QUOTIENTS BEYOND GROUP ACTIONS: LEARNING VARIABLE EQUIVALENCE WITH LOCAL GEOMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning equivalence classes removes redundancy by identifying different realizations that represent the same input. Existing group-based equivalence methods accommodate redundant variation through a shared transformation structure, using the group action to define both the equivalence classes and their quotient geometry. This restricts the resulting equivalence to classes generated by that action, limiting task-, context-, and realization-dependent or non-geometric equivalence. We propose a novel approach to representation learning that learns a quotient map directly from observed outcomes, instead of deriving it from predefined transformations, by assigning shared coordinates to equivalent realizations. In continuous domains, the learned quotient map realizes equivalence classes as smooth level sets, with its differential and the Riemannian metric defining the tangent spaces and orthogonal decomposition, yielding a local quotient geometry. We use this geometry with flow matching to generate preferred realizations within fixed equivalence classes and to define consistent motion across changing classes. Experiments show that the learned quotient captures equivalence across both image and language domains, while articulated system experiments demonstrate generation and dynamics on the learned quotient geometry.

1 INTRODUCTION

Understanding equivalence classes in data is a fundamental capability in representation learning, as it enables efficiency by collapsing redundancy. Equivalence can remove redundant variation, reduce effective dimension, and improve data efficiency and generalization when learning from data. This idea appears across diverse fields, including Group equivariant convolutional neural networks for vision (Cohen & Welling, 2016), graph learning (Maron et al., 2019), molecular modeling (Satorras et al., 2021), and matrix groups for physical systems (Finzi et al., 2021; Deng et al., 2021; Lin et al., 2024; Kim et al., 2026), where different realizations may represent the same object or result. Early approaches use data augmentation to encourage consistency across selected transformations (Chen et al., 2020), while group-equivariant models build transformations such as rotation and translation directly into the architecture (Cohen & Welling, 2016; Weiler & Cesa, 2019). Group-based quotient models go further by introducing an equivalence class as a set of realizations whose differences do not matter, and represent the entire class as a single element (Wang et al., 2022; Huh et al., 2023; Xu et al., 2026b). By collapsing equivalent realizations, quotient-space removes redundant distinctions, reducing the effective dimension. These approaches are applicable when an appropriate global transformation group is known in advance.

However, many task-dependent equivalences between realizations do not admit a single known group action that is globally valid. For example, a manipulation task may admit disconnected families of valid realizations, such as handle and body grasps of a mug. Equivalence may also not be explicitly geometric, as in language, where questions with different syntax and compositional structure can induce the same answer. Moreover, the equivalence classes can change with the task and context, as two arm configurations can be equivalent when only end-effector position matters, but distinct when orientation is also required by the task. Context-wise, vertical and horizontal reflections keep digits 2 and 5 distinct in ordinary typography, but swap their identities on a seven-segment display (2, 5). Even within a fixed task and context, the changes that preserve the outcome can also depend on the current realization, as the same reflections on a segment display may preserve identities for other digits. A single predefined group action can therefore be too restrictive to capture

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.

all valid equivalences, yet too permissive when locally valid transformations are imposed across the full problem (Petrache & Trivedi, 2023).

This limitation is not resolved simply by learning the transformation structure instead of specifying it. Learned alignment methods still relate inputs through a chosen transformation family (Jaderberg et al., 2015), while canonicalization and frame-based methods retain an underlying group action (Kaba et al., 2023; Puny et al., 2022). Thus, equivalence remains restricted only to class structures generated by the assumed transformation group. Representation-level objectives (Khosla et al., 2020; Zha et al., 2023), latent state models (Gelada et al., 2019), and bisimulation-based methods (Zhang et al., 2021) can identify invariance or task-relevant similarity, but do not explicitly construct the local geometry of the resulting equivalence classes. **The gap:** What remains missing is a construction that identifies equivalent realizations without requiring restrictive transformations between them, while providing local quotient geometry in continuous domains.

Our key observation is that task-relevant behavior itself induces equivalence, and realizations are equivalent when they produce the same relevant outcome. This avoids specifying transformations between them, but defining the relation alone does not produce a useful representation. **The challenge** is to group equivalent realizations while keeping different outcomes distinguishable and, in continuous domains, to recover how realizations can vary locally without changing the outcome. We address both by learning a quotient map whose level sets represent the equivalence classes. When the realization space and learned map are smooth with locally constant rank, these level sets form local fibers and the differential separates outcome-preserving from outcome-changing directions. This local construction requires no global group action or connected equivalence classes, allows rank to vary across regular regions, and recovers the group-based case when fibers coincide with group orbits. **Our contributions** are:

1. *Outcome-defined quotient representations beyond group actions.* We learn shared coordinates for realizations with the same task-relevant outcome, accommodating non-geometric and task-, context-, or realization-dependent equivalence.
2. *Geometry from the learned quotient map.* Where the learned map realizes the behavior relation and has locally constant rank, its level sets model the equivalence classes and its differential identifies behavior-preserving and behavior-changing directions.
3. *Fiber-preserving generation and projectable motion.* Implicit decoders and a target law with the required geometric corrections enable cost-sensitive flow matching within fixed fibers and requested motion across changing fibers.

Image, language, and robotic experiments evaluate the representation, local geometry, probability assignment, and dynamics.

2 BACKGROUND

Equivalence and invariance. Let $\mathcal{X}$ be a set and let a group G act on $\mathcal{X}$. The action assigns $g \cdot x \in \mathcal{X}$ to each $g \in G$ and $x \in \mathcal{X}$, satisfying $e \cdot x = x, (g_1 g_2) \cdot x = g_1 \cdot (g_2 \cdot x)$, where e is the identity element. The orbit of x is $G \cdot x = \{g \cdot x \mid g \in G\}$. A map $f : \mathcal{X} \to \mathcal{Y}$ is *invariant* if $f(g \cdot x) = f(x)$, and when G also acts on $\mathcal{Y}$, it is *equivariant* if $f(g \cdot x) = g \cdot f(x)$ These properties are commonly encouraged through data augmentation or imposed through group-equivariant architectures (Cohen & Welling, 2016).

Quotient spaces and quotient manifolds. When a group G acts on $\mathcal{X}$, the action induces an equivalence relation $x \sim_G x' \iff \exists\, g \in G \textit{ such that } x' = g \cdot x$. The equivalence class of x is $[x]_G = G \cdot x = \{g \cdot x \mid g \in G\}$. The corresponding quotient space is

$$\mathcal{Q} = \mathcal{X}/G = \{G \cdot x \mid x \in \mathcal{X}\}, \qquad \pi(x) = G \cdot x, \tag{1}$$

where $\pi : \mathcal{X} \to \mathcal{Q}$ is the canonical projection. Thus, each element of the quotient represents an entire orbit rather than an individual realization. The quotient therefore represents the intrinsic variability of the system after collapsing symmetry-related degrees of freedom, reducing the redundancy in both representation learning and generative modeling (Huh et al., 2023; Xu et al., 2026b).

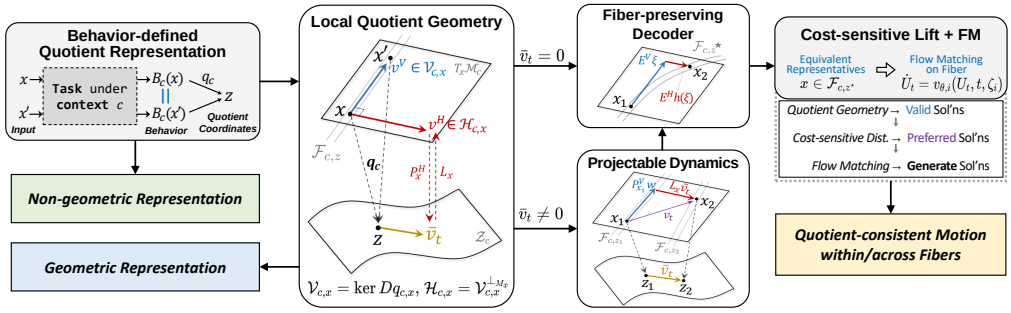


Figure 1: **Method Overview.** Task-relevant behavior defines equivalence, and a learned quotient map assigns equivalent realizations shared coordinates. For smooth realizations, Dq induces local quotient geometry. When $\bar{v}_t = 0$, a fiber-preserving decoder enables cost-guided flow matching within a selected fiber. When $\bar{v}_t \neq 0$, projectable dynamics enable flow matching along changing fibers while enforcing shared quotient motion.

When a Lie group G acts smoothly, freely, and properly on a smooth manifold $\mathcal{M}$, the orbit space $\mathcal{M}/G$ inherits a smooth quotient-manifold structure under standard regularity conditions (Lee, 2012; Xu et al., 2026b).

Conditional flow matching. Flow matching learns a time-dependent vector field whose ordinary differential equation transports a source distribution to a target distribution (Lipman et al., 2023; Chen & Lipman, 2024). For independent samples $a_0 \sim p_0$ and $a_1 \sim p_1(\cdot \mid \zeta)$, a standard conditional path is

$$a_t = (1-t)a_0 + ta_1, \qquad \dot{a}_t = a_1 - a_0. \tag{2}$$

A velocity model $v_\theta(a, t, \zeta)$ is trained with

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}\left[\|v_\theta(a_t, t, \zeta) - (a_1 - a_0)\|_2^2\right]. \tag{3}$$

At the population optimum, $v^\star(a, t, \zeta) = \mathbb{E}[a_1 - a_0 \mid a_t = a, t, \zeta]$, and integrating $\dot{a}_t = v_\theta(a_t, t, \zeta)$ defines the corresponding transport. Riemannian flow matching extends this construction to prescribed manifold geometries (Chen & Lipman, 2024).

3 Methods

As illustrated in Fig. 1, our method first defines equivalence through task-relevant behavior and learns a shared coordinate for each equivalence class (Sec. 3.1-Sec. 3.2). This set-level construction applies across domains without requiring geometric structure. With smooth realization space, the differential of the same learned map further reveals how equivalent realizations vary locally, yielding a local quotient geometry (Sec. 3.3). We use this geometry to shape representations and, when equivalence classes contain continuous degrees of freedom, to construct generation for static and dynamic quotient coordinates (Sec. 3.4-Sec. 3.6)

3.1 Behavior-dependent equivalence

Let $\mathcal{X}_c$ and $\mathcal{B}_c$ denote the set of inputs and resultant behaviors under context c, respectively, and let $B_c : \mathcal{X}_c \to \mathcal{B}_c$ denote the *task-relevant behavior map*, which summarizes the result of x that is relevant under context c. We define

$$x \sim_c x' \quad \Longleftrightarrow \quad B_c(x) = B_c(x'). \tag{4}$$

Since equality in $\mathcal{B}_c$ is reflexive, symmetric, and transitive, the relation induced by B_c is an equivalence relation. Its equivalence classes $[x]_c = \{x' \in \mathcal{X}_c \mid x' \sim_c x\}$ partition $\mathcal{X}_c$. The resulting quotient is the set of these classes

$$\mathcal{Q}_c = \mathcal{X}_c/\!\sim_c = \{[x]_c \mid x \in \mathcal{X}_c\}, \tag{5}$$

with canonical projection $\pi_c(x) = [x]_c$. This set-level quotient requires only the equivalence relation and therefore does not require a group action, topology, or smooth structure. Behavior definitions for each application and their estimation from finite observations are detailed in App. B.

3.2 LEARNING A QUOTIENT REPRESENTATION

With the equivalence classes defined, we seek to assign each class a common coordinate while preserving distinctions between classes. Accordingly, we seek a quotient map $q_c : \mathcal{X}_c \to \mathcal{Z}_c$, where the quotient coordinate space $\mathcal{Z}_c := q_c(\mathcal{X}_c) \subseteq \mathbb{R}^m$ provides coordinates for the quotient, such that

$$q_c(x) = q_c(x') \quad \iff \quad x \sim_c x'. \tag{6}$$

Under this condition, q_c assigns a single coordinate to each equivalence class and distinct coordinates to distinct classes, inducing a well-defined bijection from the quotient set to the quotient coordinate space, $\bar{q}_c : \mathcal{Q}_c \to \mathcal{Z}_c$, given by $\bar{q}_c([x]_c) = q_c(x)$. Eq. (6) is a learning objective rather than an assumed property of the fitted map. The geometry below is defined by the learned q_c and evaluated against behavior empirically. In the discrete language setting, we use the learned quotient representation directly, instantiated through a low-dimensional projection of predicted behavior with complementary reconstruction, with details in App. B and App. J. The domain-specific objectives and training procedures used to learn q_c are detailed in Appendices I-L.

3.3 LOCAL QUOTIENT GEOMETRY

With smooth realization space, local quotient geometry is required to turn equivalence from a set-level partition into a continuous structure that describes how each equivalence class is locally organized. Assume $\mathcal{M}_c \subseteq \mathcal{X}_c$ is a k-dimensional smooth manifold embedded in an ambient $\mathbb{R}^D$ and equipped with a Riemannian metric M_x. We assume that the learned quotient map q_c from Eq. (6) is C^2 on $\mathcal{M}_c$. On regular regions, we obtain the local geometry from its differential, which identifies the tangent directions of each equivalence class, together with the Riemannian metric that induces the corresponding orthogonal decomposition.

On a region where Dq_c has constant rank r, the constant-rank theorem gives a local r-dimensional representation in which q_c is a submersion (Lee, 2012). Thus, realizations sharing a quotient coordinate form smooth local manifolds of dimension $k-r$. When q_c realizes the behavior-defined equivalence relation, these manifolds locally coincide with its equivalence classes. Different constant-rank regions may have different ranks. The construction does not specify transport through rank-changing boundaries. Within each such region, we work with the corresponding local rank-r quotient structure and continue to write q_c for its restriction.

For a quotient coordinate $z \in \mathcal{Z}_c$, the equivalence *fiber* $\mathcal{F}_{c,z}$ is the set of x mapped to z,

$$\mathcal{F}_{c,z} = q_c^{-1}(z) = \{x \in \mathcal{M}_c \mid q_c(x) = z\}, \qquad \dim \mathcal{F}_{c,z} = k - r. \tag{7}$$

For a smooth submersion, the null space of $Dq_{c,x}$ is the tangent space of the fiber and defines the *vertical space* $\mathcal{V}_{c,x}$. Following the standard vertical-horizontal decomposition for Riemannian quotients (Lee, 2018), its metric (M_x)-orthogonal complement defines the *horizontal space* $\mathcal{H}_{c,x}$:

$$\mathcal{V}_{c,x} = T_x \mathcal{F}_{c,q_c(x)} = \ker Dq_{c,x}, \qquad \mathcal{H}_{c,x} = \mathcal{V}_{c,x}^{\perp_{M_x}}. \tag{8}$$

Vertical directions preserve the quotient coordinate to first order, while horizontal directions capture the complementary quotient-changing variation. Since the vertical spaces are kernels of the differential of a single constant-rank map q_c, they form an integrable distribution tangent to its fibers. Thus, these local invariant directions belong to coherent equivalence manifolds rather than defining independent pointwise subspaces. See App. C for details.

Projectors and horizontal lift. Following the standard vertical-horizontal decomposition for smooth submersions (Lee, 2018), we construct the corresponding operators from the learned map q_c. Let $S_x \in \mathbb{R}^{D\times k}$ be an M_x-orthonormal basis of $T_x\mathcal{M}_c$, so that $\delta x = S_x\xi$ represents a tangent vector in local coordinates $\xi \in \mathbb{R}^k$. We set $A_x = Dq_{c,x}S_x$, the quotient differential expressed in these tangent coordinates. On a regular rank-r region, we use r-dimensional local coordinates on the image of q_c and continue to write q_c for this coordinate representation. Then $A_x \in \mathbb{R}^{r\times k}$ has full row rank, so $A_xA_x^\top$ is invertible, with details in App. C. On such regular regions,

$$L_x = S_xA_x^\top(A_xA_x^\top)^{-1}, \qquad P_x^H = L_xA_xS_x^\top M_x, \qquad P_x^V = S_xS_x^\top M_x - P_x^H. \tag{9}$$

Here P_x^V and P_x^H are the metric projectors onto the vertical and horizontal spaces, respectively, and L_x is the horizontal lift from quotient coordinates to $T_x\mathcal{M}_c$. They satisfy $Dq_{c,x}L_x = I_r$ and $Dq_{c,x}P_x^V = 0$, so L_x realizes a specified quotient velocity while P_x^V isolates motion within the current equivalence class.

