CD4 T-cell predictions in one shared reference projection. That PCA basis includes held-out real stimulated cells for evaluation display only and is applied unchanged to every method; the full-space distribution comparison is supplied by MMD^2.

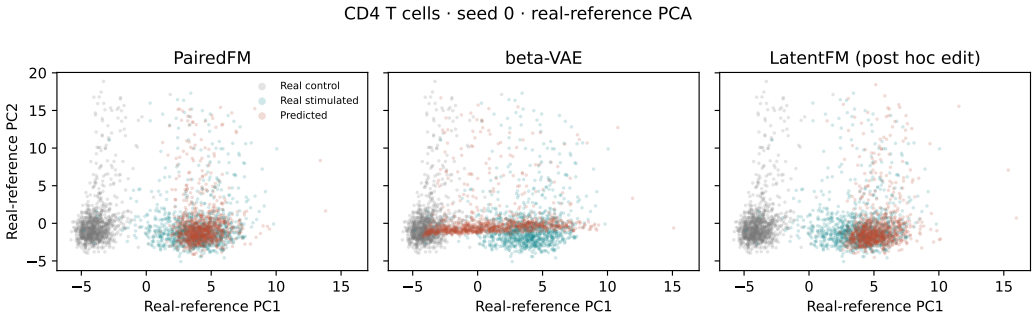


Figure 8: CD4 T-cell predictions and real controls and stimulated references on shared PCA axes. The basis is fitted to real reference cells and used only for this evaluation display.

B Proofs for PairedFM

B.1 Representation recovery

For any square-integrable vector W and square-integrable vector V measurable with respect to $\widehat{h}(Y)$,

$$\begin{aligned}\mathbb{E}\,\|W-V\|^2 &= \mathbb{E}\left\|W-\mathbb{E}[W\mid \widehat{h}(Y)]\right\|^2 \\ &\quad + \mathbb{E}\left\|\mathbb{E}[W\mid \widehat{h}(Y)]-V\right\|^2 .\end{aligned} \tag{B.1}$$

Expand the squared norm. The cross term vanishes because

$$\begin{aligned}&\mathbb{E}\langle W-\mathbb{E}[W\mid \widehat{h}(Y)],\mathbb{E}[W\mid \widehat{h}(Y)]-V\rangle \\ &\quad=\mathbb{E}\langle \mathbb{E}[W-\mathbb{E}[W\mid \widehat{h}(Y)]\mid \widehat{h}(Y)],\mathbb{E}[W\mid \widehat{h}(Y)]-V\rangle \\ &\quad=0.\end{aligned}$$

These products are integrable by Cauchy–Schwarz. Any prediction measurable with respect to $\widehat{h}(Y)$ with finite error is square-integrable because W is. Thus (B.1) shows that the conditional mean minimizes squared prediction error among all predictions determined by $\widehat{h}(Y)$.

If $\widetilde{W}$ is an independent draw from the conditional law of W given $\widehat{h}(Y)$, expansion and conditional independence give

$$\begin{aligned}\tfrac{1}{2}\mathbb{E}\left\|W-\widetilde{W}\right\|^2 &= \mathbb{E}\,\|W\|^2-\mathbb{E}\left\|\mathbb{E}[W\mid \widehat{h}(Y)]\right\|^2 \\ &= \mathbb{E}\left\|W-\mathbb{E}[W\mid \widehat{h}(Y)]\right\|^2 .\end{aligned} \tag{B.2}$$

These identities also hold with $\widehat{h}(Y)$ replaced by any conditioning variable, and under any fixed conditional law. We write $\mathrm{Var}(W)=\mathbb{E}\|W-\mathbb{E}W\|^2$ for scalar total variance of a vector, with conditioning indicated when used. Conditional laws and the required conditional copies exist for the Euclidean-valued variables here.

Proof of Theorem 1. If $R(\widehat{h},\widehat{v})=\infty$, the desired bounds are immediate. Otherwise the regression error is square-integrable for almost every time and almost every representation pair. Work at such a time and representation pair, conditional on $(\widehat{h}(Y_0),\widehat{h}(Y_1))=(c_0,c_1)$, with $\mathsf{L}_{\mathcal{V}}(t)<\infty$ and the spatial bound in force.

The endpoints are independent under this conditioning, with their respective laws given $\widehat{h}(Y_0) = c_0$ and $\widehat{h}(Y_1) = c_1$. To verify this factorization, for any bounded measurable functions a, b, independence of the original observations gives

$$\mathbb{E}[a(Y_0)b(Y_1) \mid \widehat{h}(Y_0), \widehat{h}(Y_1)] = \mathbb{E}[a(Y_0) \mid \widehat{h}(Y_0)]\,\mathbb{E}[b(Y_1) \mid \widehat{h}(Y_1)].$$

This characterizes the product conditional law.

