SHARED-GENERATOR NEURAL FLOWS FOR ZERO-SUPPORT TRANSITION PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generalizing learned state-transition mechanisms to unobserved condition pairs is a fundamental problem in scientific machine learning, dynamical modeling, and perturbation prediction. Yet learning how a state changes under observed conditions does not determine how to reverse that change or extend it to a larger shift. We study zero-support transition prediction: given a source state and the source and target conditions, the task is to predict the resulting state without examples of the requested transition. Mechanism-Preserving Transport (MPT) ties these predictions together through shared state-dependent vector fields: reversing the condition difference runs the exact flow backward, while doubling it follows the same dynamics for twice as long. To establish when these shared fields also admit a unique interpretation, we characterize the activation patterns that identify their directions up to scaling and permutation, given sufficiently rich grouped short-time observations in fixed coordinates. Our prediction experiments show that the two constraints need not help together. On PDEBench Burgers, MPT reduces direct prediction MSE by 37.8% relative to a zero-preserving conditional ODE over ten seeds. On Shapes3D, keeping reversal but relaxing linear scaling reduces latent MSE by 59.5% relative to MPT. On Norman Perturb-seq, the more flexible zero-preserving model remains more accurate after mean-matching retraining. These controlled comparisons show that preserving an inverse relation can be useful even when extending a change by linear control scaling is not. Reversal and scaling should therefore be chosen separately for unseen-transition prediction, using endpoint accuracy to determine which structure the task supports.

1 INTRODUCTION

Suppose measurements show how a state changes from setting 0 to 1 and from 1 to 2. Can a model predict the change from 0 directly to 2, or from 1 back to 0, without examples of either transition? Figure 1(a) illustrates the missing direct transition. This question arises when predicting unmeasured gene combinations from perturbation screens (Roohani et al., 2024) or the response to a new parameter shift in a physical system. In this *zero-support transition prediction* setting, the source state and both conditions are known, but examples of the requested transition are unavailable. The model may have encountered both conditions separately.

Two predictors can fit the observed changes equally well and disagree on the missing transitions, an instance of underspecification (D'Amour et al., 2022). Consider a predictor trained only on increases in a setting. A negative condition difference specifies a decrease, but does not require the prediction to undo the corresponding increase. A larger difference does not require agreement with repeated smaller changes either. A flexible conditional model may learn these relationships, but fitting the observed transitions alone does not impose them. We ask whether explicitly constraining these relationships improves predictions on unobserved transitions.

Mechanism-Preserving Transport (MPT) learns local changes that are shared across conditions (Figure 1(c)). Following the neural ODE formulation (Chen et al., 2018), it integrates these changes to predict a finite transition. The condition difference weights a set of shared vector fields, called generators. Reversing the difference reverses the vector field; doubling it follows the same dynamics for twice as long. For the exact flow, the reverse prediction then undoes the forward prediction, and

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

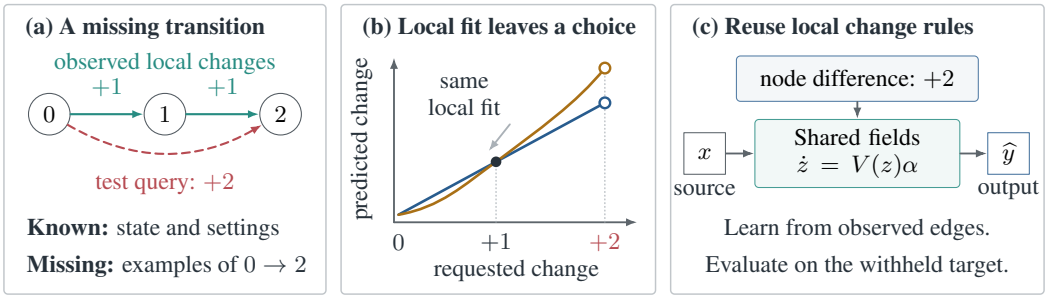

Figure 1: Missing-edge prediction. (a) Adjacent changes are observed; the direct query has known settings but no examples. (b) Schematic predictors agree on a local change and disagree on its extension. (c) MPT reuses local rules under the requested control. We score its endpoint against the target and predictions composed along observed routes. Panel (b) is illustrative, not experimental data.

two equal changes agree with one doubled change. The endpoint can still vary nonlinearly with the condition difference because the fields depend on the current state.