3.4 FIBER-PRESERVING DECODER

When the fiber dimension $k - r > 0$, each fiber contains continuous degrees of freedom that can be used in two settings. Following Fig. 1, we first consider generation within a static equivalence class ($\bar{v}_t = 0$), and then extend it to dynamic quotient coordinates ($\bar{v}_t \neq 0$).

For a fixed quotient coordinate $z^\star$, generation should vary the realization while preserving $q_c(x) = z^\star$. As illustrated in Fig. 1, the vertical space describes the fiber only infinitesimally, so a finite displacement along a vertical direction can leave the curved fiber. We therefore construct a local fiber-preserving decoder that corrects for fiber curvature while mapping finite local coordinates to realizations. For decoder i anchored at $x_i \in \mathcal{F}_{c,z^\star}$, let Ψ_i be a local manifold parameterization centered at x_i, and let E_i^V, E_i^H be orthonormal bases of the vertical and horizontal coordinate spaces. We introduce unconstrained coordinates $u \in \mathbb{R}^{k-r}$ and map them to bounded vertical displacements $\xi = \varrho_i(u)$, so generation can operate in Euclidean coordinates without leaving the decoder's valid local neighborhood. The term $E_i^V \xi$ moves along the local vertical directions, while a horizontal correction $E_i^H h_i(\xi)$ compensates for fiber curvature. Thus,

$$\varphi_i(u) = \Psi_i\big(E_i^V \xi + E_i^H h_i(\xi)\big), \qquad q_c(\varphi_i(u)) = z^\star. \tag{10}$$

The correction h_i is solved numerically on a valid local branch, preserving the selected learned fiber (Thm. 1). Details of h_i, local decoder construction, and numerical conditions are given in App. D.

3.5 COST-SENSITIVE FIBER DISTRIBUTION AND FLOW MATCHING WITHIN FIBER

The decoder identifies valid alternative realizations for a selected quotient coordinate, while preference among them depends on external criteria, such as feasibility and physical cost. We therefore define a target distribution over the selected fiber. A reference density p_{ref} on $\mathcal{M}_c$ specifies the baseline weighting of realizations, while the feasible set $\mathcal{A}_c$ restricts the distribution to valid solutions. The physical cost C_{phys} then favors lower-cost solutions, with temperature $\tau > 0$ controlling the strength of this preference. The resulting cost-sensitive fiber distribution has density

$$p_z^\star(x) = \frac{\mathbf{1}_{\mathcal{A}_c}(x) p_{\mathrm{ref}}(x) e^{-C_{\mathrm{phys}}(x)/\tau}}{Z(z) J_q(x)}, \qquad x \in \mathcal{F}_{c,z}, \tag{11}$$

where $Z(z)$ is a finite positive normalizer and J_q^{-1} is the coarea factor accounting for the local volume change when conditioning the reference density onto the fiber.

Generation operates in the local decoder coordinates u, so the target distribution on the fiber must be expressed in these coordinates. Accounting for the decoder-induced volume and overlap between local parameterizations gives

$$p_{1,i}(u) \propto \alpha_i(\varphi_i(u)) p_z^\star(\varphi_i(u)) \sqrt{\det G_i(u)}, \qquad G_i(u) = D\varphi_i(u)^\top M_{\varphi_i(u)} D\varphi_i(u). \tag{12}$$

where $G_i(u)$ is the metric induced by decoder φ_i and α_i corrects overlap between local decoder regions. The resulting coordinate distributions, together with the learned probabilities of selecting each local decoder, define the targets used for generation. Details and derivations are in App. E.

Flow matching within fiber Having defined the target distribution in decoder coordinates, we use flow matching to sample from it. For a fixed quotient coordinate $z^\star$, Gaussian source coordinates U_0 and targets $U_1 \sim p_{1,i}$ define $U_t = (1-t)U_0 + tU_1$, and we integrate

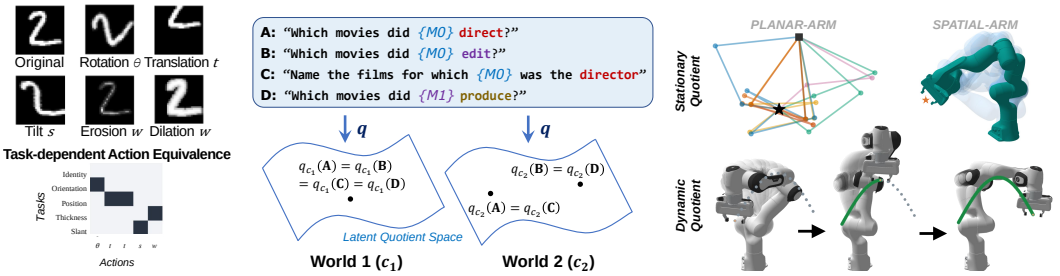
$$\dot{U}_t = v_{\theta,i}(U_t, t, \zeta_i), \qquad X_1 = \varphi_i(U_1). \tag{13}$$

Since $q_c(\varphi_i(u)) = z^\star$, the generated realization remains on the selected fiber while flow matching learns the desired distribution among its representatives.

3.6 PROJECTABLE DYNAMICS FOR DYNAMIC QUOTIENT

For dynamic quotient coordinates ($\bar{v}_t \neq 0$), the realization dynamics must instead move between fibers while remaining well defined at the quotient level. Equivalent realizations may follow different trajectories in realization space, but they must induce the same velocity of their shared quotient coordinate. This consistency across equivalent representatives is the *projectability* condition

$$Dq_{c,x} v_t(x) = \bar{v}_t(q_c(x)). \tag{14}$$



(a) **MNIST Experiment** (b) **Language Experiment** (c) **Planar-/Spatial-Arm Experiment**

Figure 2: **Experimentation across domains. (a)** Image transformations preserve different task-relevant properties, making equivalence task-dependent. **(b)** Language behavior is observed through answer patterns across finite worlds, while executable certificates define positive equivalence. **(c)** Arm configurations share either a static target or a dynamic quotient trajectory. Translucent configurations show all sampled representatives.

We enforce this condition by decomposing the realization-space velocity as

$$v_\theta(x,t) = \underbrace{L_x \bar{v}_\theta(q_c(x),t)}_{\text{inter-fiber}} + \underbrace{P_x^V w_\theta(x,t)}_{\text{intra-fiber}}. \tag{15}$$

Here L_x lifts the shared quotient velocity, while P_x^V restricts auxiliary motion to the current fiber. Since $Dq_{c,x}L_x = I$ and $Dq_{c,x}P_x^V = 0$, Eq. (15) is projectable by construction. Under unique flows on a common regular region, equivalent initial realizations therefore follow the same quotient trajectory (Thm. 3), with group-equivariant flow recovered as a special case (Cor. 4).

Flow matching across fibers For a dynamic quotient path z_s, we use a time-dependent decoder satisfying $q_c(\varphi_{i,s}(u)) = z_s$ and perform flow matching in its local coordinates,

$$\dot{U}_s = v_{\theta,i}(U_s, s, \zeta_i), \qquad X_s = \varphi_{i,s}(U_s), \qquad Dq_{c,X_s}\dot{X}_s = \dot{z}_s. \tag{16}$$

Thus the changing decoder enforces the requested quotient trajectory, while flow matching determines how the representative varies within the corresponding moving fibers. Training, conditioning, decoder construction, and inference details are given in App. F and App. L.4.

4 Experiments

The following experiments instantiate complementary layers of the framework rather than a single common benchmark: images test local quotient geometry (Fig. 2a), language tests the set-level construction without smooth realization space (Fig. 2b), and 2D planar and 3D spatial articulated systems test fiber-preserving generation under static and dynamic quotient coordinates (Fig. 2c).

4.1 Task-dependent local geometry

We train an image encoder on transformed MNIST digits (LeCun et al., 1998) and evaluate on separate MNIST images under seven behavior definitions. All retain digit identity while selectively retaining orientation, position, thickness, or slant, making the same transformation behavior-preserving or changing depending on context and, in one state-dependent setting, image region. For instance, rotation should be ignored when only identity matters, but retained when orientation also matters (Fig. 2a). Full training objectives and implementation details are given in App. I.

Comparisons and evaluation. The base encoder receives only task supervision, while augmentation consistency additionally aligns transformed views. Geometric controls impose predefined transformation structure through feature pooling, orientation-indexed C8 features, or learned spatial alignment with an STN (Jaderberg et al., 2015). SupCon (Khosla et al., 2020) organizes representations using task labels, while behavior-based contrastive learning additionally uses continuous outcome similarity. Subspace controls predict invariant directions with separate fixed, ground-truth, or response-estimated geometry heads, testing whether auxiliary geometry prediction alone is sufficient without shaping the encoder differential. The adapted Structured Flow Autoencoder (SFA) (Xu

et al., 2026a) tests whether adding generative training under the same observable supervision improves the learned representation. Training protocols, information access, and objective details are given in App. I.3. AUROC measures equivalence discrimination, subspace F1 measures agreement of invariant directions, and balanced rank accuracy (RA) measures agreement of their dimensions with equal weighting of the true dimensions 3, 4, and 5. The latter two metrics are computed from the encoder differential itself.

Results and mechanism. Ours achieves the highest means across all three metrics in Tab. 1, with substantially stronger geometric recovery. Despite comparable equivalence AUROC, ours achieves substantially higher rank accuracy than baselines, more often recovering the correct number of independent local directions that preserve the task-relevant outcome. This distinction shows that recognizing equivalent realizations does not by itself ensure recovery of their local geometry. The subspace controls also perform substantially worse on geometric recovery, showing that predicting local geometry in a separate head is not sufficient when the encoder differential itself does not reflect that geometry. Together, these results support a context-dependent local quotient geometry rather than a single global invariance, with transformed views showing the context-dependent distinctions (Fig. 3, Fig. 4). A separate task-conditioned planar study further recovers changes in local rank across workspace regions, testing geometry whose dimension varies with realization (App. K.1). Component ablations are reported in Tab. 7.

Category	Method	AUROC ↑	F1 ↑	RA (%) ↑
Standard	Base encoder	0.6283	0.7281	33.87
	Augmentation consistency	0.6496	0.8043	44.28
Geometric	C4 rotation pooling	0.6359	0.7526	38.92
	Eight-angle rotation pooling	0.6449	0.7816	45.30
	Rotation–translation pooling	0.6533	0.7660	41.59
	C8 rotation features	0.6746	0.7472	38.50
	Learned alignment (STN)	0.6842	0.7256	35.22
Contrastive	SupCon	0.7498	0.8092	51.41
	Behavior-based contrastive	0.7216	0.8167	48.39
Subspace supervision	Fixed subspace per task	0.7205	0.8153	48.34
	Ground-truth subspace head	0.7333	0.8121	48.23
	Learned subspace head	0.7146	0.8052	48.07
Generative	SFA	0.7113	0.8082	48.18
	Ours	**0.7503**	**0.9171**	**82.13**

Table 1: **MNIST Experiment on Task-dependent Local Geometry.** Values are means over six seeds, and **bold** denotes the highest mean in each column.

4.2 Behavioral equivalence in language

We test whether differently phrased questions share a representation when they request the same result. For example, *Which movies did {M0} direct?* and *Which movies were directed by {M0}?* are equivalent, whereas replacing *direct* with *edit* changes the requested answer. We derive executable questions from CFQ (Keysers et al., 2020) with controlled paraphrase, negation/quantity variants, and compositions. A world specifies a synthetic setting of fixed entities and relations, and executing each question across a shared panel of 32 distinct worlds produces a finite behavior signature used for supervision. We instantiate our language model with the BERT-base (Devlin et al., 2019) backbone, whose finite behavior signatures are mapped by Eq. (B.2) to a 128-dimensional quotient representation using text alone at inference. Full data construction, model, training, and executable evaluation details are given in App. J.

Comparisons and evaluation. We evaluate two representation pathways. *Direct text* baselines learn a 128-dimensional representation directly from BERT features, while *Behavior* baselines first predict the 512-dimensional finite behavior signature and then encode it. The autoencoder uses reconstruction, while RNC (Zha et al., 2023), weighted contrastive learning, and SupCon organize either pathway using answer-distance ordering, graded answer similarity, and binary equivalence, respectively. Ours applies the fixed projection and complementary reconstruction of Eq. (B.2) to predicted behavior. All BERT comparisons share answer supervision, sentence streams, output dimension, and 5000 backbone updates, with validation-only selection. AUROC measures equivalence discrimination and FPR_{95} measures false merges at 95% recall. Pairs are labeled equivalent when executable forms specify the same result and distinct when they disagree in at least one world.

Results. Ours achieves the highest reported AUROC and lowest FPR_{95} in Tab. 2, including displayed ties. Low FPR_{95} shows that equivalent wording can share a representation while distinctions in the requested result are retained. Under most objectives, behavior-based representations improve average AUROC and FPR_{95} over direct text, supporting the use of predicted behavior. These results support learning shared coordinates from answer behavior without requiring transformations between equivalent questions, extending the quotient representation to a non-geometric domain.

Representation	Original		Paraphrases		Negation/quantity		Compositions	
	AUROC ↑	FPR_{95} ↓	AUROC ↑	FPR_{95} ↓	AUROC ↑	FPR_{95} ↓	AUROC ↑	FPR_{95} ↓
Direct text autoencoder	0.5356	0.954	0.8148	0.576	0.9212	0.395	0.7077	0.827
Direct text + RNC	0.5288	0.957	0.8053	0.681	0.9264	0.462	0.7229	0.845
Direct text + weighted contrastive	0.9486	0.250	0.9953	0.009	0.9725	0.109	0.9792	0.091
Direct text + SupCon	0.9230	0.270	0.9921	0.021	0.9912	0.038	0.9646	0.178
Behavior autoencoder	0.5895	0.952	0.8399	0.536	0.9366	0.197	0.7679	0.698
Behavior + RNC	0.5422	0.971	0.7968	0.728	0.9188	0.509	0.7093	0.821
Behavior + weighted contrastive	0.9557	0.241	0.9951	0.009	0.9954	0.022	0.9877	0.034
Behavior + SupCon	0.9497	0.147	0.9965	0.011	0.9757	0.073	0.9837	0.057
Ours	**0.9974**	**0.010**	**0.9998**	**0.000**	**0.9997**	**0.002**	**0.9991**	**0.001**

Table 2: **Language equivalence experiment** on CFQ-derived executable questions. The four subsets test original questions, single paraphrases, meaning-changing negation or quantity edits, and compositions of rewrites.

4.3 Flow Matching on Quotient Fiber

With $z^\star$ fixed, we generate different joint configurations that reach the same target while following the cost-sensitive fiber distribution. This experiment separates two requirements of quotient-based generation, preserving the selected equivalence class and learning the distribution over alternative representatives within that class. We experiment in both 2D and 3D space, with 3 and 4 joint planar arms in 2D retaining endpoint position, with obstacles in the 3 joint setting, and the seven-DoF Franka Emika Panda manipulator in 3D retaining both position and orientation, using PyBullet for simulating collision. Analytic kinematics independently evaluate physical accuracy, separating preservation of the learned equivalence class from agreement with the true outcome.

Comparisons and evaluation. Conditional flow matching (FM) (Lipman et al., 2023) generates directly in joint space, while Riemannian Flow Matching (RFM) (Chen & Lipman, 2024) uses geodesic conditional paths on the flat torus, implemented by shortest wrapped interpolation in periodic joint-angle space. FM and RFM with final projection numerically correct generated configurations toward $q(x) = z^\star$ after sampling, while Physics-Constrained Flow Matching (PCFM) (Utkarsh et al., 2025) applies corrections throughout sampling. All methods use the same target distribution, and all numerical corrections use the same frozen quotient map q and its Jacobian. This controlled comparison holds the learned equivalence constraint and target distribution fixed, isolating the generation mechanism rather than differences in learning q. It tests whether the fiber-preserving decoder enables more accurate modeling of the representative distribution than ambient-space generation or numerical constraint correction alone. Source distributions and model inputs follow each method, with implementation details in App. K.2 and App. L.

We evaluate $2{,}048$ configurations per condition over three seeds, using 64 planar targets and 128 Panda target and scene requests. Planar success rate (SR) requires physical endpoint accuracy, feasibility, and numerical validity. Panda pose SR counts physically accurate poses over the requested samples. SW_1 measures joint-distribution error, while component total variation (TV) on Panda measures probability error across disconnected solution families. Acceptance checks, missing-output accounting, and evaluation tolerances are given in App. L.