Hold the source at a value y_0, and let $\widetilde{Y}_1$ be a conditionally independent copy of Y_1 under its target-representation law. The reverse triangle inequality and (4) imply

$$\begin{aligned}\Big\|\widehat{v}((1-t)y_0 + tY_1, t \mid c_0, c_1) - \widehat{v}((1-t)y_0 + t\widetilde{Y}_1, t \mid c_0, c_1) \\ - (Y_1 - \widetilde{Y}_1)\Big\| \\ \geq (1 - t\mathsf{L}_{\mathcal{V}}(t))_+ \left\|Y_1 - \widetilde{Y}_1\right\|,\end{aligned}$$

where $u_+ = \max(u, 0)$. Applying (B.2) to this difference and then averaging over Y_0 yields

$$\begin{aligned}\mathbb{E}[\mathrm{Var}\{\widehat{v}(Y_t, t \mid c_0, c_1) - Y_1 \mid Y_0, c_0, c_1\} \mid c_0, c_1] \\ \geq (1 - t\mathsf{L}_{\mathcal{V}}(t))_+^2\,\mathrm{Var}(Y_1 \mid c_1).\end{aligned} \tag{B.3}$$

For the variation over the source, compare two possible values y_0, y_0'. The conditional mean of the regression error at these sources differs by

$$(y_0 - y_0') + \mathbb{E}\left[\begin{matrix}\widehat{v}((1-t)y_0 + tY_1, t \mid c_0, c_1) \\ -\,\widehat{v}((1-t)y_0' + tY_1, t \mid c_0, c_1)\end{matrix}\,\middle|\, c_0, c_1\right].$$

The norm of the expectation term is at most $(1-t)\mathsf{L}_{\mathcal{V}}(t)\|y_0 - y_0'\|$. The whole difference consequently has norm at least $(1 - (1-t)\mathsf{L}_{\mathcal{V}}(t))_+\|y_0 - y_0'\|$. Apply (B.2) to independent draws from the source-representation law. It follows that

$$\begin{aligned}\mathrm{Var}(\mathbb{E}[\widehat{v}(Y_t, t \mid c_0, c_1) - (Y_1 - Y_0) \mid Y_0, c_0, c_1] \mid c_0, c_1) \\ \geq (1 - (1-t)\mathsf{L}_{\mathcal{V}}(t))_+^2\,\mathrm{Var}(Y_0 \mid c_0).\end{aligned} \tag{B.4}$$

Apply (B.1) to the regression error, first conditioning on the source and then on the representation pair. Its conditional second moment is the sum of the conditional variance over the target, the variance of its conditional mean over the source, and the squared norm of its overall conditional mean. The last term is nonnegative. Combining (B.3) and (B.4) therefore gives

$$\begin{aligned}&\mathbb{E}[\|\widehat{v}(Y_t, t \mid c_0, c_1) - (Y_1 - Y_0)\|^2 \mid c_0, c_1] \\ &\quad \geq (1 - t\mathsf{L}_{\mathcal{V}}(t))_+^2\,\mathrm{Var}(Y_1 \mid c_1) + (1 - (1-t)\mathsf{L}_{\mathcal{V}}(t))_+^2\,\mathrm{Var}(Y_0 \mid c_0).\end{aligned}$$

Average over the representation pair. Both endpoint conditional variances average to $\mathbb{E}\|Y - \mathbb{E}[Y \mid \widehat{h}(Y)]\|^2$. Integrating over time gives

$$R(\widehat{h}, \widehat{v}) \geq \beta[\mathsf{L}_{\mathcal{V}}]\mathbb{E}\left\|Y - \mathbb{E}[Y \mid \widehat{h}(Y)]\right\|^2, \tag{B.5}$$

where

$$\beta[\mathsf{L}_{\mathcal{V}}] = \int_0^1 \left[(1 - t\mathsf{L}_{\mathcal{V}}(t))_+^2 + (1 - (1-t)\mathsf{L}_{\mathcal{V}}(t))_+^2\right] dt. \tag{B.6}$$

This coefficient is at most two and strictly positive. Indeed, if it were zero, both nonnegative summands would vanish for almost every $t \in (0, 1)$. This would require

$$\mathsf{L}_{\mathcal{V}}(t) \geq \max\{1/t, 1/(1-t)\} \quad \text{for almost every } t \in (0, 1),$$

contradicting $\int_0^1 \mathsf{L}_{\mathcal{V}}(t)\,dt < \infty$. Dividing (B.5) by $\beta[\mathsf{L}_{\mathcal{V}}]$ proves (5). Finite loss and second moments justify the conditional calculations for almost every time and representation pair. All final integrands are nonnegative, so averaging and time integration also follow from Tonelli's theorem.

For the latent conclusion, take $\widetilde{X}$ as an independent draw from the conditional law of X given $\widehat{h}(Y)$, and put $\widetilde{Y} = f^\star(\widetilde{X})$. These are conditional copies, not the independent endpoints used in training. By (B.2) and the lower bound for $f^\star$,

$$\begin{aligned}\mathbb{E}\left\|X - \mathbb{E}[X \mid \widehat{h}(Y)]\right\|^2 &= \tfrac{1}{2}\mathbb{E}\left\|X - \widetilde{X}\right\|^2 \\ &\le \frac{1}{2\underline{\mathsf{L}}_f^2}\mathbb{E}\left\|Y - \widetilde{Y}\right\|^2 \\ &= \frac{1}{\underline{\mathsf{L}}_f^2}\mathbb{E}\left\|Y - \mathbb{E}[Y \mid \widehat{h}(Y)]\right\|^2 \\ &\le \frac{R(\widehat{h},\widehat{v})}{\underline{\mathsf{L}}_f^2\beta[\mathsf{L}_\mathcal{V}]}.\end{aligned}$$

□

Exact recovery at zero loss

Proof of the exact-recovery statement after Theorem 1. The observation bound (5) and $\beta[\mathsf{L}_\mathcal{V}] > 0$ give

$$\mathbb{E}\left\|Y - \mathbb{E}[Y \mid \widehat{h}(Y)]\right\|^2 = 0$$

when $R(\widehat{h},\widehat{v}) = 0$. The squared norm is nonnegative, so $Y = \mathbb{E}[Y \mid \widehat{h}(Y)]$ almost surely. Thus the observation is measurable with respect to its learned representation, up to a null event.