Reversal specifies how to undo a change. Linear control scaling makes a separate assumption: a larger change follows the same dynamics for longer. Neither relation follows from knowing the source and target conditions alone. We test their value separately, giving all predictors the same condition differences and no samples of the requested transition.

The preferred constraints differ across tasks. Shared-generator flows improve direct Burgers predictions. On Shapes3D, retaining reversal while relaxing linear scaling gives more accurate unseen-transition predictions. On Norman Perturb-seq, enforcing only zero-change identity performs better after matching the training objective to the mean-response evaluation. These results support choosing reversal and scaling separately, with their value assessed by prediction error on heldout targets.

Can the shared directions also be recovered uniquely from observations? As in dictionary learning (Cohen & Gillis, 2019), different sets of components can explain the same data. We study an observation design that can distinguish them. Groups of sufficiently rich short-time changes reveal which directions act together, and intersections between the resulting subspaces can isolate individual directions. In fixed coordinates, our theorem characterizes the patterns that identify directions up to scaling and permutation, as well as the mixtures that remain indistinguishable when identification fails. Synthetic tests examine these conditions; the external benchmarks evaluate prediction.

Our contributions are:

- **Separating constraints on unseen transitions.** Shared-generator flows enforce reversal and same-direction composition. Our controls relax these constraints separately while keeping query information fixed, so the comparisons test specific relations between predictions.
- **Evidence for task-dependent choices.** Among the tested models, Burgers favors linear control, Shapes3D favors reversal without linear scaling, and Norman favors the more flexible zero-preserving model. Endpoint and route comparisons distinguish predictive gains from internal agreement.
- **Conditions for interpreting shared directions.** Theorem 1 characterizes the ambiguity left by grouped short-time observations in fixed coordinates. Its noisy and finite-duration extensions quantify recovery error under their stated assumptions.

2 Related Work

Neural ordinary differential equations (ODEs) learn state-dependent dynamics (Chen et al., 2018). FE-NODE estimates coefficients from task observations using a shared basis (Ingebrand et al., 2024, Sections 3.1–3.2). Neural Context Flows (NCF) fits a task context and also describes parameter-induced zero-shot adaptation via inverse regression (Nzoyem et al., 2025, Appendix A.2). Our

metadata adapters use training edges only; adaptation with query-edge examples is reported separately (Appendix B.10).

TransE models relations as translations between entity embeddings (Bordes et al., 2013); learned Lie-group models represent transformations through group actions (Cohen & Welling, 2014; Zhu et al., 2021; Chau et al., 2023). Equation (1) is a control-affine neural ODE, equivalently a neural controlled differential equation (CDE) driven by a linear control path (Kidger et al., 2020).

TRS-ODEN regularizes time-reversal symmetry under a state-reversing operator (Huh et al., 2020). Deep-OSG constrains evolution-map composition across time intervals (Chen & Wu, 2023). We instead test reversal and scaling of condition differences separately, using missing transitions with the same query information.

Whether shared fields can be interpreted uniquely is a different question. Disentanglement results require assumptions on the model or observations (Locatello et al., 2019; 2020); causal representation results use intervention, mixing, or sparsity conditions for latent recovery (Lachapelle et al., 2022; Xu et al., 2024; Bing et al., 2024). We fix the state coordinates and study the subspaces revealed by grouped local-time data. Dictionary learning uses incidence and subspace intersections to recover atoms (Cohen & Gillis, 2019; Sitharam et al., 2014; Novikov & White, 2023). We characterize all basis changes preserving the grouped subspaces, including nonseparating designs.