Results. Without numerical correction, FM and RFM rarely preserve the selected target (Tab. 3). Projection-based methods recover high SR, showing that explicit constraint enforcement can recover valid target-equivalent configurations, but their representative distributions remain substantially different. Ours achieves the lowest SW_1 in all three planar settings and the highest SR with obstacles. PCFM attains the highest SR for 3 and 4 joints, but with substantially larger SW_1, giving a clear

Method	3 Joint		3 Joint+obs.		4 Joint	
	SR (%) ↑	SW_1 ↓	SR (%) ↑	SW_1 ↓	SR (%) ↑	SW_1 ↓
Conditional FM	0.079	0.0366	0.068	0.0661	0.066	0.0294
RFM	0.123	0.0341	0.111	0.0697	0.108	0.0288
FM + final proj.	99.988	0.0325	90.383	0.0617	75.683	0.0275
PCFM	**99.997**	0.2506	92.085	0.2293	**75.977**	0.2561
RFM + final proj.	99.986	0.0319	89.380	0.0665	75.545	0.0277
Ours	99.992	**0.0232**	**97.085**	**0.0328**	75.128	**0.0267**

Table 3: **Planar-arm fixed-target generation results.** SR measures physical endpoint success, while SW_1 measures joint-distribution error.

Method	Pose SR (%) ↑	SW_1 ↓	Comp. TV ↓
Conditional FM	0.000	—	—
RFM	0.000	—	—
FM + final proj.	**95.788**	0.1199	0.2545
PCFM	93.252	0.1613	0.3478
RFM + final proj.	95.165	0.1057	0.2318
Ours	95.180	**0.0211**	**0.0467**

Table 4: **Spatial-arm fixed-target generation results.** Pose SR counts missing outputs as failures. Distribution errors use accepted configurations, with — indicating no accepted outputs.

separation between target satisfaction and distribution matching. Under the same learned equivalence constraint, these results support more accurate matching of the cost-sensitive representative distribution while maintaining near-best SR.

On Panda, uncorrected FM and RFM produce no successful outputs (Tab. 4). Among corrected methods, ours achieves pose SR comparable to RFM with final projection and slightly below FM with final projection, while reducing both SW_1 and component TV by about 80% relative to the strongest distribution-matching baseline. SW_1 measures the overall joint distribution, while component TV measures probability allocation across disconnected solution families. Their joint improvement shows that the gain is not higher pose accuracy, but more accurate probability assignment both within and across equivalent solution families.

4.4 Flow Matching with Quotient Motion

We evaluate Eq. (16) on prescribed position and pose (position + orientation) paths for the Panda arm, testing flow matching with changing quotient coordinates on four- and one-dimensional regular fibers, respectively. We consider hold, straight, curved, and loop movement tasks.

Comparisons and evaluation. Ambient FM predicts joint-space velocities directly, FM + step retraction corrects intermediate states onto the moving constraint, and Constrained FM uses the projectable parameterization in Eq. (15). Ours applies flow matching in the moving decoder coordinates. Unlike RFM in the fixed-target comparison, adapted RFM (Chen & Lipman, 2024) here uses our local geometry and the same moving implicit decoder. Both therefore follow the prescribed quotient path through the same decoder, but differ in how they learn motion among equivalent configurations. Ours learns from linear interpolation in the decoder coordinates, whereas adapted RFM constructs training paths using the induced metric of the initial fiber, with arc-length interpolation for pose and coordinate distance rather than geodesic distance for position. The metric remains fixed in time, while the decoder follows the prescribed quotient path. All methods share the frozen learned outcome map, path requests, initial representatives, target distribution, and local geometry. SR measures successful simulated execution under a shared controller, and SW_1 measures terminal joint-distribution error over completed numerical rollouts. Success requires 5 mm position accuracy and, for pose, 1° orientation accuracy, with completion and execution protocols given in App. L.4.

Results and mechanism. Ours achieves the highest reported SR for both position and pose motion and the lowest SW_1 for pose (Tab. 5). FM + step retract. and constrained FM approach our SR but have higher SW_1, especially for pose. Adapted RFM gives closer SW_1 and lower position SW_1, but substantially lower SR on both tasks. Since RFM also uses the moving decoder, these differences concern learned transport and execution within shared local geometry, rather than whether the prescribed quotient motion is included. Its lower position SW_1 is conditional on completed rollouts and should therefore be read alongside execution success. The no-transport ablation retains the moving decoder while fixing the local coordinate and substantially worsens SW_1 (App. L.4), isolating the benefit of learned coordinate transport from preservation of the prescribed quotient path. A planar diagnostic verifies projectability, with equivalent configurations inducing the same quotient trajectory up to numerical error (App. K.4).

	Position tracking		Pose tracking	
Method	SR (%) ↑	SW_1 ↓	SR (%) ↑	SW_1 ↓
Ambient FM	2.08	0.01438	11.46	0.01728
FM + step retract.	95.83	0.01401	94.79	0.01639
Constrained FM	95.83	0.01400	94.79	0.01633
RFM (adapted)	88.54	**0.01045**	90.63	0.00319
Ours	**98.96**	0.01139	**97.92**	**0.00254**

Table 5: **Flow Matching with Quotient Motion.** Three-seed means of SR and SW_1.

5 Conclusion

We introduced a framework that learns equivalence from task-relevant outcomes without prescribing a group action. Its quotient representation identifies shared behavior, and in continuous domains, its differential defines local quotient geometry. Image and language experiments support equivalence recognition and task-dependent geometric recovery, while robot experiments demonstrate the complementary roles of finite fiber preservation and probability assignment. Moving-target experiments extend this construction to alternative trajectories that realize a shared evolving outcome. Together, these contributions enable a model to separate what outcome must be preserved from which realization should be used to achieve it. We plan to extend this framework to real-world manipulation, tasks spanning longer horizons, and world models that operate in learned quotient spaces.

AI USE STATEMENT

We received assistance from AI in executing baseline experiments, debugging code, constructing synthetic language augmentations, retrieving relevant literature, and suggesting minor wording and grammar corrections. For baseline experiments, LLM agents assisted with implementing and setting up baseline methods from their original descriptions and adapting them to our experimental settings. The core implementation of our proposed method and the overall experimental codebase were developed by the authors, with LLM assistance used for debugging, identifying implementation issues, and suggesting code corrections. The language augmentations use controlled rewrites of CFQ-derived questions, including active/passive voice, conjunction order, wording changes, and their compositions, together with variants that change negation or quantity (Appendix J). We take responsibility for reviewing all AI-assisted material, verifying references and research claims, and checking generated data and experimental artifacts. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses MNIST images, controlled CFQ-derived executable questions, and simulated robot arms, with no human-subject recruitment or physical-robot deployment. The main risk is that an inaccurate behavior definition or learned quotient map may collapse distinctions that matter outside the evaluated task. Accordingly, language equivalence and robotic accuracy are evaluated independently of the learned representation where possible. The reported results do not constitute guarantees of unrestricted semantic equivalence, real-world safety, or deployment reliability.

REPRODUCIBILITY STATEMENT

Section 3 defines the quotient representation, local geometry, implicit decoder, target distribution, and flow-matching constructions. App. B–App. G provide the supporting derivations, proofs, and assumptions, while App. H distinguishes the separately trained models. App. I and App. J describe data construction, splits, model architectures, training objectives, and evaluation protocols for the MNIST and language experiments. App. K and App. L specify the arm systems, learned geometry, local decoder construction, training targets, comparison methods, and metrics. App. L.4 details moving-target flow matching and simulated execution. These sections also specify model selection, seed counts, and the treatment of numerical failures, rejected samples, and incomplete trajectories.

REFERENCES

Michael Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Barrett Ames, Jeremy Morgan, and George Konidaris. IKFlow: Generating diverse inverse kinematics solutions. *IEEE Robotics and Automation Letters*, 7(3):7177–7184, 2022.

Ricky T. Q. Chen and Yaron Lipman. Flow matching on general geometries. In *The Twelfth International Conference on Learning Representations*, 2024.

Shuxiao Chen, Edgar Dobriban, and Jane H Lee. A group-theoretic framework for data augmentation. *Journal of Machine Learning Research*, 21(245):1–71, 2020.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In *International conference on machine learning*, pp. 2990–2999. PMLR, 2016.

Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J Guibas. Vector neurons: A general framework for SO(3)-equivariant networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12200–12209, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.

Herbert Federer. *Geometric measure theory*. Springer, 2014.

Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite markov decision processes. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pp. 162–169. AUAI Press, 2004. ISBN 0974903906.

Marc Finzi, Max Welling, and Andrew Gordon Wilson. A practical method for constructing equivariant multilayer perceptrons for arbitrary matrix groups. In *International conference on machine learning*, pp. 3318–3328. PMLR, 2021.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. DeepMDP: Learning continuous latent space models for representation learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2170–2179. PMLR, 2019.

Nikolaos Gkanatsios, Jiahe Xu, Matthew Bronars, Arsalan Mousavian, Tsung-Wei Ke, and Katerina Fragkiadaki. 3d flowmatch actor: Unified 3d policy for single-and dual-arm manipulation. *arXiv preprint arXiv:2508.11002*, 2025.

In Huh, Changwook Jeong, Jae Myung Choe, Young-Gu Kim, and Dae Sin Kim. Isometric quotient variational auto-encoders for structure-preserving representation learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pp. 39075–39087, 2023.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Sékou-Oumar Kaba, Arnab Kumar Mondal, Yan Zhang, Yoshua Bengio, and Siamak Ravanbakhsh. Equivariance with learned canonicalization functions. In *International Conference on Machine Learning*, pp. 15546–15566. PMLR, 2023.

Tsung-Wei Ke, Nikolaos Gkanatsios, and Katerina Fragkiadaki. 3d diffuser actor: Policy diffusion with 3d scene representations. In *8th Annual Conference on Robot Learning*, 2024.

Mete Kemertas and Tristan Aumentado-Armstrong. Towards robust bisimulation metric learning. *Advances in Neural Information Processing Systems*, 34:4764–4777, 2021.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*, 2020.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Chankyo Kim, Sicheng Zhao, Minghan Zhu, Tzu-Yuan Lin, and Maani Ghaffari. Equivariant neural networks for general linear symmetries on Lie algebras. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=TqL8MOHtW7.

Jonas Köhler, Leon Klein, and Frank Noe. Equivariant flows: Exact likelihood generative learning for symmetric densities. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5361–5370. PMLR, 2020.

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

John M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, 2 edition, 2012.

John M Lee. *Introduction to Riemannian manifolds*, volume 2. Springer Cham, 2018.

Oliver Limoyo, Filip Marić, Matthew Giamou, Petra Alexson, Ivan Petrović, and Jonathan Kelly. Generative graphical inverse kinematics. *IEEE Transactions on Robotics*, 41:1002–1018, 2025.

Tzu-Yuan Lin, Minghan Zhu, and Maani Ghaffari. Lie Neurons: Adjoint-equivariant neural networks for semisimple Lie algebras. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 30529–30545. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/lin24aa.html`.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International conference on learning representations (ICLR)*, 2023.

Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and equivariant graph networks. In *International Conference on Learning Representations*, 2019.

Mircea Petrache and Shubhendu Trivedi. Approximation-generalization trade-offs under (approximate) group equivariance. *Advances in Neural Information Processing Systems*, 36:61936–61959, 2023.

Omri Puny, Matan Atzmon, Heli Ben-Hamu, Ishan Misra, Aditya Grover, Edward J. Smith, and Yaron Lipman. Frame averaging for invariant and equivariant network design. In *International Conference on Learning Representations*, 2022.

Vıctor Garcia Satorras, Emiel Hoogeboom, and Max Welling. E (n) equivariant graph neural networks. In *International conference on machine learning*, pp. 9323–9332. PMLR, 2021.

Patrice Simard, Bernard Victorri, Yann LeCun, and John Denker. Tangent prop—a formalism for specifying selected invariances in an adaptive network. In *Advances in Neural Information Processing Systems*, volume 4, 1991.

Utkarsh Utkarsh, Pengfei Cai, Alan Edelman, Rafael Gómez-Bombarelli, and Christopher Vincent Rackauckas. Physics-constrained flow matching: Sampling generative models with hard constraints. In *Advances in Neural Information Processing Systems*, 2025.

Dian Wang, Robin Walters, Xupeng Zhu, and Robert Platt. Equivariant q learning in spatial action spaces. In *Conference on Robot Learning*, pp. 1713–1723. PMLR, 2022.

Maurice Weiler and Gabriele Cesa. General e (2)-equivariant steerable cnns. *Advances in neural information processing systems*, 32, 2019.

Yidan Xu, Yixin Wang, and XuanLong Nguyen. Structured flow autoencoders: Learning structured probabilistic representations with flow matching. In *International Conference on Learning Representations*, 2026a.

Yixian Xu, Yusong Wang, Shengjie Luo, Kaiyuan Gao, Tianyu He, Di He, and Chang Liu. Quotient-space diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Hongyu Zang, Xin Li, Leiji Zhang, Yang Liu, Baigui Sun, Riashat Islam, Remi Tachet des Combes, and Romain Laroche. Understanding and addressing the pitfalls of bisimulation-based representations in offline reinforcement learning. *Advances in Neural Information Processing Systems*, 36:28311–28340, 2023.

Kaiwen Zha, Peng Cao, Jeany Son, Yuzhe Yang, and Dina Katabi. Rank-N-Contrast: Learning continuous representations for regression. In *Advances in Neural Information Processing Systems*, 2023.

Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

A Related work

Equivalence, invariance, and representations. Data augmentation encourages consistency under selected transformations (Chen et al., 2020), while group-equivariant architectures build a known transformation action into the model (Cohen & Welling, 2016; Weiler & Cesa, 2019; Maron et al., 2019; Satorras et al., 2021; Finzi et al., 2021). Canonicalization learns representatives while retaining an underlying transformation group (Kaba et al., 2023). IQVAE similarly assumes a specified symmetry group and combines group-invariant encoding, quotient reconstruction, and a Riemannian isometry objective (Huh et al., 2023). Our equivalence is instead defined from task-relevant behavior, without requiring an explicit transformation between equivalent realizations. In smooth realization spaces, regular level sets of the learned quotient map provide the corresponding local geometry, with group-based quotients recovered as a special case.

Local differential constraints. Tangent propagation constrains network derivatives along selected infinitesimal transformations (Simard et al., 1991). Our image objectives also use derivatives to identify behavior-preserving directions, but these directions arise as kernels of a single learned quotient map. On constant-rank regions, they therefore form an integrable distribution tangent to its fibers, while independently predicted projectors do not necessarily define coherent equivalence manifolds (App. C).

Flow matching on manifolds and quotients. Flow matching and rectified flow learn deterministic probability transport through time-dependent velocity fields (Lipman et al., 2023; Liu et al., 2023). Riemannian Flow Matching extends this construction to supplied manifold geometries (Chen & Lipman, 2024), while stochastic interpolants connect deterministic and stochastic transport between distributions (Albergo et al., 2025). Our use of flow matching follows these transport objectives after expressing the target distribution in local decoder coordinates. The distinction lies in learning the quotient geometry, constructing a fiber-preserving decoder, and defining the cost-sensitive target distribution.

QSDM constructs quotient diffusion from an isometric group action and horizontally lifts the quotient dynamics (Xu et al., 2026b). Its horizontal construction removes the need to predict vertical components associated with the group action. Our setting instead learns equivalence from task-relevant behavior and retains within-fiber variation because equivalent realizations can differ in feasibility or physical cost. Projectability requires equivalent representatives to induce the same quotient velocity while allowing different motions within the fiber. Group-equivariant dynamics are recovered under the assumptions of Cor. 4.

Control and reinforcement learning. Bisimulation uses reward and transition behavior to identify task-relevant similarity between states (Ferns et al., 2004; Zhang et al., 2021; Kemertas & Aumentado-Armstrong, 2021; Zang et al., 2023). Our equivalence is instead defined directly from task-relevant behavior under the fixed context of each experiment. Equivariant Q-learning assumes known transformations of observations and actions and imposes the corresponding equivariance (Wang et al., 2022). Diffusion and flow policies model multimodal robot actions or trajectories (Chi et al., 2025; Ke et al., 2024; Gkanatsios et al., 2025), while our arm experiments study generation and dynamics under a selected target or quotient path using a frozen learned quotient map.

Attribution of the mathematical ingredients. The quotient and regular-fiber arguments use standard differential topology (Lee, 2012), while the vertical-horizontal decomposition uses Riemannian geometry (Lee, 2018). The implicit decoder uses the implicit function theorem, and the language construction uses standard linear algebra. The target distribution uses the coarea and area formulas (Federer, 2014), while flow matching follows existing deterministic transport objectives (Lipman et al., 2023; Chen & Lipman, 2024). Our contribution is the construction connecting outcome-defined quotient representations, local geometry from the learned quotient map, and cost-sensitive fiber-preserving generation with projectable motion.