Under (2), the inverse $h^\star$ of $f^\star$ on $\mathcal{Y}$ is continuous. To verify this, suppose $f^\star(x_j) \to f^\star(x)$ but $x_j \not\to x$. A subsequence stays a positive distance from x. Compactness of $\mathcal{X}$ gives a further subsequence converging to some $x' \neq x$. Continuity then gives $f^\star(x') = f^\star(x)$, contrary to injectivity. This proves inverse continuity. Since $\mathcal{Y}$ is compact and closed, extending $h^\star$ outside $\mathcal{Y}$ by a fixed element of $\mathcal{X}$ gives a measurable function. Consequently $X = h^\star(Y)$ is measurable with respect to $\widehat{h}(Y)$, up to a null event. Compactness makes X integrable, and hence

$$\mathbb{E}[X \mid \widehat{h}(Y)] = h^\star(Y) = X \quad \text{almost surely.}$$

Choose a Borel version of $\mathbb{E}[Y \mid \widehat{h}(Y)]$ as a function of the representation. Such a version exists because the variables take values in Euclidean spaces. On a measurable set of full latent probability, its value at $\widehat{h}(f^\star(x))$ equals $f^\star(x)$. If two points in this set have the same representation, this function gives the same observation at both points. Injectivity of $f^\star$ makes the points equal. Thus $x \mapsto \widehat{h}(f^\star(x))$ is injective on this set. This argument uses neither a lower distance bound for $f^\star$ nor endpoint continuity of the velocity field. □

B.2 Representation separation at positive loss

The representation recovery bound measures information retained in the representation. The next statement limits how often separated observations receive nearby representations.

Proposition 1 (Approximate representation separation)**.** *Suppose the spatial bound (4) holds. In addition, suppose there is a finite $\mathsf{L}_{\mathcal{V},h} \ge 0$ such that every $(h, v) \in \mathcal{H} \times \mathcal{V}$ satisfies*

$$\|v(z, t \mid c_0, c_1) - v(z, t \mid c_0, c_1')\| \le \mathsf{L}_{\mathcal{V},h}\,\|c_1 - c_1'\|$$

for almost every t, all $z \in \mathbb{R}^D$, and $c_0, c_1, c_1' \in h(\mathcal{Y})$. The conclusion also holds if these spatial and representation bounds are assumed only for the learned pair. For $a \in (0, 1)$, let

$$I_a = \{t \in (0, 1) : t\mathsf{L}_\mathcal{V}(t) \le 1 - a\},$$

and write $|I_a|$ for its Lebesgue measure. This measure is strictly positive. For independent observations Y, Y',

$$\mathbb{E}\left[\left(a\,\|Y - Y'\| - \mathsf{L}_{\mathcal{V},h}\left\|\widehat{h}(Y) - \widehat{h}(Y')\right\|\right)_+^2\right] \le \frac{2R(\widehat{h},\widehat{v})}{|I_a|}. \tag{B.7}$$

Under (2) and the lower bound in Assumption 1, the observation distance on the left can be replaced by $\underline{\mathsf{L}}_f\|X - X'\|$ for independent latent draws. The representation bound applies also when a spatial state is evaluated with the other target's representation.

Proof. The set I_a is measurable because $\mathsf{L}_{\mathcal{V}}$ is. If $|I_a| = 0$, then $\mathsf{L}_{\mathcal{V}}(t) > (1-a)/t$ for almost every $t \in (0,1)$, which contradicts integrability of $\mathsf{L}_{\mathcal{V}}$. Thus $|I_a| > 0$.

If $R(\widehat{h}, \widehat{v}) = \infty$, the bound is immediate. Otherwise the regression errors are square-integrable for almost every time. Take three independent observations $Y_0, Y_1, \widetilde{Y}_1$ and form the pairs (Y_0, Y_1) and $(Y_0, \widetilde{Y}_1)$. Subtract the two regression errors. Changing first the spatial input and then the target representation, the Lipschitz bounds give

$$\begin{aligned}&\left\| \begin{aligned} &\widehat{v}((1-t)Y_0 + tY_1, t \mid \widehat{h}(Y_0), \widehat{h}(Y_1)) \\ &- \widehat{v}((1-t)Y_0 + t\widetilde{Y}_1, t \mid \widehat{h}(Y_0), \widehat{h}(\widetilde{Y}_1)) - (Y_1 - \widetilde{Y}_1) \end{aligned} \right\| \\ &\quad \geq \left((1 - t\mathsf{L}_{\mathcal{V}}(t)) \left\| Y_1 - \widetilde{Y}_1 \right\| - \mathsf{L}_{\mathcal{V},h} \left\| \widehat{h}(Y_1) - \widehat{h}(\widetilde{Y}_1) \right\| \right)_+ .\end{aligned}$$

For $t \in I_a$, the coefficient $1 - t\mathsf{L}_{\mathcal{V}}(t)$ is at least a. Replacing it by a gives a smaller lower bound that is independent of time.