For physical fields, we use PDEBench's Burgers data and a Fourier neural operator (FNO) baseline (Takamoto et al., 2022; Li et al., 2021). GEARS provides a gene-space reference on the Norman combinatorial perturbation screen (Norman et al., 2019; Roohani et al., 2024). These tasks provide different observations, which we specify next.

3 METHOD

3.1 OBSERVATIONS AND QUERIES

Conditions are nodes of a directed graph $G = (\mathcal{E}, \mathcal{P}_{\mathrm{train}})$, with $\mathcal{P}_{\mathrm{train}} \subseteq \mathcal{E} \times \mathcal{E}$. States lie in $D \subseteq \mathbb{R}^d$, and condition e has code $\beta_e \in \mathbb{R}^K$: d is the state dimension and K the number of control coordinates and shared fields. For images and PDEs, a training edge $e \to e'$ supplies matched states (x, x'). Their shared context is an underlying instance: a base factor assignment for images or an initial field for PDEs. A test query supplies x and both codes; the output is a predicted target state. No examples or coefficients fitted on that edge are available.

For $0 \to 2$, the observed route $0 \to 1 \to 2$ supplies a second prediction by composing edge maps; the reverse query $1 \to 0$ has no such route in a forward-only graph. We compare direct and routed predictions for accuracy and cost.

Norman supplies cell populations under each condition, without before-and-after measurements of the same cells. We transport control cells under a perturbation code and compare their decoded mean with the heldout condition mean. Random-pair and mean-matching objectives use these unpaired samples differently; the prediction target in both cases is the population mean.

Condition metadata determine the codes in the external benchmarks. In synthetic experiments that learn codes from the training graph, one code is fixed to remove the common translation ambiguity. State representations remain fixed during transport fitting.

3.2 SHARED DYNAMICS AND TRAINING

MPT uses $\alpha = \beta_{e'} - \beta_e$ to weight shared fields $v_i : D \to \mathbb{R}^d$, with parameters θ and matrix $V_\theta(x) = [v_1(x), \ldots, v_K(x)] \in \mathbb{R}^{d \times K}$. They are fitted jointly across training edges. We define the query map by the autonomous ODE

$$\dot{z}(t) = \sum_{i=1}^{K} \alpha_i v_i(z(t)), \qquad z(0) = x, \qquad T_\alpha(x) = z(1). \tag{1}$$

Here $t \in [0, 1]$ is transport time, not the physical observation time in the PDE data. We write $\widehat{T}_\alpha$ for the fixed-step fourth-order Runge–Kutta (RK4) prediction. For locally Lipschitz fields and all

$a, b \in \mathbb{R}$ for which the required trajectories stay in D, the exact flows consequently satisfy

$$T_0 = \mathrm{Id}, \qquad T_{-\alpha} = T_\alpha^{-1}, \qquad T_{b\alpha} \circ T_{a\alpha} = T_{(a+b)\alpha}. \tag{2}$$

For equally spaced codes at settings 0, 1, and 2, these identities require $1 \to 0$ to undo $0 \to 1$ and $0 \to 2$ to agree with $0 \to 1 \to 2$. They follow from the architecture, not endpoint fitting (Proposition 2). Distinct controls need not commute; state dependence permits nonlinear endpoints (Appendix A.1). RK4 approximates the identities.

For paired image and PDE training, let $\widehat{P}^{\mathrm{pair}}_{e,e'}$ be the uniform empirical distribution of matched training states on edge $e \to e'$. We minimize the empirical endpoint loss

$$\mathcal{L}_{\mathrm{fit}} = \frac{1}{|\mathcal{P}_{\mathrm{train}}|} \sum_{(e,e') \in \mathcal{P}_{\mathrm{train}}} \mathbb{E}_{(x,x') \sim \widehat{P}^{\mathrm{pair}}_{e,e'}} \left[\frac{1}{d} \left\| \widehat{T}_{\beta_{e'} - \beta_e}(x) - x' \right\|_2^2 \right]. \tag{3}$$

Each update samples an edge uniformly and matched contexts with replacement. The factor $1/d$ matches the implemented coordinate-mean loss. No path or preservation penalty is needed to impose the exact-flow relations; we test those penalties in Appendix B.6 and Section 4.5. Norman uses the unpaired objectives in Appendix B.13.