B BEHAVIOR AND COHERENT REPRESENTATION

Behavior and finite evidence. Equality of any specified map B_c defines an equivalence relation. For images, B_c retains identity and context-selected transformation quantities, while for arms it is the endpoint or pose. For language, $B_c(x)$ denotes the executable behavior of question x, equivalently its answer function over admissible worlds. Given a finite panel $\mathcal{W} = \{w_1, \ldots, w_L\}$, we observe the finite signature

$$B_c^{\mathcal{W}}(x) = \big(\phi_{\mathrm{ans}}(\mathrm{answer}(x, w_\ell))\big)_{\ell=1}^{L}. \tag{B.1}$$

Equality of B_c implies equality of $B_c^{\mathcal{W}}$, but agreement on a finite panel does not establish executable equivalence. Accordingly, language positives require executable equivalence certificates, negatives require an observed answer disagreement, and unresolved pairs remain unresolved (App. J). The image and arm protocols specify their observed outcomes in App. I and App. K.1.

Projection and complementary reconstruction. For domains where a lower-dimensional quotient coordinate is useful, we construct one directly from the predicted task-relevant behavior. In the language setting, the model predicts the finite observed behavior $\widehat{B}^{\mathcal{W}}(x)$, which we project to a lower-dimensional coordinate while using a complementary reconstruction to preserve the relation between this coordinate and the modeled behavior. A reconstruction head then models behavior information complementary to this coordinate. Crucially, this complementary component is restricted so that it cannot alter the quotient coordinate itself.

For full-row-rank $C \in \mathbb{R}^{d_q \times m}$, let $C^\dagger = C^\top (CC^\top)^{-1}$. We define

$$q(x) = C\big(\widehat{B}^{\mathcal{W}}(x) - b\big), \qquad \psi(z) = b + C^\dagger z + (I - C^\dagger C) h_\psi(z), \qquad \widetilde{B}^{\mathcal{W}} = \psi \circ q. \tag{B.2}$$

Here $\widehat{B}^{\mathcal{W}}(x)$ is the predicted finite behavior signature, b centers its representation, and C selects the d_q coordinates retained by the quotient. The term $C^\dagger z$ maps the retained coordinates back to behavior space, while $(I - C^\dagger C)h_\psi(z)$ models complementary behavior components predictable from z without changing the quotient coordinate. The language instantiation uses orthonormal rows of C, giving $C^\dagger = C^\top$, with the construction of C and b detailed in App. J.

This separation makes the role of the quotient coordinate explicit. Since $CC^\dagger = I_{d_q}$ and $C(I - C^\dagger C) = 0$,

$$C(\psi(z) - b) = z. \tag{B.3}$$

Thus, reconstruction through the complementary component cannot modify the retained quotient coordinate. When differentiable, $CD\psi_z = I_{d_q}$, so both ψ and $D\psi_z$ are injective. Consequently, $q(x) = q(x')$ if and only if $\widetilde{B}(x) = \widetilde{B}(x')$, and $\ker D\widetilde{B}_x = \ker Dq_x$. The completion therefore preserves both equality of the modeled quotient coordinates and their local null directions.

These identities establish consistency between the quotient coordinate and its complementary reconstruction. Agreement with the observed finite behavior also depends on behavior-prediction and reconstruction error. If $\|B^{\mathcal{W}} - \widehat{B}^{\mathcal{W}}\| \le \varepsilon_B$ and $\|\widehat{B}^{\mathcal{W}} - \widetilde{B}^{\mathcal{W}}\| \le \varepsilon_R$ on the compared support, and ψ is L_ψ-Lipschitz, then

$$\|B^{\mathcal{W}}(x) - B^{\mathcal{W}}(x')\| \le 2(\varepsilon_B + \varepsilon_R) + L_\psi \|q(x) - q(x')\|. \tag{B.4}$$

Thus quotient-coordinate agreement controls agreement of the finite observed behavior only up to prediction and reconstruction error. Executable equivalence remains defined by B_c, rather than finite-panel agreement in $B_c^{\mathcal{W}}$.

C LOCAL GEOMETRY AND REPRESENTATION CONDITIONS

Regular fibers. The map $[x]_q \mapsto q(x)$ identifies the quotient set with $q(\mathcal{M})$. For a C^2 submersion $q : \mathcal{M}^k \to \mathcal{Z} \subset \mathbb{R}^r$, local coordinates (z, u) satisfy $q(z, u) = z$ by the submersion theorem (Lee, 2012). Therefore $\dim \mathcal{F}_z = k - r$ and $T_x\mathcal{F}_z = \ker Dq_x$. For redundant outcome coordinates, use local coordinates on the constant-rank image. Different rank regions are treated separately. This argument does not supply a smooth quotient across rank-changing boundaries or one chart covering disconnected components.

Integrability. For vertical C^1 fields a, b, $aq^j = bq^j = 0$, so $[a, b]q^j = 0$ and $Dq[a, b] = 0$. The local level sets above are integral manifolds of this distribution. By contrast, $\mathrm{span}\{\partial_{x_1}, \partial_{x_2} + x_1\partial_{x_3}\}$ has a smooth rank-two projector but bracket ∂_{x_3} outside its span. Thus independent pointwise projectors need not define equivalence surfaces. A connected integral manifold need not exhaust a disconnected level set.

Metric projectors and the horizontal lift. For an M_x-orthonormal supported frame S_x, write $A = Dq_x S_x$ with full row rank. In support coordinates, $\Pi^H = A^\top(AA^\top)^{-1}A$ and $\Pi^V = I - \Pi^H$ are complementary orthogonal projectors, with $A\Pi^V = 0$. Transporting through S_x gives Eq. (9), including $Dq_x L_x = I_r$ and $Dq_x P_x^V = 0$. Every supported solution to $Dq_x v = b$ is $v = S_x(A^\top(AA^\top)^{-1}b + \eta)$, $\eta \in \ker A$. The two terms are orthogonal, proving that $L_x b$ is the unique minimum-M_x-norm lift. Likewise, for a supported prediction f,

$$\arg\min_{Dq_x v = b} \tfrac{1}{2}\|v - f\|_{M_x}^2 = f + L_x(b - Dq_x f) = L_x b + P_x^V f. \tag{C.1}$$

Thus equality-constrained velocity projection and the projectable parameterization coincide under these assumptions.

Numerical rank. Surjectivity requires $\sigma_{\mathrm{sur}}(A) = \sqrt{\lambda_{\min}(AA^\top)} > 0$ on the used region. A small singular value or a finite secant is not an exact null direction. With damping, $DqL_\lambda = I - \lambda(AA^\top + \lambda I)^{-1}$, so the right-inverse identity is only approximate. This differs from damped iterations that converge to an exact nonlinear constraint root.

When an encoder represents the quotient. For differentiable E, if $DE_x P_x^V = 0$ and $DE_x|_{\mathcal{H}_x}$ is injective, then $\ker DE_x = \mathcal{V}_x$. Indeed, decompose $v = v_V + v_H$; $DE_x v = 0$ forces $v_H = 0$. In submersion coordinates, $\partial_u E = 0$ then gives $E(z, u) = \bar{E}(z)$ locally. Global factorization additionally requires constancy across all pieces of each fiber, and class separation requires injectivity of $\bar{E}$. For a horizontal orthonormal frame H_x, the needed injectivity is

$$\sigma_{\mathrm{inj}}(DE_x H_x) = \inf_{\|h\|=1} \|DE_x H_x h\| > 0. \tag{C.2}$$

The horizontal objective in Eq. (I.2) uses each of these directions. For a unit-normalized p-dimensional encoding, the derivative has image in a $(p-1)$-dimensional tangent space, requiring $r \le p - 1$. The MNIST losses encourage these properties but do not make them exact architectural identities.

D Implicit decoding and numerical conditions

Let Ψ_i be an injective C^2 support chart around an anchor x_i with $q(x_i) = z^\star$. Choose orthonormal vertical and horizontal coordinate frames E_i^V, E_i^H at the anchor and define

$$F_i(h, \xi) = q\big(\Psi_i(E_i^V \xi + E_i^H h)\big) - z^\star. \tag{D.1}$$

Assume $D_h F_i(0, 0)$ is nonsingular. The implicit function theorem gives a unique local C^2 correction $h_i(\xi)$ through the anchor. Choose R_i so that the closed vertical ball lies inside this valid graph neighborhood and set

$$\varrho_i(u) = R_i u(1 + \|u\|^2)^{-1/2}, \qquad \varphi_i(u) = \Psi_i(E_i^V \varrho_i(u) + E_i^H h_i(\varrho_i(u))). \tag{D.2}$$

Theorem 1 (Fiber-preserving decoding)**.** *On the specified valid branch, every decoded coordinate trajectory $X_t = \varphi_i(U_t)$ satisfies $q(X_t) = z^\star$.*

Proof. The correction satisfies $F_i(h_i(\xi), \xi) = 0$, and ϱ_i maps $\mathbb{R}^{d_i}$ into its valid vertical ball. Substitution proves the claim. For $K_H = Dq\, D\Psi_i E_i^H$ and $K_V = Dq\, D\Psi_i E_i^V$, differentiation gives $Dh_i = -K_H^{-1} K_V$, hence

$$D\varphi_i = D\Psi_i(E_i^V - E_i^H K_H^{-1} K_V) D\varrho_i, \qquad Dq_{\varphi_i(u)} D\varphi_i(u) = 0. \tag{D.3}$$

All differentials here are evaluated at the current graph point. □

The radial map is a diffeomorphism onto the open ball, with $\varrho_i^{-1}(\xi) = \xi/\sqrt{R_i^2 - \|\xi\|^2}$ and $\det D\varrho_i = R_i^{d_i}(1 + \|u\|^2)^{-(d_i+2)/2} > 0$. The graph is an injective immersion because projection onto E_i^V recovers ξ; therefore so is φ_i, and its pullback metric is positive definite. This justifies its chart-volume factor.

Numerical solver limitations. Small residual and full rank do not by themselves identify the intended local branch. For example, $e^v \sin h = 0$ admits both $h = 0$ and $h = \pi$. The experiments therefore use finite branch, residual, and support checks and report failures, coverage, and true outcome accuracy separately. The identities above apply to the exact implicit branch. For approximate solves, a small residual alone does not control derivative error. The volume factor must correspond to the decoder being evaluated. Orthogonal changes of the vertical frame preserve the physical decoder when its coordinates and graph are transformed consistently. A decoder mixture retains its local label and does not require a single vector field on overlaps.

E Conditional measures and chart probabilities

Conditioning on a regular fiber. Fix a regular submersion q and a reference density p_{ref} relative to the declared support volume. For this derivation, the explicit feasibility restriction in Eq. (11) is absorbed into p_{ref} by restriction and normalization, assuming positive feasible reference mass. In local coordinates $\Gamma(z, u)$ with $q(\Gamma(z, u)) = z$, write the support metric as $G = \left(\begin{smallmatrix} G_{zz} & G_{zu} \\ G_{uz} & G_{uu} \end{smallmatrix}\right)$. Then $J_q^2 = \det((G^{-1})_{zz}) = \det G_{uu} / \det G$. The coordinate volume formula and a partition of unity give coarea (Federer, 2014):

$$\int f(x) J_q(x) \, \mathrm{d}\,\mathrm{vol}_{\mathcal{M}}(x) = \int_{\mathcal{Z}} \int_{\mathcal{F}_z} f(x) \, \mathrm{d}\,\mathrm{vol}_{\mathcal{F}_z}(x) \, \mathrm{d}z. \tag{E.1}$$

Applying it to $f p_{\mathrm{ref}}/J_q$ gives the conditional reference density proportional to p_{ref}/J_q on the fiber. At a requested regular value we use this coarea-defined version, assuming a finite positive normalizer; disintegration alone defines conditionals only almost everywhere. Cost tilting by $e^{-\beta \mathcal{C}}$ gives Eq. (11). For $\tau = 1/\beta > 0$, its Gibbs optimality follows from

$$\mathbb{E}_P \mathcal{C} + \tau \, \mathrm{KL}(P \| \mu_{\mathrm{ref}}(\cdot \mid z)) = \tau \, \mathrm{KL}(P \| P_z^\star) - \tau \log \mathbb{E}_{\mu_{\mathrm{ref}}(\cdot \mid z)} e^{-\mathcal{C}/\tau}. \tag{E.2}$$

Nonnegativity of KL proves uniqueness when the displayed objective is finite. The case $\beta = 0$ retains the conditional reference. This is a fiberwise preference law, not a trajectory-optimality claim.

Chart mixture. Let $\{\varphi_i\}$ be a finite cover of the target support by valid injective charts, and let $\alpha_i \geq 0$ form a partition of unity subordinate to their images. With $j_i = \sqrt{\det(D\varphi_i^\top M D\varphi_i)}$, define

$$\begin{gathered} b_i(u) = \frac{\alpha_i(\varphi_i(u)) p_{\mathrm{ref}}(\varphi_i(u)) e^{-\beta \mathcal{C}(\varphi_i(u))} j_i(u)}{J_q(\varphi_i(u))}, \qquad Z_i = \int b_i(u) \, \mathrm{d}u, \\ \lambda_i = \frac{Z_i}{\sum_j Z_j}, \qquad p_{1,i} = \frac{b_i}{Z_i}. \end{gathered} \tag{E.3}$$

Omit zero-mass charts. For any integrable f, the area formula gives

$$\sum_i \lambda_i \int f(\varphi_i(u)) p_{1,i}(u) \, \mathrm{d}u = \int_{\mathcal{F}_z} f(x) p_z^\star(x) \sum_i \alpha_i(x) \, \mathrm{d}\,\mathrm{vol}_{\mathcal{F}_z}(x) = \mathbb{E}_{P_z^\star} f.$$

Thus $1/J_q$ conditions support volume, j_i converts fiber volume to coordinate volume, and α_i removes overlap multiplicity. Charts may cover different components, but chart identity is not a component label. Correct within-chart laws alone do not ensure correct component masses. Missing components remain a support limitation; an incomplete cover recovers only its covered target. For $d_i = 0$, use counting measure and a discrete chart law.

Importance sampling and rejection. For known joint proposal $\lambda_i^{\mathrm{prop}} r_i(u) > 0$ on target support, assign a raw draw weight $W(i, u) = b_i(u)/(\lambda_i^{\mathrm{prop}} r_i(u))$. Cancellation under proposal expectation

gives the target integral above. With independent draws and finite positive expected weight, the strong law gives consistency of normalized masses and weighted integrals. The same argument applies chartwise to independent stratified proposals. Rejected raw draws must keep zero weight; equal accepted counts do not estimate the same masses.

A subsequent physical acceptance function $a(x)$ changes the law to

$$P_z^a(\,\mathrm{d}x) = \frac{a(x)}{\mathbb{E}_{P_z^\star} a}\, P_z^\star(\,\mathrm{d}x). \tag{E.4}$$

It preserves the normalized target iff a is positive and constant almost surely. A deterministic gate therefore preserves it only when acceptance is almost sure. No reweighting recovers a region with zero acceptance. A zero-total-weight bank is undefined and must be reported as insufficient support. Bounded rejection, shortages, and any best-of-K selection are consequently part of the evaluation protocol.

F Conditional flow matching

Training target. Let stored chart coordinates u_{in} have normalized weights w_{in}. The planar target is

$$p_{1,i}^{\varepsilon}(u) = \sum_n w_{in} \mathcal{N}(u \mid u_{in}, \varepsilon^2 I_{d_i}), \qquad \varepsilon = 0.03. \tag{F.1}$$

It approximates the ideal law in Eq. (E.3). Independent $U_0 \sim \mathcal{N}(0, I_{d_i})$, $U_1 \sim p_{1,i}^{\varepsilon}$, and uniform t define $U_t = (1-t)U_0 + tU_1$. Training minimizes

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\big[\|v_{\theta,i}(U_t, t, \zeta_i) - (U_1 - U_0)\|^2 / d_i\big], \qquad d_i > 0, \tag{F.2}$$

with the map and banks frozen; a separate chart head fits their normalized masses. The spatial implementation instead uses source standard deviation 0.5 and an atomic weighted target (App. L). The following smooth-endpoint result requires $\varepsilon > 0$.