Conditional on Y_0, the two regression errors are independent and identically distributed. By (B.2), the expectation of their squared difference is twice their mean conditional variance and is at most

$$2\mathbb{E} \left\| \widehat{v}(Y_t, t \mid \widehat{h}(Y_0), \widehat{h}(Y_1)) - (Y_1 - Y_0) \right\|^2 .$$

Square the preceding lower bound, average, and integrate over I_a. The lower bound contributes $|I_a|$ times the left side of (B.7). The upper bound contributes at most $2R(\widehat{h}, \widehat{v})$, since the squared regression error is nonnegative and $I_a \subset (0,1)$. Dividing by $|I_a|$ proves the result. The latent version follows from $\|Y - Y'\| \geq \underline{\mathsf{L}}_f \|X - X'\|$ and monotonicity of the positive part. □

Proof of Theorem 2. Proposition 1 establishes $|I_a| > 0$ and, using the lower bound in Assumption 1,

$$\mathbb{E}\left[\left(a\underline{\mathsf{L}}_f \|X - X'\| - \mathsf{L}_{\mathcal{V},h} \left\| \widehat{h}(f^\star(X)) - \widehat{h}(f^\star(X')) \right\| \right)_+^2 \right] \leq \frac{2R(\widehat{h}, \widehat{v})}{|I_a|}.$$

If $R(\widehat{h}, \widehat{v}) > 0$, Markov's inequality gives

$$\Pr\left\{ a\underline{\mathsf{L}}_f \|X - X'\| - \mathsf{L}_{\mathcal{V},h} \left\| \widehat{h}(f^\star(X)) - \widehat{h}(f^\star(X')) \right\| > \sqrt{\frac{2R(\widehat{h}, \widehat{v})}{|I_a|\eta}} \right\} \leq \eta.$$

On the complementary event, rearranging and dividing by $\mathsf{L}_{\mathcal{V},h} > 0$ gives the lower bound in (7). If $R(\widehat{h}, \widehat{v}) = 0$, the nonnegative quantity inside the expectation is zero almost surely, giving the same lower bound with zero additive error.

For every $x, x' \in \mathcal{X}$, the encoder bound and Assumption 1 give

$$\left\| \widehat{h}(f^\star(x)) - \widehat{h}(f^\star(x')) \right\| \leq \mathsf{L}_{\mathcal{H}} \|f^\star(x) - f^\star(x')\| \leq \mathsf{L}_{\mathcal{H}} \overline{\mathsf{L}}_f \|x - x'\| .$$

Combining the two inequalities proves the result. □

B.3 Identification and the remaining coordinate ambiguity

Proof of Corollary 1. Since $R(\widehat{h}, \widehat{v}) = 0$, Proposition 1 gives, for every $a \in (0,1)$,

$$\mathbb{E}\left[\left(a \|Y - Y'\| - \mathsf{L}_{\mathcal{V},h} \left\| \widehat{h}(Y) - \widehat{h}(Y') \right\| \right)_+^2 \right] = 0.$$

The nonnegative quantity inside the expectation is therefore zero almost surely. Apply this conclusion for $a = 1 - 1/j$, $j \geq 2$, and intersect the resulting countably many events of probability one. Letting $j \to \infty$ gives

$$\|Y - Y'\| \leq \mathsf{L}_{\mathcal{V},h} \left\| \widehat{h}(Y) - \widehat{h}(Y') \right\| \quad \text{almost surely.}$$

Both sides are continuous functions of $(y, y') \in \mathcal{Y}^2$ by (3). The independent pair (Y, Y') has support $\mathcal{Y}^2$. If the inequality failed at one pair, continuity would make it fail on a relative open neighborhood with positive probability, a contradiction. Consequently,

$$\|y - y'\| \le \mathsf{L}_{\mathcal{V},h} \left\|\widehat{h}(y) - \widehat{h}(y')\right\| \qquad (y, y' \in \mathcal{Y}). \tag{B.8}$$

Taking $y = f^\star(x)$ and $y' = f^\star(x')$, Assumption 1 and (B.8) give

$$\underline{\mathsf{L}}_f \|x - x'\| \le \|f^\star(x) - f^\star(x')\| \le \mathsf{L}_{\mathcal{V},h} \left\|\widehat{h}(f^\star(x)) - \widehat{h}(f^\star(x'))\right\|.$$

The encoder bound gives

$$\left\|\widehat{h}(f^\star(x)) - \widehat{h}(f^\star(x'))\right\| \le \mathsf{L}_{\mathcal{H}} \|f^\star(x) - f^\star(x')\| \le \mathsf{L}_{\mathcal{H}} \overline{\mathsf{L}}_f \|x - x'\|.$$

Dividing the first inequality by $\mathsf{L}_{\mathcal{V},h}$ and combining with the second proves (8). The positive lower bound implies injectivity. The image is $\widehat{h}(\mathcal{Y})$ by definition, proving bijectivity.