3.3 Controls for reversal and scaling

The controls separate reversal from scaling by removing constraints from MPT (Table 1). Starting from a conditional neural ODE (Chen et al., 2018), Zero-NODE requires zero change at zero control. Autonomous-zero removes the integration-time input: its rule depends only on state and control. Autonomous-odd adds reversal while allowing nonlinear control scaling. MPT adds linear scaling.

The comparisons test specific choices. Autonomous-zero versus autonomous-odd tests adding reversal; autonomous-odd versus MPT tests adding linear scaling. All receive the same condition difference. These ODE controls are defined here. A shared conditional predictor imposes none of the flow identities; SharedLowRank makes the displacement linear in the control without integrating a field. Appendix B.2 gives their prediction rules.

Table 1: Flow controls defined in this paper. "Yes" denotes an architectural constraint; "–" means it is not imposed. The identities in Eq. (2) concern exact flows where the trajectories exist. Same-direction composition does not imply commutation of distinct controls.

Model	No explicit time input	Zero-change identity	Sign reversal	Same-direction composition
Zero-NODE	–	Yes	–	–
Autonomous zero	Yes	Yes	–	–
Autonomous odd	Yes	Yes	Yes	–
MPT	Yes	Yes	Yes	Yes

The closest published methods test alternative ways to share dynamics. FE-NODE (Ingebrand et al., 2024) and NCF (Nzoyem et al., 2025) obtain query coefficients or contexts through adapters fitted on training edges. FE-NODE combines independently integrated basis displacements; MPT combines fields before integration. Appendix B.10 specifies the adaptations and separate support-based comparisons. All conditional ODE controls fit endpoints, without a density or log-Jacobian objective.

3.4 Evaluation criteria

The relations in Table 1 concern predictions, not their accuracy. Endpoint error is the primary criterion. Leakage measures unintended coordinate change; path error compares direct and composed predictions. The identity map makes both zero while missing every nonzero intervention. Target error, movement, and joint target-and-preservation success distinguish this trivial solution from useful transport. Routed endpoint error also tests whether composing observed maps reaches the target. Appendix B.3 defines these metrics.

4 EXPERIMENTS

We test whether the constraints in Table 1 reduce prediction error on unseen transitions. Burgers tests larger forward changes, advection adds reversals, and Shapes3D tests reversal without linear scaling. Norman evaluates population means rather than paired states. Route and loss comparisons examine alternative explanations (Appendix Table 11).

4.1 PROTOCOL

Data. Burgers uses PDEBench terminal fields (Takamoto et al., 2022), paired by initial condition across five viscosities. Periodic advection is our generated control, with exact Fourier translations at five speeds and a fixed duration. Shapes3D (Burgess & Kim, 2018) uses 16-dimensional frozen autoencoder states from 64×64 RGB images; 17 nodes vary orientation, scale, object hue, and wall hue. Norman Perturb-seq (Norman et al., 2019) supplies unpaired expression samples under gene perturbations. Codes are normalized PDE parameters, image-factor shifts, or multi-hot gene interventions. Splits and preprocessing appear in Appendix B.1; heldout transition targets enter evaluation only.

Baselines. FE-NODE (Ingebrand et al., 2024) and NCF (Nzoyem et al., 2025) use our training-edge adapters, denoted $+\alpha$. The PDE operator control adapts FNO (Li et al., 2021); the gene-space reference uses GEARS (Roohani et al., 2024). SharedLowRank, shared conditional predictors, Identity, and additive singles are elementary controls specified here, alongside the ODE controls in Table 1. Shared conditional predictors use dense MLPs for latent states and periodic CNNs for PDE fields. Appendix Table 12 records the implementation and origin of each name.