Theorem 2 (Compatible fiber transport)**.** *For a finite bank and $\varepsilon > 0$, the population field $v^\star(u,t) = \mathbb{E}[U_1 - U_0 \mid U_t = u]$ has a smooth representative whose exact ODE transports the Gaussian source to $p_{1,i}^{\varepsilon}$. A valid decoder then gives $X_1 \sim \varphi_{i\#} p_{1,i}^{\varepsilon}$ and $q(X_1) = z^\star$.*

Proof. Conditional expectation minimizes the square loss, and differentiating $\mathbb{E} f(U_t)$ for compactly supported smooth f gives the weak continuity equation with velocity $v^\star$ (Lipman et al., 2023; Albergo et al., 2025). For the finite Gaussian mixture in Eq. (F.1), each component of U_t has variance $s_t^2 = (1-t)^2 + \varepsilon^2 t^2$, which is bounded away from zero for $\varepsilon > 0$. Hence $v^\star$ is smooth with at most linear growth, giving a unique global flow with endpoint law $p_{1,i}^{\varepsilon}$. Decoding gives the stated pushforward, while Theorem 1 gives $q(X_1) = z^\star$. □

This is a specialization of established flow-matching and interpolant constructions, not a guarantee for a fitted finite-step field or an atomic endpoint. Sampling local decoder i with its target mass gives the associated physical mixture. Smoothing and subsequent screening may change the ideal unsmoothed law.

Ambient carrier implementation. For a metric-orthonormal frame V_i, let $a = V_i u$ and $P_i = V_i V_i^\top M_i$. Then $\|P_i v_\theta^{\mathrm{car}}(V_i u_t, t) - (a_1 - a_0)\|_{M_i}^2 = \|V_i^\top M_i P_i v_\theta^{\mathrm{car}}(V_i u_t, t) - (u_1 - u_0)\|^2$. Thus the projected carrier objective equals intrinsic-coordinate regression. The nonlinear fiber trajectory is $\varphi_i(u_t)$, not the affine carrier path. A zero-dimensional chart needs neither an ODE nor the dimension-normalized loss.

Training and inference. Fit and freeze the outcome map; build valid local charts for each training request; compute raw-proposal weights from Eq. (E.3); normalize positive masses; and fit chart probabilities and conditional velocities. For a new request, construct its charts, predict their masses, sample a chart and Gaussian source, integrate its coordinate ODE, decode, and apply the declared numerical/feasibility gate. Conditioning must distinguish the target requests and chart geometry being modeled. Chart setup and numerical screening are part of inference cost.

G PROJECTABLE DYNAMICS AND GROUP EQUIVARIANCE

Theorem 3 (Projectable transport)**.** *Suppose classical fields $v_t, \bar{v}_t$ satisfy $Dq_x v_t(x) = \bar{v}_t(q(x))$ and have unique flows $\Phi_{s,t}, \bar{\Phi}_{s,t}$ on a common regular domain and interval. Then*

$$q \circ \Phi_{s,t} = \bar{\Phi}_{s,t} \circ q, \qquad q_\#(\Phi_{s,t\#}\mu_s) = \bar{\Phi}_{s,t\#}(q_\#\mu_s) \tag{G.1}$$

for initial laws supported on that domain.

Proof. The chain rule gives $\frac{\mathrm{d}}{\mathrm{d}t} q(\Phi_{s,t}(x)) = \bar{v}_t(q(\Phi_{s,t}(x)))$. This quotient ODE starts at $q(x)$, so uniqueness proves the first identity; applying it to a random initial point proves the second. Conversely, differentiating the commuting identity at $t = s$ yields projectability. □

The parameterization Eq. (15) satisfies the condition because $DqL = I$ and $DqP^V = 0$. Every supported projectable field admits this decomposition. These are continuous-time statements; discretization and approximate inverse errors are evaluated separately.

Horizontal motion alone. For $q(x_1, x_2) = x_1$ and Euclidean metric, $v = (x_2, 0)$ is horizontal but has representative-dependent quotient velocity $Dq\, v = x_2$. Its flow $(x_1 + tx_2, x_2)$ sends equal initial quotient values to different ones. The correct horizontal term is the lift of one shared quotient field, $L_x \bar{v}_t(q(x))$.

Corollary 4 (Group-equivariant special case)**.** *Under the theorem's flow assumptions, suppose a smooth free proper isometric group action has orbits equal to the fibers of q. If $w_\theta(g \cdot x, t) = Dg_x w_\theta(x, t)$, then the field in Eq. (15) obeys $\Phi_{s,t}(g \cdot x) = g \cdot \Phi_{s,t}(x)$.*

Proof. Differentiating $q(g \cdot x) = q(x)$ gives $Dq_{g \cdot x} Dg_x = Dq_x$. Thus Dg_x maps vertical spaces onto vertical spaces; isometry also preserves their orthogonal complements. Consequently $P^V_{g \cdot x} Dg_x = Dg_x P^V_x$ and, by uniqueness of the horizontal lift, $L_{g \cdot x} = Dg_x L_x$. The total field is equivariant, so flow uniqueness proves the result. □

This recovers the standard equivariant-flow implication (Köhler et al., 2020). It concerns fixed context; a simultaneous action on scene and configuration requires additional context compatibility. An arbitrary cost-sensitive fiber law need not be group-invariant.

H EXPERIMENTAL SCOPE

Each domain instantiates the same behavior-defined relation with a separately trained model. The arm experiments first fit and freeze the outcome map, then train fixed-target generation or moving-target dynamics separately. Image, language, fixed-target arm, and moving-target arm objectives are therefore evaluated as separate instantiations of the framework rather than a jointly trained model.

I MNIST EXPERIMENT

I.1 BEHAVIOR, RENDERER, AND DATA

For source image I_0 with digit label y, the dimensionless transformation vector is $p = (\theta, t_x, t_y, s, w) \in \mathbb{R}^5$. A fixed image-dependent orthogonal matrix C_{I_0} from image moments defines local coordinates $a = C_{I_0}^\top p$. The renderer $F_{I_0}(a)$ uses bilinear affine warping with angle 120θ degrees, normalized-grid translations $(16t_x/28, 16t_y/28)$, and shear $4s/15$. Thickness blends the warped image with its 3×3 dilation or erosion with weight $\min(|w|, 0.8)$. These are coordinates of prescribed image transformations, not a claim that all image pixels form one regular quotient manifold.

All seven contexts retain digit identity and the additional quantity in Table 6. Set $r_{\mathrm{img}} = (t_x^2 + t_y^2)^{1/2}$ and $g_{\mathrm{img}}(r) = u(r)^2[3 - 2u(r)]$, with $u(r) = \min(1, \max(0, (r - 0.28)/0.12))$. The gate is zero for $r \le 0.28$, one for $r \ge 0.40$, and interpolates smoothly between these regions. The four diagnostic

Context	Additional retained quantity	Regular vertical dimension
Identity	None	5
Orientation	$(\sin \pi\theta, \cos \pi\theta)$	4
Position	(t_x, t_y)	3
Thickness	w	4
Slant	s	4
Radial position	r_{img}	4 for $r_{\mathrm{img}} > 0$
State-dependent orientation	$g_{\mathrm{img}}(r_{\mathrm{img}})\theta$	5 inside, otherwise 4

Table 6: MNIST contexts and local invariant dimensions. All contexts retain digit identity, and "inside" denotes $r_{\mathrm{img}} \leq 0.28$.

offsets are

$$\Delta_{\mathrm{img}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0.18 & 0.20 & -0.12 & 0 & 0 \\ -0.16 & -0.10 & 0.20 & 0.15 & 0 \\ 0.10 & 0 & 0 & -0.18 & 0.22 \end{pmatrix}. \tag{I.1}$$

Each response concatenates a ten-dimensional one-hot label and eight context-masked geometric channels, giving 72 channels over the four diagnostics. Orientation, position, thickness, and slant use the offset transformation. The radial and state-dependent channels instead add prescribed response offsets $\Delta_{\pi,1}$ and $0.05\Delta_{\pi,1}$ to the original radial and gated-orientation responses. They do not recompute the radius after a translated image transformation. This preserves the intended invariant directions of the base response and defines the deterministic signature $B_c^{\mathrm{img}}(I_0, a)$.

A training tuple contains an anchor, two equivalent images, and one distinct image. Equivalent handwriting instances share the source digit class and retained quantities. Within the transition annulus, equivalent samples keep rotation and position fixed. Distinct images use a different digit for identity and the same digit with a different discretized retained target in other contexts. The auxiliary class counts are $(10, 40, 90, 30, 30, 30, 50)$. Rotation boundaries are $(-0.33, 0, 0.33)$, position, slant, and thickness use $(-0.22, 0.22)$, and radius uses $(0.30, 0.52)$. These labels supervise an auxiliary task and do not replace continuous behavioral equivalence.

We partition the 60,000 MNIST training images into 50,000 for training and two separate sets of 5,000 for validation, with seeds fixed at 1729 (LeCun et al., 1998). We use 600 images from each validation set; one selects the numerical rank condition and the other selects the trained model. Equivalent and distinct images come from the same partition as the corresponding input. Evaluation uses the 10,000 MNIST test images across all seven contexts, giving 70,000 input–context pairs; no test images are used for training or validation-only selection. Training and evaluation use the same distribution of transformations, with $(\theta, t_x, t_y, s) \in [-0.75, 0.75]^4$ and $w \in [-0.60, 0.60]$. Both slant and thickness vary in each image. For the context retaining r_{img}, we sample $r_{\mathrm{img}} \in [0.18, 0.75]$. For state-dependent orientation, we sample r_{img} from $[0.04, 0.22]$, $[0.30, 0.38]$, or $[0.44, 0.72]$ with probabilities 0.4, 0.2, and 0.4, respectively.

I.2 Recorded encoder and objectives

The encoder $E_c(I) \in \mathbb{R}^{64}$ has four 3×3 convolutional layers with widths $(32, 64, 96, 128)$, batch normalization, GELU, pooling after the second layer, and 4×4 spatial pooling. A linear projection maps the 2048 spatial features into 128 dimensions, with $\mathrm{LayerNorm}$ and GELU. A 16-dimensional context embedding and two-layer head produce the representation. The original models match approximately 755k parameters by adjusting head widths; SFA adds generative parameters. Inference uses image and context only.

For $\bar{E}_c = E_c / \max(\|E_c\|, 10^{-8})$, centered finite differences of step 0.04 on 24 anchors give the encoder differential $J_E = D_a \bar{E}_c(F_{I_0}(a))$. Response differences of step 0.05 supply a detached projector P_V from right singular vectors below absolute threshold 0.10. Analytic response derivatives are used for evaluation and the ground-truth control. Write $P_H = I_5 - P_V$, $d = \operatorname{tr} P_V$, and $r = 5 - d$. For an orthonormal horizontal frame H with $HH^\top = P_H$, let $\sigma_j(A) = \sqrt{\lambda_j(A^\top A)}$.

The losses used by the MNIST experiment runs are

$$\mathcal{L}_V^{\text{img}} = \mathbb{E}\left[\frac{\|J_E P_V\|_F^2}{d} \,\middle|\, d > 0\right], \qquad \mathcal{L}_H^{\text{img}} = \mathbb{E}\left[\frac{1}{r}\sum_{j=1}^{r}[0.7 - \sigma_j(J_E H)]_+^2 \,\middle|\, r > 0\right]. \tag{I.2}$$

The horizontal objective uses all r directions. These objectives do not enforce exact equivalence throughout the space (App. C).

A soft projector head predicts $\widehat{P} = U\,\mathrm{diag}(m)U^\top$ using a QR-orthonormal frame and sigmoid weights. Its losses are $\mathcal{L}_P = \mathbb{E}\|\widehat{P} - P_V\|_F^2/5$ and $\mathcal{L}_d = \mathbb{E}(\mathbf{1}^\top m - d)^2$. A finite-perturbation term $\mathcal{L}_f$ matches normalized encodings after a length-0.12 step along the normalized projection of $(0.41, -0.53, 0.31, 0.67, -0.23)$, skipping a zero projection. Such a straight tangent step is a finite approximation, not an exact curved-fiber motion.

A paired-feature head predicts $T_{ij}(a) = R_{ij}a + t_{ij}$ with a Cayley-parameterized orthogonal R_{ij}. Define $D_T(T, T') = \|R - R'\|_F^2/25 + \|t - t'\|^2/5$. The transition losses are

$$\begin{aligned} \mathcal{L}_{\text{id}} &= \mathbb{E}D_T(T_{ii}, \text{id}), & \mathcal{L}_{\text{inv}} &= \mathbb{E}D_T(T_{ji} \circ T_{ij}, \text{id}), \\ \mathcal{L}_{\text{comp}} &= \mathbb{E}D_T(T_{ik}, T_{jk} \circ T_{ij}), & \mathcal{L}_{\text{coord}} &= \mathbb{E}\|T_{ij}(a_i) - a_j\|^2/5, \\ \mathcal{L}_{\text{transport}} &= \mathbb{E}\|R_{ij}P_{V,i}R_{ij}^\top - P_{V,j}\|_F^2/25. \end{aligned} \tag{I.3}$$

These losses use known transformation coordinates and detached projectors on equivalent same-context tuples; they encourage consistency but do not certify an exact generative atlas.

The complete MNIST experiment objective is

$$\begin{aligned} \mathcal{L}_{\text{img}} = {} & 2\mathcal{L}_{\text{task}} + 0.3\mathcal{L}_{\text{contrast}} + 0.5\mathcal{L}_{\text{response}} + 0.8\mathcal{L}_P + 0.03\mathcal{L}_d + 0.1\mathcal{L}_f \\ & + 5\mathcal{L}_V^{\text{img}} + 12\mathcal{L}_H^{\text{img}} + 0.1(\mathcal{L}_{\text{id}} + \mathcal{L}_{\text{inv}}) + 0.2\mathcal{L}_{\text{comp}} + 0.3(\mathcal{L}_{\text{coord}} + \mathcal{L}_{\text{transport}}). \end{aligned} \tag{I.4}$$

The task loss is context-specific cross-entropy and response regression is mean squared error, averaged across the four tuple members. The within-context contrastive temperature is 0.1, with positive weights proportional to $\exp[-\|B_i^{\text{img}} - B_j^{\text{img}}\|/(0.08\sqrt{72})]$, normalized per anchor.

Objective ablations. We ablate the geometry-specific components of Eq. (I.4) while retaining all other losses and the same training protocol. Projector/dimension supervision removes $\mathcal{L}_P$ and $\mathcal{L}_d$, finite-perturbation invariance removes $\mathcal{L}_f$, and encoder-Jacobian supervision removes $\mathcal{L}_V^{\text{img}}$ and $\mathcal{L}_H^{\text{img}}$. For compactness, we denote the latter pair by $\mathcal{L}_J$ in Tab. 7. Transition supervision, denoted by $\mathcal{L}_T$, removes the identity, inverse, composition, coordinate, and projector-transport losses in Eq. (I.3).

Variant	AUROC ↑	Subspace F1 ↑	Rank (%) ↑
Full model	**0.750**	**0.917**	**82.1**
w/o projector/dimension supervision	0.749	0.812	61.2
w/o finite-perturbation invariance	0.748	0.904	81.8
w/o encoder-Jacobian supervision	0.685	0.610	14.7
w/o transition supervision	0.734	0.881	74.4

Table 7: Component-removal ablation of the MNIST training objective. Each row removes one loss group from the full model while retaining all others. Values are six-seed means.

Removing encoder-Jacobian supervision produces the largest degradation across all three metrics, reducing AUROC from 0.750 to 0.685, subspace F1 from 0.917 to 0.610, and rank accuracy from 82.1% to 14.7%. This supports directly shaping the encoder differential as the main mechanism for recovering behavior-preserving directions and their dimension. Removing projector/dimension supervision also reduces geometric recovery while leaving AUROC nearly unchanged, indicating that these auxiliary objectives primarily support local geometry rather than pairwise equivalence discrimination. Removing transition supervision gives smaller but consistent reductions across the three metrics, supporting its complementary role in aligning local geometry across equivalent realizations. Finite-perturbation invariance has only a small effect on the reported means, suggesting that it acts as an auxiliary consistency term rather than the primary source of geometric recovery.