For another zero-loss pair $(h, v) \in \mathcal{H} \times \mathcal{V}$, the same argument yields

$$\|y - y'\| \le \mathsf{L}_{\mathcal{V},h} \|h(y) - h(y')\| \qquad (y, y' \in \mathcal{Y}).$$

Both encoders are therefore injective on $\mathcal{Y}$. Matching their representations for the same observation defines a bijection between the representation images. The matching is unique because each representation determines exactly one observation. Finally,

$$\|h(y) - h(y')\| \le \mathsf{L}_{\mathcal{H}} \|y - y'\| \le \mathsf{L}_{\mathcal{H}} \mathsf{L}_{\mathcal{V},h} \left\|\widehat{h}(y) - \widehat{h}(y')\right\|,$$
$$\left\|\widehat{h}(y) - \widehat{h}(y')\right\| \le \mathsf{L}_{\mathcal{H}} \|y - y'\| \le \mathsf{L}_{\mathcal{H}} \mathsf{L}_{\mathcal{V},h} \|h(y) - h(y')\|.$$

These inequalities prove the Lipschitz bounds in both directions. □

B.4 Recovery during generation

Proof of Theorem 3. Finite loss makes the squared regression error integrable in time for almost every sampled pair. For each such pair, the triangle inequality, the spatial bound, and Cauchy–Schwarz give

$$\begin{aligned}
&\int_0^1 \left\|\widehat{v}(0, t \mid \widehat{h}(Y_0), \widehat{h}(Y_1))\right\| dt \\
&\quad \le \int_0^1 \left\|\widehat{v}(Y_t, t \mid \widehat{h}(Y_0), \widehat{h}(Y_1)) - (Y_1 - Y_0)\right\| dt \\
&\qquad + \|Y_1 - Y_0\| + \int_0^1 \mathsf{L}_{\mathcal{V}}(t) \|Y_t\| \, dt \\
&\quad \le \left(\int_0^1 \left\|\widehat{v}(Y_t, t \mid \widehat{h}(Y_0), \widehat{h}(Y_1)) - (Y_1 - Y_0)\right\|^2 dt\right)^{1/2} \\
&\qquad + \|Y_1 - Y_0\| + \max\{\|Y_0\|, \|Y_1\|\} \int_0^1 \mathsf{L}_{\mathcal{V}}(t) \, dt < \infty.
\end{aligned}$$

The last inequality uses $\|Y_t\| \le (1 - t)\|Y_0\| + t\|Y_1\| \le \max\{\|Y_0\|, \|Y_1\|\}$. The origin integral is measurable in the observation pair by composition and time integration. The set of pairs in $\mathcal{Y} \times \mathcal{Y}$ where it is finite is therefore measurable and has probability one. All resulting representations lie in $\widehat{h}(\mathcal{Y})$, where the spatial bound applies.

Fix an observation pair in this set, with representation values (c_0, c_1). Integrability of $\mathsf{L}_{\mathcal{V}}$ gives a finite partition of $[0, 1]$ for which the integral of $\mathsf{L}_{\mathcal{V}}$ over each interval is less than one. To see this, absolute continuity of the Lebesgue integral provides $\eta > 0$ such that every measurable set of measure less than η has integral of $\mathsf{L}_{\mathcal{V}}$ less than $1/2$; a finite partition into intervals of length less than η has the required property. On one such interval $[a, b]$, the integral map

$$\gamma \longmapsto \left(t \longmapsto z_a + \int_a^t \widehat{v}(\gamma(s), s \mid c_0, c_1) \, ds\right)$$

maps continuous paths to continuous paths. Its integrand is integrable because a continuous path is bounded and

$$\|\widehat{v}(\gamma(s), s \mid c_0, c_1)\| \le \|\widehat{v}(0, s \mid c_0, c_1)\| + \mathsf{L}_{\mathcal{V}}(s)\,\|\gamma(s)\|\,.$$

In the supremum norm, the integral map has Lipschitz constant at most $\int_a^b \mathsf{L}_{\mathcal{V}}(s)\,ds < 1$. Starting with the constant path $\gamma(t) = z_a$, successive iterations therefore converge uniformly: the differences between successive paths are bounded by a geometric series. The Lipschitz bound allows passage to the limit inside the time integral, so the limit solves the integral equation. The same contraction bound gives uniqueness. Applying this construction on the finitely many consecutive intervals gives the unique absolutely continuous solution on $[0, 1]$. On the measurable set above, the first initial state is y_0. Each iterate is jointly measurable in time and (y_0, y_1) by composition and time integration, with finite integrals by the preceding bounds. Pointwise limits preserve measurability; each interval endpoint supplies the next measurable initial state. Extend by the constant path $t \mapsto y_0$ on the complement. The resulting path and endpoint are measurable without assumptions on unsupported representations.