Metrics. The primary endpoints use mean squared error (MSE): Shapes3D latent MSE, PDE direct field MSE after training-set standardization, and Norman DE20 MSE on condition means. Paired-task errors average over coordinates and contexts before queries, with equal query weights. DE20 uses the twenty largest absolute target-minus-control gene changes. Decoded pixel MSE, non-target leakage, route disagreement, all-gene MSE, and error on the least-changing half of genes are secondary. Appendix B.3 gives the formulas. The prespecified Norman objective comparison measures the change in the MPT–Zero-NODE DE20 gap.

Selection and uncertainty. Image and PDE comparisons use ten common seeds; Burgers autonomy uses three prespecified seeds. Seeds vary context splits and fitting, including the Shapes3D encoder. Reported dispersion is across seed means, not edges or cells. A 10/10 paired direction gives two-sided sign-test $p = 0.001953$. Norman uses five prespecified condition splits and three training seeds per split; uncertainty is over split means. Each Shapes3D candidate receives 12,000 updates. Frozen encoders achieve test reconstruction MSE 4.75×10^{-4} and probe factor accuracy 0.99992.

PDE models receive at most 12,000 updates, with parameter counts within 3.4% of MPT. Observed-edge validation selects checkpoints; validation chords, untrained edges used only for selection, choose among three learning rates. Test edges are separate. PDE evaluation uses all test contexts and 64 RK4 steps, with 16/32/64-step solver checks. Appendix B records task-specific selection rules and optimization differences.

4.2 PDE FIELDS: DIRECT SCALE AND SIGN EXTRAPOLATION

Burgers tests a larger viscosity change from observed local shifts. Every test query has an observed route, allowing us to distinguish direct extrapolation from the accumulated error of learned local maps (Table 2). MPT has lower direct error than each trainable comparator in all ten seeds ($p = 0.001953$ per comparison). Its direct and routed outputs differ by 3.08×10^{-10}, largely imposed by scalar autonomous control. The evidence for useful prediction is its 2.36% relative endpoint error and displacement of 0.961 times the true displacement. Identity has the same route agreement but fifteen times the endpoint MSE. Appendix B.11 reports solver-refinement checks.

Table 2: Field MSE after training-set standardization, mean ± standard deviation over seeds 300–309. Burgers direct and advection all-heldout direct are primary; routed and reverse-only errors are secondary. Burgers values are in 10^{-4}; advection values are unscaled. Baseline sources and adaptations are specified in Section 4 and Appendix Table 12. Lower is better.

	Burgers		Advection	
Model	Direct	Routed	All direct	Reverse direct
MPT	$\mathbf{1.162 \pm .082}$	$1.162 \pm .082$	$\mathbf{(3.216 \pm .224) \times 10^{-3}}$	$\mathbf{(2.097 \pm .150) \times 10^{-3}}$
Zero-NODE	$1.869 \pm .669$	$\mathbf{1.073 \pm .099}$	2.880 ± 1.970	3.384 ± 2.659
FE-NODE $+\alpha$	10.01 ± 3.21	$1.206 \pm .087$	$12.506 \pm .277$	$12.724 \pm .282$
NCF $+\alpha$	$1.884 \pm .411$	$1.255 \pm .099$	$1.957 \pm .022$	$1.999 \pm .020$
SharedLowRank	16.56 ± 2.04	$1.348 \pm .181$	$12.530 \pm .286$	$12.747 \pm .291$
Shared CNN	6.501 ± 2.259	$1.307 \pm .083$	$1.882 \pm .071$	$1.859 \pm .141$
Conditional ODE	2.155 ± 1.248	$1.209 \pm .134$	$1.911 \pm .023$	$1.948 \pm .022$
FNO	$4.456 \pm .943$	1.466 ± 1.499	$1.896 \pm .046$	$1.888 \pm .083$
Identity	$17.66 \pm .27$	$17.66 \pm .27$	$1.867 \pm .025$	$1.816 \pm .027$