I.3 BASELINES, OPTIMIZATION, AND METRICS

Every model uses task supervision. Augmentation consistency additionally matches augmented normalized encodings. C4 rotation pooling averages features over four right-angle rotations, while eight-angle rotation pooling uses eight evenly spaced rotations. Rotation–translation pooling averages features over twenty views formed by four rotations and five translations. These finite pooling constructions do not guarantee exact continuous rotation or translation invariance. The C_8 feature model keeps the eight ordered orientation features instead of pooling. STN learns an affine alignment (Jaderberg et al., 2015). These geometric baselines do not receive the behavioral response loss. SupCon uses shared task labels (Khosla et al., 2020), while the behavior-based contrastive baseline uses the same task, response regression, and response-weighted contrastive terms as our base objective.

The fixed subspace baseline uses one projector per context. The ground-truth and learned subspace-head baselines receive analytic and response-derived directions, respectively. These controls use base, projector, dimension, and finite-perturbation terms, without the encoder-Jacobian and transition losses. They are not encoders with exact oracle invariance. These controls compare behavior supervision and separate geometry prediction with the complete objective under shared training protocols. These controls use different heads and objectives, so the comparison tests the complete construction. The non-SFA methods use six seeds, ten epochs, batch size 128, $3{,}900$ updates after dropping incomplete batches, AdamW learning rate 3×10^{-4}, weight decay 10^{-4}, and gradient clipping at 5. The geometry heads use learning rate 2×10^{-3}; both learning rates follow $\eta_e = \eta_0[1 + \cos(\pi e/10)]/2$, $e = 0, \dots, 10$. Shared loss weights, training identities, and held-out samples are matched; checkpoint selection uses validation data only. Equal updates do not imply equal compute for multi-view or finite-difference objectives.

Structured Flow Autoencoder control. The SFA row adapts the continuous Gaussian-latent model of Xu et al. (2026a) to the same image and context inputs. Its clean-image posterior mean uses the benchmark encoder, with additional time and diagonal-variance heads and a two-layer width-512 conditional velocity decoder. The decoder receives the noisy image, latent sample, context, and time. Training retains the released single-sample velocity-regression objective, minimum noise scale 10^{-4}, and latent norm coefficient 0.01. It also receives the shared task, response, contrastive, projector, dimension, and finite-perturbation losses, but not our encoder-differential or transition terms. Thus the comparison tests generative training under shared observable supervision, rather than reproducing unsupervised SFA. Three development pilots select a generative-loss weight of 10 from $\{0.1, 1, 10\}$. Six fresh seeds use ten epochs and the same data and update budget, with additional parameters and pilot compute rather than an equal-compute claim. Evaluation uses the normalized deterministic clean posterior mean and the same metric definitions, with geometric scores evaluated on $1{,}024$ held-out anchors across all seven contexts. The generative decoder is not used to produce the evaluated representation.

Equivalence AUROC scores pairs by negative Euclidean distance between normalized encodings. For geometric evaluation, let $J_E = U \operatorname{diag}(\sigma_1, \dots, \sigma_5) V^\top$, with $\sigma_1 \geq \cdots \geq \sigma_5 \geq 0$ and $V = [v_1, \dots, v_5]$. The selected invariant directions define

$$\widehat{P}_V = \sum_{j=1}^{5} \mathbf{1}[\sigma_j \leq \max(a, b\sigma_1)] v_j v_j^\top .$$

The first validation set selects (a, b) from

$$a \in \{0, 0.01, 0.02, 0.04, 0.08, 0.12, 0.16, 0.24, 0.32, 0.48, 0.64, 0.96, 1.28\},$$
$$b \in \{0.05, 0.10, 0.15, 0.22, 0.32\},$$

using rank accuracy with equal weighting of the true dimensions 3, 4, and 5. Ties use lower a, then lower b. The metrics are

$$F_{\text{enc}} = \frac{2 \operatorname{tr}(\widehat{P}_V P_V)}{\widehat{d} + d}, \qquad R_{\text{enc}} = \mathbb{E}\mathbf{1}[\widehat{d} = d], \qquad d = \operatorname{tr} P_V, \quad \widehat{d} = \operatorname{tr} \widehat{P}_V . \tag{I.5}$$

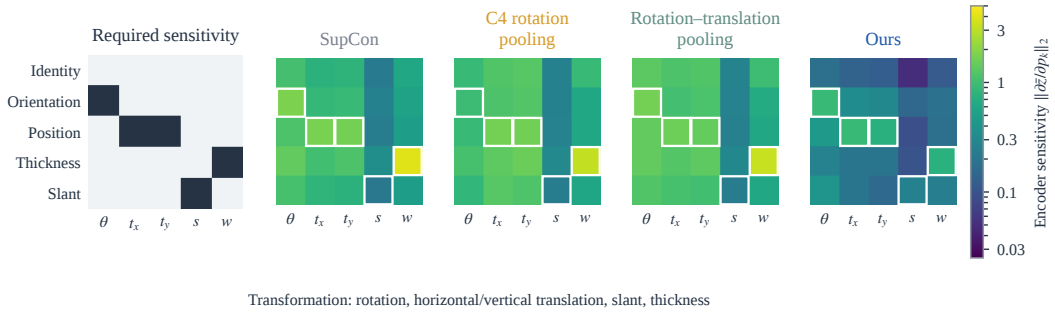


Figure 3: **Task-dependent encoder sensitivity.** Rows are five task contexts and columns are rotation, horizontal/vertical translation, slant, and thickness. Black reference cells and white outlines mark behavior-changing coordinates. Measured cells show Jacobian-column norms of the normalized encoder, averaged over the same 40 held-out MNIST images and six seeds, using finite-difference step 0.04 and one shared logarithmic scale. Ours more selectively suppresses the unmarked coordinates.

The same numerical rank condition applies to all contexts, including identity. For all methods, the other validation set selects the trained model using

$$S = \frac{0.5}{7}\sum_{c=1}^{7}\mathrm{AUROC}_c + \frac{0.25}{3}\sum_{k\in\{3,4,5\}}\Pr(\widehat{d}=d \mid d=k) + \frac{0.25}{7}\sum_{c=1}^{7}R_{\mathrm{enc},c}.$$

All models use the selected numerical rank condition at evaluation, with geometric metrics computed on 1,024 separate test images across all seven contexts. Results are reported as seed means; differences without uncertainty estimates are descriptive.

I.4 Encoder sensitivity and context-dependent organization

We use separately trained models with the same objectives and training protocols as Table 1. The pooling baselines shown in both figures are C4 rotation pooling and rotation–translation pooling. Figure 3 shows that ours suppresses sensitivity to task-irrelevant transformations while retaining sensitivity to the requested quantities. The same rotation is suppressed under identity context and retained under orientation context, illustrating selective rather than uniform invariance. The slant context retains sensitivity to slant with lower sensitivity to translation and thickness, while rotation remains more sensitive.

Figure 4 complements this local measurement with views of the same handwritten sevens across the experiment's rotation range. Under identity context, ours groups each source's rotated views compactly; under orientation context, a clear angle-dependent color gradient emerges across sources. This illustrates context-dependent organization within one digit class, rather than separation between digit classes.

J Language experiment

Data and equivalence labels. Executable questions are derived from CFQ (Keysers et al., 2020). Rewrites change active/passive voice, conjunction order, or wording; their unseen compositions test generalization. Further pairs change negation or quantity. Source-question families and their derivatives are disjoint across splits: 3098/464/210 source questions and 110674/16456/17650 sentences for training/development/test. This is a controlled CFQ-derived task, not the standard MCD benchmark. The four test subsets contain 21,594 labeled pairs and 724 unresolved pairs. Positive certificates identify executable programs after canonical variable renaming and atom ordering, retaining inequalities and negation/quantity guards. An observed answer disagreement certifies a negative. Finite-world agreement alone supplies no positive certificate, and unresolved pairs are excluded from AUROC and FPR_{95}.

Worlds and information access. Each world contains 16 entities; query answers are their binary indicators. Unary facts are sampled independently with probability 0.5, and binary-relation den-

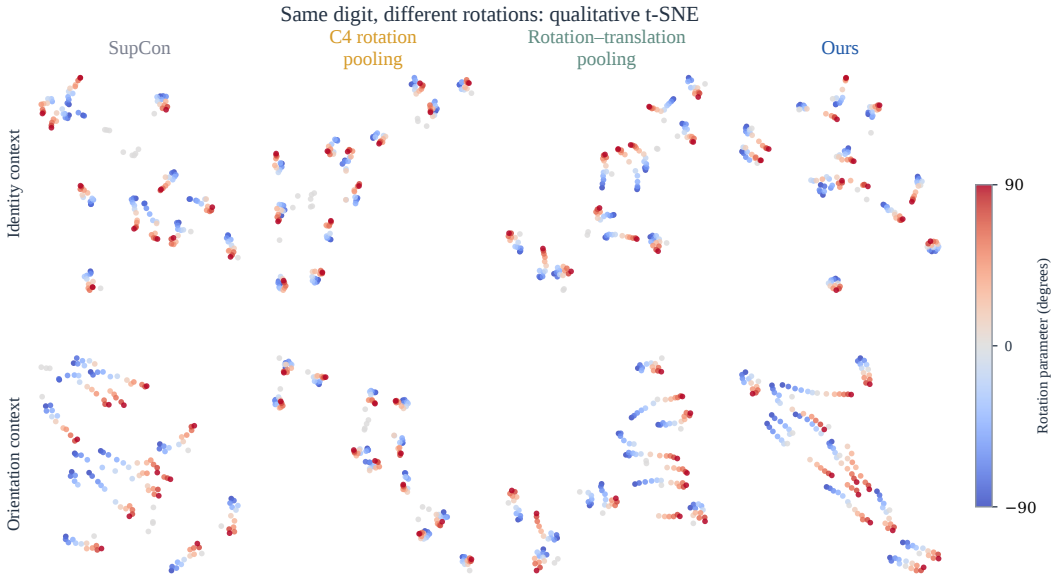


Figure 4: **Identity and orientation contexts in t-SNE.** Each panel uses the same 12 held-out handwritten sevens at 13 rotations from -90° to 90°, matching the experiment's rotation range and colored by angle. Normalized features from checkpoint seed 1 are embedded independently with identical settings (perplexity 20, PCA initialization, random seed 73, no whitening). Ours shows compact source groups under identity context and a directional rotation pattern under orientation context. The plots are qualitative; independently fitted coordinates do not share a distance scale.

sities cycle through $0.05, 0.15, 0.35$. Responses across 32 worlds instantiate $B_c^{\mathcal{W}}$ as 512 answer channels used for behavior supervision. The selected panel is drawn from 352 candidates by a training-query-only greedy distinction criterion with diminishing reward for already distinguished pairs. Final labels use certificates and disagreements on the frozen selected/random panels and 256 fresh worlds generated after model and threshold selection. All methods use the same final pairs; panel selection costs additional executions. Every inference path uses text only, without queries, identifiers, observed test answers, or an execution engine.

BERT model and objectives. BERT-base-uncased (Devlin et al., 2019) mean-pools nonspecial, nonpadding tokens and predicts the 512-dimensional finite behavior signature $\widehat{B}_c^{\mathcal{W}}(x)$ through a width-512 GELU head. Our 128-dimensional quotient code instantiates Eq. (B.2), with the centering vector and orthonormal projection rows obtained from principal components of training behavior vectors only. The complementary reconstruction uses a width-512 GELU residual. This construction grounds the quotient coordinate in predicted task-relevant behavior while preserving the retained coordinates under reconstruction.

We compare two alternative representation pathways. Direct-text controls learn a 128-dimensional code directly from BERT features, while behavior controls first predict the behavior vector and then learn an answer-to-code mapping. All BERT methods receive answer supervision. Our model combines behavior prediction, complementary reconstruction with weight 0.1, and the pair objective. The pair objective minimizes squared distance between normalized codes for certified equivalent pairs, applies a unit-margin squared hinge to witnessed negatives, and for unresolved pairs matches code distance to the distance between normalized centered observed behaviors with weight 0.1. The loss differentiates through the deployed text-to-code pathway. The autoencoder omits the pair objective, while SupCon, weighted contrastive learning, and RNC use their respective representation objectives. RNC (Zha et al., 2023) orders pairs by L^1 behavior distance, retains ties, and trains with negative Euclidean distances between unnormalized codes. Evaluation normalizes the codes of all methods.

Development pilots use 1,500 updates to select answer MSE versus binary cross-entropy and pair weight in $\{0.3, 1\}$. RNC additionally selects temperature in $\{0.1, 0.5, 2\}$. Final models restart from

pretrained weights for seeds $17, 29, 43$ and train for $5{,}000$ updates with batch size 128, backbone and head learning rates 2×10^{-5} and 10^{-4}, weight decay 0.01, and 10% warmup. Validation every 500 steps selects mean AUROC over the original, single-rewrite, and negation/quantity subsets, breaking ties by answer MSE. Backbone, sentence streams, code dimension, and update budgets are shared, while head parameter counts can differ.

Matched pair-loss comparison. To separate the representation pathway from differences in pairwise supervision, we additionally compare models trained with the same pair objective. Direct text learns the code directly from BERT features, while predicted behavior first maps text to the behavior representation and then learns the code. Both controls share our pair objective, answer supervision, sentence streams, 128-dimensional representation, and $5{,}000$ backbone updates. Our model instead applies the fixed projection and complementary reconstruction of Eq. (B.2) to the predicted behavior representation. Ours achieves the highest mean AUROC across all four subsets in Tab. 8, showing that the advantage remains when pairwise supervision and the backbone training budget are matched. This comparison evaluates the complete projection and reconstruction construction and does not separately isolate its individual components.

Representation	Original	Paraphrases	Negation/quantity	Compositions
Direct text with exact pair loss	0.988424	0.997579	0.998974	0.998391
Predicted behavior with learned code	0.987644	0.996028	0.999702	0.997669
Ours	**0.997368**	**0.999787**	**0.999721**	**0.999140**

Table 8: BERT comparison with the same answer supervision and pair loss on 32 selected worlds. Values are three-seed mean AUROC. Higher is better, with the highest means in bold.

Qwen backbone extension. We further test whether the same quotient construction transfers to the larger Qwen3-Embedding-0.6B backbone (Zhang et al., 2025). Qwen uses last-token pooling and one shared equivalence instruction, with all backbone parameters fine-tuned. The InfoNCE baseline uses its native $1{,}024$-dimensional embeddings with symmetric multi-positive pairs while masking self and unresolved pairs. Our model instead predicts the 512-dimensional behavior vector and applies the 128-dimensional projection and complementary reconstruction of Eq. (B.2). Because our model additionally receives behavior-vector supervision, this experiment tests transfer of the construction to a stronger language backbone rather than providing a supervision-matched ablation.

Four $1{,}000$-update pilots per condition select learning rate and temperature for InfoNCE or pair weight for our model. Final models use the same family-separated data, three seeds, $5{,}000$ updates, batch size 128, and validation every 500 updates. A predefined convergence rule extends all conditions together when final mean development improvement exceeds 0.002. Because this experiment uses the same held-out test cohort as the BERT study, we treat it as a backbone extension rather than an independent replication. The results in Tab. 9 show that the behavior-defined quotient construction remains effective with the larger backbone.

	Original		Paraphrases		Negation/quantity		Compositions	
Representation	AUROC ↑	FPR_{95} ↓	AUROC ↑	FPR_{95} ↓	AUROC ↑	FPR_{95} ↓	AUROC ↑	FPR_{95} ↓
InfoNCE	0.9802	0.075	0.9998	**0.000**	0.9971	0.001	0.9730	0.179
Ours	**0.9995**	**0.005**	**0.9999**	**0.000**	**0.9999**	**0.000**	**0.9995**	**0.001**

Table 9: Additional language equivalence results with the fully fine-tuned Qwen3-Embedding-0.6B backbone. Values are three-seed means on a 0-1 scale. Best results are **bolded**, including ties at the displayed precision.

Evaluation metrics. Similarity is negative Euclidean distance between unit-normalized codes. FPR_{95} is the minimum ROC false-positive rate at recall of at least 95%, computed separately for each seed and subset before averaging.

K PLANAR-ARM EXPERIMENT

K.1 TASK, REFERENCE, AND LEARNED GEOMETRY

For $x \in \mathbb{T}^n$, use Euclidean angular metric, $\gamma_j = \sum_{a \le j} x_a$, and

$$p(x) = \sum_{j=1}^{n} \ell_j (\cos \gamma_j, \sin \gamma_j)^\top, \qquad \mathcal{C}(x) = \frac{1}{n} \sum_j [1 - \cos(x_j - x_{\mathrm{rest},j})]. \tag{K.1}$$

Lengths are $(0.45, 0.35, 0.20)$ for three joints and $(0.40, 0.40, 0.12, 0.08)$ for four. Targets lie in the annulus $[0.35, 0.55]$. The obstacle is a disk centered at $(0.24, 0.10)$ with effective radius 0.075, tested against every link segment. The rest vector is $(0.4, -0.8, 0.6, -0.3)$, truncated as needed, and the reference is uniform angular volume. Position fibers have regular dimension $n - 2$. Feasibility restricts the sampling law without redefining endpoint equivalence.