The sampled interpolation is absolutely continuous, with derivative $Y_1 - Y_0$ and the same initial point as Z_t. For almost every time, the norm of their difference satisfies

$$\begin{aligned}\frac{d}{dt}\,\|Z_t - Y_t\| &\le \left\|\widehat{v}(Z_t, t \mid \widehat{h}(Y_0), \widehat{h}(Y_1)) - (Y_1 - Y_0)\right\| \\ &\le \mathsf{L}_{\mathcal{V}}(t)\,\|Z_t - Y_t\| + \left\|\widehat{v}(Y_t, t \mid \widehat{h}(Y_0), \widehat{h}(Y_1)) - (Y_1 - Y_0)\right\|.\end{aligned}$$

The first inequality is the norm derivative inequality for an absolutely continuous vector; it holds almost everywhere, including the zero set of the difference. The function $t \mapsto \int_0^t \mathsf{L}_{\mathcal{V}}(r)\,dr$ is absolutely continuous. Multiplying by $\exp(-\int_0^t \mathsf{L}_{\mathcal{V}}(r)\,dr)$ and applying the product rule for absolutely continuous functions, then integrating from zero to t, gives

$$\begin{aligned}\|Z_t - Y_t\| \le \int_0^t \exp\!\left(\int_s^t \mathsf{L}_{\mathcal{V}}(r)\,dr\right) \\ \times \left\|\widehat{v}(Y_s, s \mid \widehat{h}(Y_0), \widehat{h}(Y_1)) - (Y_1 - Y_0)\right\|\,ds.\end{aligned} \tag{B.9}$$

At $t = 1$, Cauchy–Schwarz gives

$$\begin{aligned}\|Z_1 - Y_1\|^2 &\le \left(\int_0^1 \exp\!\left(2\int_s^1 \mathsf{L}_{\mathcal{V}}(r)\,dr\right)\,ds\right) \\ &\qquad \times \int_0^1 \left\|\widehat{v}(Y_s, s \mid \widehat{h}(Y_0), \widehat{h}(Y_1)) - (Y_1 - Y_0)\right\|^2\,ds \\ &= \kappa[\mathsf{L}_{\mathcal{V}}] \int_0^1 \left\|\widehat{v}(Y_s, s \mid \widehat{h}(Y_0), \widehat{h}(Y_1)) - (Y_1 - Y_0)\right\|^2\,ds.\end{aligned}$$

Since $\kappa[\mathsf{L}_{\mathcal{V}}] \le \exp(2\int_0^1 \mathsf{L}_{\mathcal{V}}(r)\,dr) < \infty$, taking expectations proves (10). If the loss is zero, the last integrand vanishes for almost every time, for almost every observation pair. Equation (B.9) then gives $Z_t = Y_t$ for every t, on one full-probability set of pairs. For a general pair law Π, use a measurable set of admissible observation pairs of full Π-probability where the spatial bound holds and the origin integral is finite. The displayed estimate proves this finiteness under Π. The same construction applies, using only representations from this pair law. No step uses independence, so taking expectations under Π proves the assertion for $R_\Pi(\widehat{h}, \widehat{v})$. □

C Proofs for Grouped PairedFM

Throughout this appendix, the latent state and pairing law are those in (11). The map $f^\star$ remains continuous and injective on the compact product $\mathcal{U} \times \mathcal{G}$. All assertions about a group g are with respect to the law of G; none requires a positive probability for each individual group.

C.1 Recovery within a group

We first establish the within-group recovery bound. The statement is immediate if $R_\Pi(\widehat{h}, \widehat{v}) = \infty$, so assume the loss is finite. Conditional on $G = g$, the observations $f^\star(U_0, g)$ and $f^\star(U_1, g)$ are independent and have the same law. Their conditional loss is finite for almost every g, because it is nonnegative and its average is $R_\Pi(\widehat{h}, \widehat{v})$. The observations have finite second moments by compactness and continuity. Each group uses observations in $f^\star(\mathcal{U} \times \{g\}) \subseteq \mathcal{Y}$. The spatial bound (4) therefore holds for their representation pairs with the same function $\mathsf{L}_\mathcal{V}(t)$, and hence the same positive coefficient $\beta[\mathsf{L}_\mathcal{V}]$. No encoder Lipschitz bound is used here. The observation part of Theorem 1, applied to this conditional law, therefore gives

$$\mathbb{E}\Big[\Big\|Y - \mathbb{E}[Y \mid \widehat{h}(Y), G]\Big\|^2 \mid G = g\Big] \\ \leq \frac{1}{\beta[\mathsf{L}_\mathcal{V}]}\mathbb{E}\left[\int_0^1 \Big\|\widehat{v}(Y_t, t \mid \widehat{h}(Y_0), \widehat{h}(Y_1)) - (Y_1 - Y_0)\Big\|^2 \, dt \,\middle|\, G = g\right].$$

Here the conditional mean given $(\widehat{h}(Y), G = g)$ is the conditional mean for the observation law in that group. Averaging over g bounds the mean squared recovery error by $R_\Pi(\widehat{h}, \widehat{v})/\beta[\mathsf{L}_\mathcal{V}]$. Measurable versions of these conditional means exist because all variables take values in Euclidean spaces.