Zero-NODE is the closest direct comparator by mean error; MPT lowers its MSE by 37.8%. Along observed routes, Zero-NODE instead reaches 1.073×10^{-4}, below MPT's 1.162×10^{-4}. Its movement ratio is 0.938, ruling out a near-identity explanation for the routed result.

On Burgers seeds 300–302, the reused MPT fits have 34.2% and 28.0% lower direct MSE than autonomous-zero and autonomous-odd, with three paired wins against each (Table 29). Both controls fit observed edges slightly better and have competitive routed errors. The gap concerns extension to a larger control.

At eight RK4 steps on the separate cost seeds 300/304/309, MPT direct inference takes 8.01 ms per 128 fields versus 23.92 ms along the route, at nearly equal endpoint MSE. Zero-NODE's route is more accurate; SharedLowRank's route is much faster (Table 32).

Controlled advection adds reversal: all observed edges are forward, with code 0.25, so reverse queries have no observed directed route. Zero-NODE fits the observed edges to validation MSE 4.23×10^{-5}, but its direct MSE is 1.619 on larger positive shifts and 3.384 on reversals; its positive routed error is only 0.00801. Fitting one positive code and anchoring zero change leave other codes unconstrained. MPT links them through linear autonomous control and has the lowest reverse MSE in every seed. FE-NODE has the lowest positive routed error, 4.455×10^{-4}. A separate stability test exposes a limitation: the learned MPT field amplifies a heldout forward high-frequency perturbation by 11.23 on average, whereas exact translation preserves Fourier magnitudes. Appendix B reports directional and solver checks.

4.3 SHAPES3D: REVERSAL WITHOUT LINEAR SCALING

Does linear scaling help beyond reversal in learned image coordinates? The key comparison is MPT versus autonomous-odd, which retains reversal but allows nonlinear control. All ten frozen autoencoders pass the reconstruction and probe criteria; Table 3 scores the same 30 edges. MPT has lower latent MSE than SharedLowRank and Zero-NODE in all ten seeds, despite their lower observed-edge error. The odd model improves further, lowering MPT's latent MSE by 59.5% and image MSE by 51.4%, both with ten paired wins. The primary MPT-minus-odd difference is 1.094×10^{-3}, with paired bootstrap 95% interval $[0.812, 1.371] \times 10^{-3}$.

Odd control improves all four direction–composition strata on average (Appendix B.9). Target-factor error falls from 0.01779 to 0.00940, and joint target-and-preservation success rises from 71.6% to 86.1%. Mean latent movement is similar, 0.1000 versus 0.0973, ruling out a near-identity explanation. Leakage means favor odd control, but their paired interval includes zero.

Zero-NODE predicts positive transitions accurately but has large reverse errors. The odd model retains the inverse relation and gives the lowest image endpoint error. This ranking favors reversal without linear scaling under the tested procedures. The two autonomous controls share initialization and sampling, but the reused MPT fits use a fixed learning rate and a different clipping threshold; their comparison also includes optimization differences.

Table 3: Shapes3D missing direct edges, mean ± sample standard deviation over seeds 100–109. Latent MSE is primary; image MSE and factor leakage are secondary. Each seed averages the same 30 edges. Autonomous controls are new fits; the other rows reuse frozen fits. Lower is better. Bold marks the lowest mean, not a significance test.