For convenience in the experiments, we use the inverse-temperature parameter $\beta = 1/\tau$, with $\beta = 0$ denoting no cost preference. Independent reference quadrature parameterizes distal absolute angles $\gamma_3, \ldots, \gamma_n$ and both analytic elbow branches, with weights $\mathbf{1}_{\mathcal{A}}(x) e^{-\beta \mathcal{C}(x)} / |\ell_1 \ell_2 \sin x_2|$. This already accounts for the reference-coordinate change of measure. Reference IK, derivatives, and component labels do not train the learned outcome map or construct its generation banks.

Outcome training. The map takes $(\sin x, \cos x)$ through four width-256 SiLU hidden layers to two outputs. The fixed split has $80{,}000/10{,}000/10{,}000$ uniformly sampled configurations. Original fits use $50{,}000$ updates, batch $2{,}048$, AdamW learning rate 3×10^{-4}, weight decay 10^{-6}, cosine decay, gradient clipping at 1, and validation endpoint MSE for selection. Learned geometry uses $q = \widehat{B}$, with no compression, and differentiates this frozen map. Exact geometry uses analytic kinematics.

The refinement warm-starts the map using the same observed triples $x^0 = x$, $x^\pm = \mathrm{wrap}(x \pm hv)$, $y^a = p(x^a)$ for every control, where $h = 0.01$ and one random unit v is fixed per center.

$$\mathcal{L}_{\mathrm{geom}} = \mathbb{E}\left[\frac{1}{6} \sum_{a \in \{0,+,-\}} \|q(x^a) - y^a\|^2 + \frac{\lambda_{\mathrm{dir}}}{2} \left\| Dq_x v - \frac{y^+ - y^-}{2h} \right\|^2 \right]. \tag{K.2}$$

There are $240{,}000$ training observations. Each $\lambda_{\mathrm{dir}} \in \{0, 0.1, 1\}$ receives $15{,}000$ updates at learning rate 10^{-4}, selected by validation endpoint MSE. The target uses observed differences, not analytic Jacobians. A C^3 directional Taylor expansion bounds difference bias by $Kh^2/6$ when the third derivative is bounded by K. Uniform directions satisfy $\mathbb{E}_v \|(Dq - Dp)v\|^2 = \|Dq - Dp\|_F^2 / n$. A finite set of such labels does not certify uniform derivative accuracy.

Task-dependent rank. A separate four-output context-conditioned model fits $B_{\mathrm{pos}} = (p_1, p_2, 0, 0)$, a duplicate position context, $B_{\mathrm{pose}} = (p_1, p_2, \cos \gamma_n, \sin \gamma_n)$, and $B_{\mathrm{gate}} = (p_1, p_2, g(p_1) \cos \gamma_n, g(p_1) \sin \gamma_n)$ with $g(a) = \mathbf{1}[a > 0]$. Position and pose have generic ranks two and three. The gated task changes rank across its two smooth regions. Points with $|p_1| \le 0.02$ are excluded only from smooth differential metrics. At the fixed primary relative singular-value threshold 0.03 (audit thresholds $0.01, 0.03, 0.10$), four-joint rank accuracy is 99.7% for position and 85.6% for the gated task. This outcome-regression study is separate from fixed-position generation and the MNIST objective.

K.2 LOCAL DECODER, TRAINING TARGETS, AND GENERATION

The original split uses $256/32/64$ training/validation/test endpoints per setting, seeds $0, 1, 2$, training preferences $\beta \in \{0, 1, 4\}$, and main evaluation at $\beta = 1$. For each target, project $1{,}024$ random angular configurations using up to 30 minimum-norm Newton steps, tolerance 2×10^{-6}, step cap 0.4, backtracking, and singular-value floor 10^{-5}. Feasible projected anchors initialize a farthest-point cover. SVD of Dq supplies frames. The local decoder uses locally wrapped joint coordinates. Horizontal correction allows 12 iterations, residual tolerance 3×10^{-6} in the original float32 evaluation, and norm limit 0.7. The initial radius 0.9 shrinks by 0.65 on failed shell tests, followed by coverage repair.

Method	3 Joint	3 Joint+obs.	4 Joint
w/o coarea	0.1297 ± 0.0012	0.1738 ± 0.0019	0.0438 ± 0.0003
w/o volume	0.0725 ± 0.0016	0.1053 ± 0.0097	0.0351 ± 0.0012
Ours	0.0560 ± 0.0026	0.0806 ± 0.0112	0.0275 ± 0.0007

Table 10: Probability-correction ablation at $\beta = 1$, with means and sample standard deviations across three seeds. Errors are CDF W_1 for three joints and periodic sliced W_1 for four joints.

Normalize $\widetilde{\alpha}_i(x) = [1 - \| \operatorname{wrap}(x - x_i)\|^2/(0.75R_i)^2]_+^2$ across local decoders where the denominator is positive. Each local decoder receives $N_i = 256$ raw proposals from $r_i = \mathcal{N}(0, 0.6^2 I)$. With acceptance indicator A_{in}, its bank uses

$$W_{in} = \frac{A_{in}\alpha_i(\varphi_i(u_{in}))j_i(u_{in})e^{-\beta\mathcal{C}(\varphi_i(u_{in}))}}{J_q(\varphi_i(u_{in}))r_i(u_{in})}, \quad \widehat{Z}_i = N_i^{-1}\sum_n W_{in}, \quad w_{in} = W_{in}/\sum_m W_{im}. \tag{K.3}$$

Rejected proposals keep zero weight, and zero-mass local decoders are omitted. Finite root, inverse-coordinate, and coverage checks approximate valid support rather than certify complete coverage. The target adds independent coordinate noise with standard deviation 0.03.

The local geometry encoder has width 128 and one four-head attention layer, receiving periodic anchors, vertical projectors, endpoint, $\beta/4$, obstacle flag, and radii. A three-hidden-layer width-256 SiLU network with 32 sinusoidal time features predicts ambient carrier velocity, projected by $V_i V_i^\top$ (App. F). A separately parameterized local geometry encoder and scalar head predict masses, trained by cross-entropy against normalized bank masses. Generators and original mass heads use $30{,}000$ updates, batch 512, AdamW learning rate 2×10^{-4}, and weight decay 10^{-5}. Validation uses each family's loss plus 0.1 local-mass KL on eight fixed validation targets, selecting within each family. Inference constructs the local decoders, predicts masses, integrates a Gaussian source for 32 Heun steps, and decodes. It requires geometry queries and root solves.

Comparison implementations. The main baselines in Tab. 3 receive the target, obstacle, and preference, but no local reference configurations, local probabilities, or implicit decoder during generation. FM uses a standard Gaussian source in joint space, while RFM uses a uniform source in periodic joint angles and shortest wrapped geodesic interpolation. Both fit the same weighted training targets after the original 0.03 local-coordinate smoothing and implicit decoding. Thus the target preparation is shared, even though the local geometric construction is not supplied during generation. The networks use width 256, $30{,}000$ updates, batch 512, learning rate 2×10^{-4}, and validation-only selection. Main-baseline PCFM applies one damped Newton correction per update, with damping 10^{-6} and guided interpolation disabled, followed by the common final projection. FM and RFM use 32 Heun steps, while PCFM uses 64 updates, giving 64 velocity evaluations in each case. Numerical corrections require additional quotient-map and Jacobian evaluations.

K.3 Evaluation and transport ablations

Table 3 requests $2{,}048$ outputs per endpoint at $\beta = 1$ and reports three-seed means over 64 targets. Physical success requires endpoint error below 10^{-3} of total reach, finite output, collision-free links, and numerical validity. The learned residual measures preservation of the selected learned fiber, while physical accuracy is evaluated independently using analytic kinematics. Periodic SW_1 averages 64 one-dimensional Wasserstein distances after embedding $(\sin x, \cos x)$ and is not intrinsic arclength distance. The three-joint CDF W_1 in Table 10 instead uses $\omega = \operatorname{wrap}(\sum_j x_j - \operatorname{atan2}(z_2^\star, z_1^\star))$ on $[-\pi, \pi)$. These metrics have different scales. SW_1 includes all returned configurations, including finite failures, and is undefined if any output is nonfinite. Component TV retains an invalid-return category. Success and distribution error must therefore be read together.

The separate probability ablation in Tab. 10 reports means and sample standard deviations across three seeds. Removing coarea or decoder-induced volume changes the distribution despite using the same implicit constraint. Its scores and seed deviations describe this separate ablation, not the main comparison in Tab. 3.

Velocity model	$e_{\mathrm{struct}} \downarrow$	$e_{\mathrm{pred}} \downarrow$	$e_{\mathrm{phys}} \downarrow$
Ambient	3.21e−3	4.14e−5	3.29e−3
Horizontal only	3.79e−3	4.14e−5	3.86e−3
Soft constraint	2.72e−3	4.14e−5	2.80e−3
Ours (projectable)	8.37e−9	4.14e−5	4.28e−5
Equality proj.	8.37e−9	4.14e−5	4.28e−5
Proj. (early)	1.24e−8	2.57e−3	2.57e−3

Table 11: Four-joint moving-target dynamics with learned geometry. Three-seed mean squared errors. Equality projection is algebraically identical to ours, and early uses an undertrained quotient field.

K.4 Moving-target dynamics

A separate supervised velocity study freezes the outcome map and uses physical time $s \in [0,1]$. For endpoint pair (z_0, z_1), curvature κ, and unit normal $n_\perp$, define

$$a(s) = 3s^2 - 2s^3, \quad z_{\mathrm{ref}} = z_0 + a(s)(z_1 - z_0) + \kappa \sin(\pi a(s)) n_\perp, \quad b^* = \dot{z}_{\mathrm{ref}} + 2(z_{\mathrm{ref}} - z). \tag{K.4}$$

Short straight, long straight, and long curved families use $256/32/64$ train/validation/test path pairs per arm, with seeds $0, 1, 2$. Endpoint angular separation is 0.2 radians for short paths and 1 radian for long paths. Curved paths use $|\kappa| = 0.02$. A three-layer width-64 SiLU model $b_\chi(z, s, z_0, z_1, \kappa)$ fits b^* on $65{,}536$ examples for $8{,}000$ updates, batch 512, selected by validation MSE. The early control freezes its checkpoint after 250 updates and keeps the same structural lift. A separate three-layer width-128 SiLU joint field f_ν receives $(\sin x, \cos x, s, \alpha/4, b_\chi)$. Here $\alpha \in \{0, 1, 2, 4\}$ weights $w_j = -\sin(x_j - x_{\mathrm{rest},j}) + 0.25\cos(2\pi s + 0.7j)$. During training $\widetilde{b} = b_\chi + 0.1\epsilon$, $\epsilon \sim \mathcal{N}(0, I_2)$, and the target is $L_x \widetilde{b} + \alpha P_x^V w$. Analytic IK supplies $131{,}072$ training joint states shared across controls, while operators use the stated frozen learned or exact map. Both networks use AdamW with learning rate 3×10^{-4} and weight decay 10^{-5}. The joint field additionally uses clipping at 5 and cosine decay to 0.1 of the initial rate.

Ambient and soft models receive $12{,}000$ updates minimizing $\mathbb{E}[\|f_\nu - (L_x \widetilde{b} + \alpha P_x^V w)\|^2/n + \lambda_{\mathrm{dyn}} \|Dq_x f_\nu - \widetilde{b}\|^2/2]$. Select $\lambda_{\mathrm{dyn}} \in \{0, 1, 10, 100\}$ by validation velocity plus projectability MSE, with zero denoting Ambient. Inference uses unperturbed b_χ, comparing direct, horizontal-only, soft, and structural fields $L_x b_\chi + P_x^V f_\nu$. Equality-constrained projection is algebraically identical to the structural field by Eq. (C.1).

Each test path starts from 32 representatives and uses 64 Heun steps, with time averages using trapezoidal quadrature. With $\bar{z}_\chi(s)$ independently integrating the quotient ODE, the main metrics are

$$\begin{aligned} e_{\mathrm{struct}} &= \mathbb{E}_{s,j} \|q(x_j(s)) - \bar{z}_\chi(s)\|^2, \quad e_{\mathrm{pred}} = \mathbb{E}_s \|\bar{z}_\chi(s) - z_{\mathrm{ref}}(s)\|^2, \\ e_{\mathrm{phys}} &= \mathbb{E}_{s,j} \|p(x_j(s)) - z_{\mathrm{ref}}(s)\|^2. \end{aligned} \tag{K.5}$$

Structural consistency, requested-path prediction, and physical tracking are different properties. In Tab. 11, projectable dynamics reduces structural error from 2.72×10^{-3} for the soft constraint to 8.37×10^{-9}, and physical error from 2.80×10^{-3} to 4.28×10^{-5}, with the same quotient-prediction error. Horizontal projection alone removes directions without fixing the induced quotient velocity. Equality projection agrees with ours by Eq. (C.1), while the early quotient field retains low structural error but gives higher physical error. These tests do not establish a time-varying Gibbs marginal or collision-free execution of the generated static configurations.

L Spatial-arm experiment

Task and geometry. The Panda benchmark generates static seven-joint configurations for full end-effector pose, with one-dimensional regular fibers. Joint coordinates are normalized by midpoint and half-range without wrapping. The reference is uniform volume in this bounded joint box, restricted by joint limits and the declared collision model. Posture cost is mean squared normalized displacement from the midpoint. Targets are forward poses of feasible configurations sampled uniformly in $[-0.99, 0.99]^7$. The free scene has no external obstacle. Clutter cycles through one axis-aligned box

with center/half-size $(0.45, 0, 0.35)/(0.065, 0.12, 0.10)$ or $(0.38, \pm 0.28, 0.40)/(0.09, 0.06, 0.13)$, in meters. Bullet checks the Panda collision meshes with the model's self-collision exclusions, fingers fixed at 0.04 m, and the hand TCP. Penetration beyond 10^{-8} m counts as collision. Pose residuals concatenate position in meters and rotation-log coordinates in radians with unit relative scale. Exact geometry uses analytic kinematics, while learned geometry differentiates a frozen pose-valued network. The latter has four width-512 SiLU layers, takes normalized joints and sine/cosine angles, and predicts position plus two orthonormalized rotation columns. The structural rank bound of six is not learned rank discovery.

Observed triples supply $600{,}000$ outcomes from $200{,}000$ training centers, with no analytic Jacobian labels. Symmetric unit-direction perturbations use step 0.01 in normalized joints, reduced near box boundaries. The map uses $120{,}000$ updates, batch $4{,}096$, AdamW learning rate 3×10^{-4}, weight decay 10^{-6}, gradient clipping at 5, and a cosine schedule ending at 0.1 of the initial rate. Position/rotation-entry weights are $100/1$ for outcome and directional-difference losses, with the latter weighted by 1. Weighted validation outcome MSE selects the checkpoint. The learned sampler never uses true pose accuracy as an acceptance criterion.

Local geometry and generator protocol. Each scene has $256/32/64$ nominal train/validation/test requests, with seeds $0, 1, 2$. Within each seed, comparisons share the available training banks. Evaluation covers 64 validation and 128 test requests across the two scenes, retaining the requested sample counts in the validity denominator. Training preferences are $0, 8, 32$. The original evaluation additionally uses $1, 4, 16$, while Table 4 reports $\beta = 1$.

Local decoder construction uses projected starts, local exploration and continuation, inverse-coordinate support checks, and at most $1{,}024$ local decoders. Dense training banks use 256 proposals per local decoder. Deployment uses 16 probes per local decoder for mass, coordinate-moment, cost, and acceptance features, alongside anchors, projectors, radii, target pose, and obstacle geometry. A width-128, four-head attention encoder predicts local context. The scalar velocity network has three width-256 SiLU layers and 32 sinusoidal time features. Learned logits correct probe-based log masses. Our source standard deviation is 0.5, with 16 Heun steps and no target smoothing. Thus the positive-bandwidth theorem in App. F is not a literal theorem for this atomic endpoint.