C.2 Common coordinates and exact recovery

Proof of Theorem 4: representation. For a Borel set $B \subseteq \mathbb{R}^k$, define

$$\nu_g(B) = \int_\mathcal{U} \mathbf{1}\{\widehat{h}(f^\star(u, g)) \in B\} \, P_U(du).$$

This is the conditional law of either representation given $G = g$. It is a measurable function of g, by measurability of $\widehat{h}$ and $f^\star$ and integration. Conditional on G, the representations are independent with this same law, because U_0 and U_1 are independent and independent of G. Consequently

$$\mathbb{E}[\nu_G(B)^2] = \Pr(\widehat{h}(Y_0) \in B, \widehat{h}(Y_1) \in B).$$

The assumed independence of $\widehat{h}(Y_0)$ and $\widehat{h}(Y_1)$, together with their equal marginal laws, gives

$$\begin{aligned}\Pr(\widehat{h}(Y_0) \in B, \widehat{h}(Y_1) \in B) &= \Pr(\widehat{h}(Y_0) \in B)^2 \\ &= (\mathbb{E}[\nu_G(B)])^2.\end{aligned}$$

Thus $\mathrm{Var}(\nu_G(B)) = 0$, and $\nu_g(B) = \Pr(\widehat{h}(Y_0) \in B)$ for almost every g.

Apply this identity to the countable collection of sets

$$B = \prod_{j=1}^k (-\infty, b_j], \qquad b_j \in \mathbb{Q}.$$

Intersecting the corresponding full-probability sets of groups leaves a set of probability one on which all these equalities hold at once. These rectangles determine a probability law on $\mathbb{R}^k$: arbitrary real lower rectangles follow by decreasing rational approximation, and equality extends to their generated Borel sigma-field by uniqueness of probability measures. Therefore ν_g equals the marginal representation law for every group in this full-probability set.

Intersect further with the full-probability set where $u \mapsto \widehat{h}(f^\star(u, g))$ belongs to $\mathcal{A}$. Choose one group g_* in the intersection. For every other group in the intersection, the maps $u \mapsto \widehat{h}(f^\star(u, g))$ and $u \mapsto \widehat{h}(f^\star(u, g_*))$ belong to $\mathcal{A}$ and have the same output law under $U \sim P_U$. Condition (12) gives

$$\widehat{h}(f^\star(U, g)) = \widehat{h}(f^\star(U, g_*)) \quad P_U\text{-almost surely.}$$

The set where this equality fails is measurable in (u, g). Integrating its indicator with respect to $P_U \otimes P_G$ gives

$$\widehat{h}(Y) = \widehat{h}(f^\star(U, g_*)) \quad \text{almost surely.}$$

Apply the same identity to a second group draw G' independent of (U, G). Outside the union of two null events, both group representations equal $\widehat{h}(f^\star(U, g_*))$, proving (13). This part of the argument does not require zero loss.

Now suppose $R_\Pi(\widehat{h}, \widehat{v}) = 0$. Its nonnegative conditional loss is zero for almost every group. Choose g_* above also from this full-probability set. The within-group recovery bound then supplies a measurable function b such that

$$b(\widehat{h}(f^\star(U, g_*))) = f^\star(U, g_*) \quad P_U\text{-almost surely.}$$

Specifically, take b to be a measurable version of the conditional mean of $f^\star(U, g_*)$ given $\widehat{h}(f^\star(U, g_*))$; the bound makes its squared error zero. Combining this identity with the displayed identity for $\widehat{h}(Y)$ gives $b(\widehat{h}(Y)) = f^\star(U, g_*)$ almost surely. The map $u \mapsto f^\star(u, g_*)$ is continuous and injective on a compact set. Its image is compact, and its inverse on that image is continuous, by the argument in Appendix B.1. Extend this inverse outside the image by any fixed element of $\mathcal{U}$. Applying the measurable extension to $b(\widehat{h}(Y))$ yields U almost surely. Thus U is measurable with respect to $\widehat{h}(Y)$, up to a null event, and

$$\mathbb{E}[U \mid \widehat{h}(Y)] = U = h^\star_U(Y) \quad \text{almost surely.}$$

For almost every g, the two identities

$$\widehat{h}(f^\star(u, g)) = \widehat{h}(f^\star(u, g_*)), \qquad b(\widehat{h}(f^\star(u, g_*))) = f^\star(u, g_*)$$

hold on a measurable set of full P_U-measure. If two points of this set have the same representation in group g, the first identity gives the same representation in group g_*. Applying b and then injectivity of $u \mapsto f^\star(u, g_*)$ makes the points equal. Therefore $u \mapsto \widehat{h}(f^\star(u, g))$ is injective on a set of full P_U-measure for almost every g. No injectivity of the learned encoder was assumed. □

C.3 Transfer between independent observations

Proof of Theorem 4: generation. The independent marginal observations can be written as $Y_a = f^\star(U_a, G_a)$ and $Y_b = f^\star(U_b, G_b)$, with joint latent law

$$(U_a, G_a, U_b, G_b) \sim P_U \otimes P_G \otimes P_U \otimes P_G.$$

It follows that (U_a, U_b, G_a) has exactly the same law as the training variables (U_0, U_1, G). In particular, $(Y_a, f^\star(U_b, G_a))$ has the training pair law (11). The common-representation identity, applied to both product-distributed pairs (U_b, G_b) and (U_b, G_a), gives

$$\widehat{h}(Y_b) = \widehat{h}(f^\star(U_b, G_b)) = \widehat{h}(f^\star(U_b, G_a)) \quad \text{almost surely.}$$