Model	Latent MSE (10^{-3})	Image MSE (10^{-3})	Leakage (10^{-5})
MPT fit only	1.839 ± 0.308	2.099 ± 0.167	5.285 ± 8.872
Autonomous odd	$\mathbf{0.745 \pm 0.313}$	$\mathbf{1.020 \pm 0.171}$	$\mathbf{2.987 \pm 4.325}$
Autonomous zero	4.120 ± 1.503	3.332 ± 0.590	84.83 ± 50.31
SharedLowRank	4.551 ± 0.303	4.966 ± 0.468	22.49 ± 12.80
Zero-NODE	5.874 ± 1.473	4.169 ± 0.707	175.45 ± 119.39

Task-matched FE-NODE and NCF reach no-support latent MSE 4.689×10^{-3} and 1.708×10^{-2}; test-edge adaptation improves both (Appendix B.10). Figure 3 shows the predeclared decoded examples.

4.4 Route agreement and endpoint accuracy

For queries with an observed route, does direct prediction improve on composing its edge maps? On long Shapes3D routes, direct and composed MPT predictions differ by 2.16×10^{-5}, versus 5.74×10^{-4} for SharedLowRank. SharedLowRank nevertheless has lower routed endpoint error: 2.195×10^{-3} versus 2.504×10^{-3}, with wins in 9/10 seeds. Path regularization does not close the gap. Agreement between predictions leaves their shared error unresolved; Theorem 6 separates these terms. A cost study restricts comparison to the 14 of 30 queries reachable through directed training edges. On this forward subset, Zero-NODE's direct predictor is more accurate and faster than MPT (Appendix B.16). Table 3 also includes the remaining reverse queries.

4.5 Ablations and alternative explanations

The preceding comparisons use endpoint training. Adding a path penalty gives slightly higher heldout MSE on Shapes3D, so the observed gain does not require that regularizer. In the nonlinear synthetic system, fit-only MPT has lower training, heldout, and path error than SharedConditionalMLP at the same update budget; the identity control has zero path error but heldout MSE 1.161. Earlier dSprites comparisons also match models on predictive fit before comparing leakage (Appendix B.6).

4.6 Norman Perturb-seq: consistency without a response advantage

Norman compares the fully constrained flow with Zero-NODE, which enforces only zero-change identity, on population means without paired states. Five condition splits each contain 24 observed singles and 32/11/11 training/validation/test doubles. A training-only truncated singular value decomposition (SVD) retains about 30% of variance in 2,048 selected genes using 48 coordinates. Random-pair MSE penalizes output variance as well as mean error, which could favor a predictor that collapses cell variation. Exact mean matching over the 2,048 control cells removes this penalty (Appendix B.13). Both protocols select on validation condition means; Table 4 compares their response and preservation errors.

Table 4: Norman heldout double perturbations, mean ± standard deviation over five split-level means, each averaging three training seeds. Random-pair rows reuse historical fits; mean matching uses new fits. DE20 is primary; gene MSE and non-DE error are secondary. All scores use the same gene space. Bold compares models within each objective.

Model / objective	DE20 MSE	Gene MSE	Non-DE error
MPT / random pair	$0.09389 \pm .01664$	$0.003816 \pm .000367$	$0.000344 \pm .000071$
Zero-NODE / random pair	$\mathbf{0.06867 \pm .00992}$	$\mathbf{0.002782 \pm .000432}$	$\mathbf{0.000272 \pm .000077}$
MPT / mean matching	$0.07559 \pm .02690$	$0.003871 \pm .001231$	$0.000386 \pm .000103$
Zero-NODE / mean matching	$\mathbf{0.06119 \pm .01486}$	$\mathbf{0.003035 \pm .000720}$	$\mathbf{0.000311 \pm .000090}$

Mean matching lowers the average DE20 error of both models. Zero-NODE remains more accurate in all five splits. The MPT-minus-Zero-NODE difference is 0.01440, with split-bootstrap 95% interval $[0.00526, 0.02527]$. The prespecified objective-by-model interaction is -0.01082, with interval $[-0.03080, 0.00686]$, so the estimated narrowing of the gap is uncertain. Zero-NODE also has lower gene MSE and non-DE error; its response gain does not come at the expense of preservation.