For the exact-geometry comparison in Tab. 12, the local-geometry baselines use the same map, banks, request features, and local probability-model architecture with separately fitted weights. They receive $20{,}000$ updates, batch 512, and common validation selection. The objective adds 0.1 mass cross-entropy to velocity regression. AdamW uses learning rate 3×10^{-4}, weight decay 10^{-5}, and cosine decay to 0.1 of the initial rate. Selection every $1{,}000$ updates uses 256 deployed samples on four validation targets per scene, scoring joint SW_1 against the validation bank plus shortage fraction. FM corrects its endpoint by final projection, while Tangent FM projects velocities and retracts both Heun stages. PCFM adapts root-mode constraint correction (Utkarsh et al., 2025). Adapted RFM uses arc-length interpolation under the induced metric of the one-dimensional fiber (Chen & Lipman, 2024). It predicts joint velocities with tangent projection and intermediate correction, rather than our scalar local-coordinate velocity. Disconnected solution families use the shared local probability model, not geodesic transport between families.

Main comparison. The baselines in Tab. 4 receive the target pose, obstacle, and preference, but no local reference configurations, local probabilities, or implicit decoder during generation. Training uses the same weighted joint configurations, update budget, and validation requests. FM uses a global standard Gaussian source, while RFM uses a uniform source in periodic joint angles with shortest wrapped geodesic interpolation. For RFM, angular coordinates are physical joint angles relative to the fixed joint-limit midpoint, not normalized joint coordinates interpreted as angles. The selected target, joint limits, and collisions remain separate constraints. Final projection and PCFM use the frozen quotient map and its Jacobian, with no physical target accuracy supplied during generation. PCFM uses one damped Newton correction per update, with damping 10^{-6} and guided interpolation disabled, followed by the common terminal correction. FM and RFM use 16 Heun steps, while PCFM uses 32 updates, giving 32 velocity evaluations per proposal. The correction steps require additional geometric computation, so this does not imply equal total cost.

Evaluation. Each target/preference requests $2{,}048$ configurations with a gate capped at eight times that many proposals. Pose SR counts returned configurations within 1 mm and 1° against the re-

Method	Exact geometry		
	Valid (%) ↑	SW_1 ↓	Component TV ↓
FM + final projection	**100.000**	0.01621	0.02942
Tangent FM	**100.000**	0.01616	0.02933
PCFM (adapted)	**100.000**	0.01602	0.03017
RFM (adapted)	**100.000**	0.01623	0.02945
IKFlow + projection	99.696	0.11668	0.22438
GGIK + projection	99.346	0.11441	0.22373
Ours	**100.000**	**0.01505**	**0.02844**

Table 12: **Fixed-target generation on the 7-DoF Panda arm with exact geometry.** Three-seed means across 128 held-out target/scene requests and six posture preferences. Validity counts feasible outputs within 1 mm and 1° against the requested total. SW_1 and component TV measure distribution and disconnected-component probability errors. IKFlow and GGIK are evaluated only with exact geometry. Bold marks column-best means.

quested total, with missing outputs counted as failures. SW_1 compares screened outputs in full normalized joint coordinates against a reference distribution obtained by adaptive quadrature over analytic IK branches. Unlike the planar score, it has no periodic embedding. Screening checks learned-target, joint-limit, numerical, and collision validity, while true pose accuracy is evaluated afterward. Pose SR therefore measures physical pose accuracy after these checks, rather than independently evaluating collision-free execution. A nonfinite return or no accepted output makes its distribution score undefined. Component TV is half the absolute generated/reference mass difference over disconnected reference families, assigning outputs by nearest reference configuration within normalized distance 0.05 and retaining unassigned mass in a zero-reference category. The quadrature parameterizes analytic IK branches by joint seven and uses adaptive feasibility intervals and coarea weights. Refined quadrature and independent projected starts audit its coverage without certifying global completeness. The comparison does not establish a runtime advantage or collision-free execution trajectories.

L.1 Exact-geometry comparison

Table 12 removes learned-map error by supplying analytic kinematics under the original six-preference evaluation protocol. IKFlow's conditional invertible model and GGIK's graph variational autoencoder are adapted to the same preference, scene, and training distribution (Ames et al., 2022; Limoyo et al., 2025), with projected validation selection. GGIK additionally receives robot-point supervision, so these external controls are reported only with exact geometry. They need no local decoder construction during generation. The common terminal solver permits 60 iterations, bounded updates, and the same numerical/collision checks. Preserving the learned fiber does not by itself ensure physical target accuracy. This comparison removes disagreement between the learned map and analytic kinematics, while retaining the question of probability assignment among equivalent configurations.

L.2 Implicit correction and linear displacement

The matched study fixes the learned map, 128 requests, 32 anchors per request, both tangent signs, nine displacement fractions, and three seeds. At 70% radius, implicit correction reduces physical position error from 4.430 to 0.638 mm, with 95% crossed seed/request bootstrap intervals $[4.051, 4.846]$ and $[0.614, 0.663]$. Orientation error decreases from 0.709° to 0.257° and acceptance rises from 60.53% to 91.46%. This is the largest tested displacement below the fixed 75% support boundary. Acceptance uses the supplied map, branch/support checks, joint limits, and collisions, not true pose error. Valid implicit outputs meet the 10^{-9} learned-residual threshold, with finite failed solves retained in residual means. The vertical space preserves the quotient coordinate only to first order. The improvement under finite displacement therefore supports the horizontal correction for fiber curvature, rather than treating a tangent direction as a valid finite motion.

L.3 Probability corrections

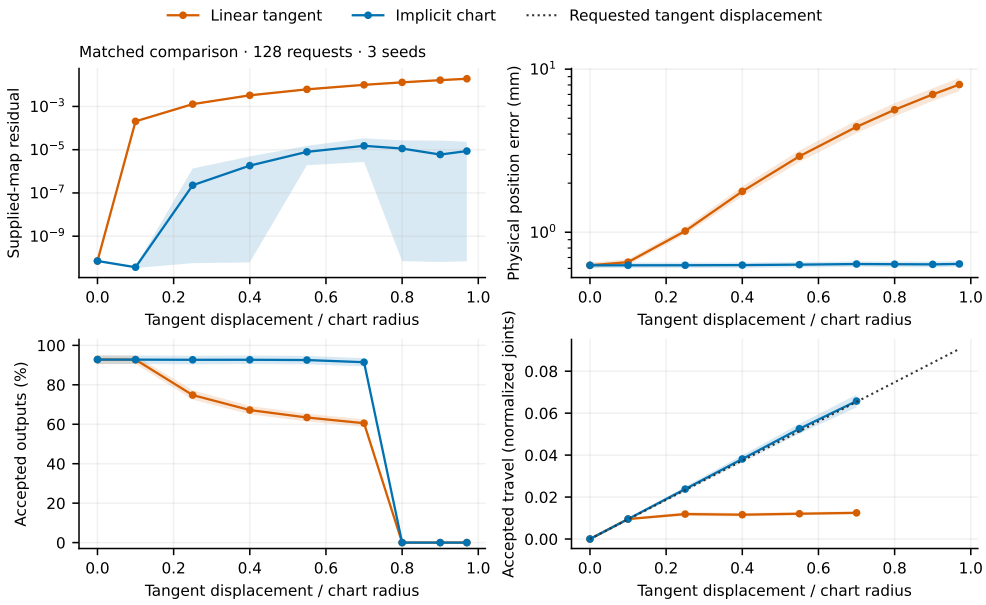


Figure 5: Linear displacement and implicit correction under identical learned geometry. Shading gives 95% seed/request bootstrap intervals. All tested displacements are retained. Above the support boundary both methods reject and accepted travel is undefined.

Variant	Valid (%) ↑	SW_1 ↓	Component TV ↓
w/o coarea	100	0.04424	0.08051
w/o volume	100	0.09961	0.19085
Ours	100	**0.01505**	**0.02844**

Table 13: Probability corrections on Panda with exact geometry. Three-seed means show different representative distributions despite equal validity.

Table 13 retrains probabilities and transport with only coarea or volume removed, retaining the exact geometry, local decoders, requests, and other corrections. This ablation uses exact geometry, separately from the learned-geometry comparison in Tab. 4. Every variant returns full counts at 100% validity. Full-minus-ablation SW_1 differences are $-0.02919/-0.08456$ for no-coarea/no-volume, with 95% crossed seed/request bootstrap intervals $[-0.03354, -0.02526]/[-0.08940, -0.08016]$, keeping preferences within request blocks. Differences favor full in every seed and at each preference. Mean missed reference mass is 0.0074%. On shared components containing 99.59% of reference probability and at least 16 samples per method/component, weighted SW_1 is $0.00958/0.02721/0.05786$ for full/no-coarea/no-volume. Thus the corrections affect probability assignment within solution families as well as across disconnected families. Equal validity does not determine the representative distribution.

L.4 Flow matching with changing quotient coordinates

Task and shared information. The Panda follows requested hold, straight, curved, and loop paths, with either endpoint position or full pose defining the quotient coordinate. The regular fibers have dimensions four and one, respectively. The learned outcome map is frozen, and numerical continuation constructs time-dependent anchors and local frames along each requested path. All methods receive the same path, posture preference, local geometry, reference anchor and its derivative, and initial representatives. Training and evaluation paths are separate, and the target distribution favors one or two preferred terminal configurations within the same local decoder region. Each task uses $256/32/64$ training/validation/test paths and three flow seeds, $0, 1, 2$, with the learned outcome map fixed at seed 0. This study tests transport within the valid moving decoder region, whereas the fixed-target experiment tests probability assignment across disconnected components.

Moving decoder and training targets. Write $\varphi_{i,s}$ for the implicit decoder at normalized physical time s, with $q_c(\varphi_{i,s}(u)) = z_s$. The initial coordinate U_0 follows a bounded Gaussian source, and U_1 is independently drawn from the terminal weighted training bank. Specifically, $U_0 = \xi/\sqrt{1 + \|\xi\|^2/2.5^2}$ with $\xi \sim \mathcal{N}(0, 0.5^2 I)$. For the uniform ambient reference and terminal

posture cost, bank weights use the decoder-induced volume and coarea factors of Eq. (12), restricted to the valid local region. Define

$$\begin{aligned} U_s &= (1-s)U_0 + sU_1, \qquad X_s = \varphi_{i,s}(U_s), \\ W_s &= \partial_s \varphi_{i,s}(U_s) + D_u\varphi_{i,s}(U_s)(U_1 - U_0), \\ \mathcal{L}_{\text{moving}} &= \mathbb{E}\left[\frac{1}{d_i}\left\|v_{\theta,i}(U_s, s, \zeta_i) - (U_1 - U_0)\right\|^2\right]. \end{aligned} \tag{L.1}$$

The expectation is over time, path context, and the paired endpoints, with d_i the fiber dimension. Inputs and velocity targets use normalization fitted on training data, with one scalar target scale per model. Here ζ_i contains the shared path, posture, and local geometry, including the current reference anchor and its derivative. The coordinate velocity is integrated and decoded at each time, yielding a complete joint trajectory rather than a single terminal configuration. Differentiating the implicit constraint gives

$$Dq_{c,X_s}\partial_s\varphi_{i,s}(U_s) = \dot{z}_s, \qquad Dq_{c,X_s}D_u\varphi_{i,s}(U_s) = 0. \tag{L.2}$$

Thus coordinate transport cannot change the induced quotient velocity. The decoder motion need not be horizontal, but satisfies $\partial_s\varphi_{i,s} = L_{X_s}\dot{z}_s + P^V_{X_s}\partial_s\varphi_{i,s}$, giving the quotient-changing and quotient-preserving structure of Eq. (15). These identities hold in valid regular regions, with numerical residual, rank, and support checks during integration. The terminal distribution is evaluated separately from path preservation, without imposing a cost-sensitive marginal at every intermediate time.

Comparison implementations. Ambient FM fits the physical target W_s in joint space, predicting a residual with respect to the shared moving reference. FM with step retraction uses the same ambient training updates, with its checkpoint selected under the numerical corrections used at inference, and corrects both integration stages into the valid local decoder region on the selected fiber. Constrained FM fits $L_x\dot{z}_s + P^V_x w_\theta$ to W_s and integrates its coordinate velocity through the same implicit decoder as ours. These are controlled velocity models with the same information. Adapted RFM (Chen & Lipman, 2024) uses the metric on the initial fiber for both position and pose motion. For pose,

$$g_{i,0}(u) = \frac{\|D_u\varphi_{i,0}(u)\|^2}{R_i^2}, \qquad \ell_i(u) = \int_0^u \sqrt{g_{i,0}(a)}\, da, \tag{L.3}$$

where R_i is the local radius. It fits linear interpolation in ℓ_i between the same paired endpoints, then applies ℓ_i^{-1} and the shared moving decoder. For position, write $a = u/\sqrt{1+\|u\|^2}$, $u(a) = a/\sqrt{1-\|a\|^2}$, and $F_i(a) = \varphi_{i,0}(u(a))$. The induced metric and conditional velocity are

$$\begin{aligned} g_i(a) &= \frac{D_aF_i(a)^\top D_aF_i(a)}{R_i^2}, \\ v_{\text{RFM}}(a, s \mid a_1) &= -\frac{\|r\|^2 g_i(a)^{-1}r}{(1-s)r^\top g_i(a)^{-1}r}, \qquad r = a - a_1, \end{aligned} \tag{L.4}$$

where a_1 is the terminal coordinate. The training objective is $\mathbb{E}[\|v_\theta - v_{\text{RFM}}\|^2_{g_i}/d_i]$, with $d_i = 4$ and the same target normalization. Conditional paths are computed numerically, and $D_a u$ maps predicted velocities to u coordinates for integration with the shared moving decoder. This construction uses coordinate distance $\|a - a_1\|$ rather than geodesic distance. In both settings, the metric is fixed on the initial fiber, while the decoder follows the changing target. The velocity models share network width, training budgets, and validation-only selection, with different interpolation coordinates and the same endpoint pairs. They use three width-256 hidden layers and 20,000 AdamW updates, batch 512, learning rate 3×10^{-4}, weight decay 10^{-5}, gradient clipping at 10, and cosine decay to 0.1 of the initial rate. Validation every 2,000 updates uses 64 draws per path, selecting mean SW_1 plus implicit decoder reconstruction error, both divided by local radius, plus ten times the numerical failure fraction. Initialization is also a candidate. Every method receives the same reference information, with no additional ground-truth joint trajectory supplied to a baseline.

Evaluation and execution. All methods use 128 Heun steps and the same 128 initial draws for each of the 64 test paths per seed, giving 24,576 requested numerical rollouts per task and method across three seeds. Track requires a complete, numerically valid path with physical position error

	Position tracking		Pose tracking	
Variant	SR (%) ↑	SW_1 ↓	SR (%) ↑	SW_1 ↓
Ours	**98.96**	**0.01139**	**97.92**	**0.00254**
w/o learned transport	97.92	0.01446	**97.92**	0.01739

Table 14: **Ablation of coordinate transport.** Fixing the local coordinate retains the moving implicit decoder. Three-seed means with the metrics of Tab. 5.

below 5 mm at every evaluated time and, for pose, orientation error below 1°. Failed requests and incomplete paths count as failures. SW_1 compares completed terminal configurations in normalized joint coordinates against an independently computed reference distribution with the same target weighting. It measures the terminal distribution before controller execution. SR measures successful simulated execution of the generated joint trajectory in PyBullet under a shared joint-velocity controller, with the same tracking tolerances, joint limits, and self-collision checks. Execution uses the first 16 test paths and first two initial draws per seed, fixed before evaluation, giving 96 episodes per task and method. Invalid plans count as failures. The controller runs at 250 Hz with 1 kHz physics, using the planned joint velocity plus joint-position error with gain 20, the robot's velocity and effort limits, and no reset after initialization. Analytic kinematics evaluate physical accuracy independently of the frozen learned map. SR therefore includes failures on the fixed execution subset, while SW_1 is conditional on completed numerical rollouts over all test paths. Their denominators differ.

The constraint $Dq_{c,x}v = \dot{z}_s$ fixes quotient motion but leaves motion within each fiber, so preserving the path does not determine the terminal representative distribution. The comparison in Tab. 5 therefore concerns distribution matching and execution, not only whether the requested quotient path is enforced. Adapted RFM also uses the moving decoder, so its lower execution success does not imply a failure of the projectability identity.

Role of learned transport. The control without learned transport fixes $U_s = U_0$ but retains $X_s = \varphi_{i,s}(U_0)$, so the moving decoder still follows the requested quotient path. In Tab. 14, learned transport reduces terminal SW_1 by 21.2% and 85.4% for position and pose, respectively. Without learned transport, SR is 97.92% in both settings, while ours achieves 98.96% for position and 97.92% for pose. These results separate high execution success from accurate probability assignment among equivalent realizations.