Thus generation from Y_a using $\widehat{h}(Y_b)$ gives the same initial state and representations as generation from Y_a using $\widehat{h}(f^\star(U_b, G_a))$. The pair $(Y_a, f^\star(U_b, G_a))$ has the training law, so Theorem 3 gives existence and uniqueness almost surely. The two choices of observation therefore give the same path. Applying the pair-law version of (10) yields

$$\mathbb{E}\,\|Z_1 - f^\star(U_b, G_a)\|^2 \le \kappa[\mathsf{L}_\mathcal{V}] R_\Pi(\widehat{h}, \widehat{v}),$$

which is (14). If $R_\Pi(\widehat{h}, \widehat{v}) = 0$, this nonnegative squared error is zero almost surely. The path statement of Theorem 3 also gives

$$Z_t = (1 - t) f^\star(U_a, G_a) + t f^\star(U_b, G_a) \quad \text{for every } t \in [0, 1], \quad \text{almost surely.}$$

This proves the remaining assertions. More generally, this proof requires only the common-representation identity, spatial regularity, and finite grouped loss; it does not otherwise use how the common-representation identity was established. □

A concrete class of coordinates Condition (12) does not require the maps in $\mathcal{A}$ to be injective. In one dimension it holds for the class of nondecreasing real-valued maps, for every input law. To prove this, let a, b be such maps with $a(U)$ and $b(U)$ identically distributed. For each real t, the sets

$$\{u : a(u) \le t\}, \qquad \{u : b(u) \le t\}$$

are lower sets in the ordered domain: membership of a point implies membership of every smaller point in that domain. Any two lower sets are nested. Indeed, if each contained a point absent from the other, comparing the two points would contradict the lower-set property of one of the sets. Equality of the output laws gives equal P_U-measure to the two lower sets. Since they are nested, their symmetric difference has measure zero. Hence

$$\mathbf{1}\{a(U) \le t\} = \mathbf{1}\{b(U) \le t\} \quad \text{almost surely.}$$

Take the countable intersection of these events over rational t. If $a(U)$ and $b(U)$ differed there, a rational number strictly between them would make the two indicators differ. Thus $a(U) = b(U)$ almost surely, proving (12).

For a vector $U = (U^{(1)}, \dots, U^{(r)})$, the same property holds for the class

$$a(u) = \big(a_1(u^{(1)}), \dots, a_r(u^{(r)})\big), \qquad a_j \text{ nondecreasing.}$$

Equality of the vector output laws implies equality of each coordinate's output law. The scalar proof then gives $a_j(U^{(j)}) = b_j(U^{(j)})$ almost surely for each j. Intersecting these finitely many probability-one events proves equality of the vectors. No density or independence of the coordinates of U is needed. This class excludes mixing between coordinates; it is one sufficient restriction, not a property of arbitrary smooth encoders.

Independent representations do not align coordinates. This example uses independent $U \sim \mathrm{Unif}[-1, 1]$ and $G \sim \mathrm{Unif}\{0, 1\}$, the observation map $f^\star(u, g) = (u, g, 0)$, and the grouped pairing law (11).

Take

$$\widehat{h}(y) = \cos(\pi y^{(2)}) y^{(1)}, \qquad \widehat{v}(z, t \mid c_0, c_1) = \big(\cos(\pi z^{(2)})(c_1 - c_0), 0, 0\big).$$

On the observation support, $\widehat{h}(f^\star(u, g)) = (-1)^g u$. Every training interpolation has second coordinate g, and therefore

$$\begin{aligned}\widehat{v}(Y_t, t \mid \widehat{h}(Y_0), \widehat{h}(Y_1)) &= \big((-1)^g((-1)^g U_1 - (-1)^g U_0), 0, 0\big) \\ &= (U_1 - U_0, 0, 0).\end{aligned}$$

The loss is zero. On supported representation pairs $|c_1 - c_0| \le 2$, and the elementary bound $|\cos a - \cos b| \le |a - b|$ gives

$$\|\widehat{v}(z, t \mid c_0, c_1) - \widehat{v}(z', t \mid c_0, c_1)\| \le 2\pi |z^{(2)} - z'^{(2)}| \le 2\pi \|z - z'\|.$$

Thus the required spatial regularity holds. Symmetry of the uniform law makes the conditional representation law uniform on $[-1, 1]$ for both values of g. Conditional on either group the two representations are independent with this same law, so their unconditional joint law is also the product of two uniform laws. In particular, $\widehat{h}(Y_0)$ and $\widehat{h}(Y_1)$ are independent. Nevertheless, $\widehat{h}(f^\star(U, 0)) = U$ and $\widehat{h}(f^\star(U, 1)) = -U$ differ almost surely because $\Pr(U = 0) = 0$. The two group maps have the same output law and different values, so they cannot both belong to a class satisfying (12).

The failure also changes generation. Starting at Y_a leaves its second and third coordinates fixed. Its first-coordinate velocity is constant and equal to

$$(-1)^{G_a}(\widehat{h}(Y_b) - \widehat{h}(Y_a)) = (-1)^{G_a + G_b} U_b - U_a.$$

Thus its endpoint is $\big((-1)^{G_a+G_b} U_b, G_a, 0\big)$. When the two groups differ, this equals $(-U_b, G_a, 0)$, rather than $f^\star(U_b, G_a)$, except on a null event. Representation independence alone therefore cannot give the stated factor transfer.