Mean matching also increases predicted coordinate variance, from 0.344 to 0.689 for MPT and from 0.00977 to 0.677 for Zero-NODE, near the source value 0.679. Zero-NODE remains more accurate, which rules out variance collapse alone as an explanation for its earlier advantage. Under random-pair training, direct–sequential disagreement is lower for MPT: 0.01128 versus 0.15969, at movement ratios 1.028 and 1.011. Additive singles has essentially zero disagreement but mean DE20 0.11310.

One gene-space GEARS run per split gives native DE20 0.10898 ± 0.03196. Appendix B.12 reports the decoder-matched scores and the different ordering in the fixed-split study. The five splits measure sensitivity within one screen, not biological replication; a unanimous split direction gives exact two-sided sign-test $p = 0.0625$.

5 When Can the Shared Directions Be Identified?

The experiments assess the combined action. Interpreting its shared directions requires observations that distinguish alternative bases. Endpoint accuracy alone does not: a change of basis can leave the action unchanged, and fixed-duration endpoints need not determine the field. The following result uses grouped short-time observations in fixed coordinates. Synthetic tests examine these premises separately from the external prediction comparisons.

Each direction is a whole field, a rule for local change across states, not an arrow at one state. Consider three independent fields activated in pairs $\{1, 2\}$, $\{1, 3\}$, and $\{2, 3\}$. Rich short-time mixtures reveal each pair's plane. The first two planes intersect in direction 1; the other intersections recover 2 and 3 without pure interventions. If direction 1 never appears without 2, these directions can mix without changing the observed subspaces. The theorem characterizes this distinction for any support family.

There are three spaces: states $x \in D \subseteq \mathbb{R}^d$, coefficients $a \in \mathbb{R}^K$, and the real vector space $\mathcal{F}(D)$ of fields $D \to \mathbb{R}^d$. The map $Va = \sum_i a_i v_i$ produces a whole field; $V_\theta(x)$ collects field values at one state. Write $[n] = \{1, \dots, n\}$ and, for $S \subseteq [K]$, $E_S = \{a \in \mathbb{R}^K : a_j = 0 \text{ for } j \notin S\}$. There are $R \geq 1$ observation groups, with active index sets $S_r \subseteq [K]$, $K \geq 1$. If group r's coefficients span E_{S_r}, its initial velocities reveal $W_r = V(E_{S_r})$. These exact subspaces are assumed known; Appendix A.3 treats estimation from probes. Here $GL(K)$ denotes invertible real $K \times K$ matrices.

Theorem 1 (Residual gauge and support-incidence identification)**.** *Let $V : \mathbb{R}^K \to \mathcal{F}(D)$ be injective and linear, let $\mathcal{S} = (S_r)_{r=1}^R$ satisfy $S_r \subseteq [K]$ and $\bigcup_r S_r = [K]$, and set $W_r = V(E_{S_r})$. Define $\mathcal{I}_i = \{r : i \in S_r\}$, $C_i = \{j : \mathcal{I}_i \subseteq \mathcal{I}_j\}$, and $\mathcal{D}_K = \{\mathrm{diag}(d_1, \dots, d_K) : d_i \neq 0\}$. Then*

$$\begin{aligned} \mathcal{G}_{\mathcal{S}} &:= \{B \in GL(K) : B(E_{S_r}) = E_{S_r} \ \forall r\} \\ &= \{B \in GL(K) : B_{ji} = 0 \text{ if } \mathcal{I}_i \not\subseteq \mathcal{I}_j\}. \end{aligned} \tag{4}$$

The residual group satisfies

$$\mathcal{G}_{\mathcal{S}} = \mathcal{D}_K \quad \Longleftrightarrow \quad C_i = \{i\} \text{ for every } i. \tag{5}$$

Under these equivalent conditions, every injective linear $\widetilde{V} : \mathbb{R}^K \to \mathcal{F}(D)$ with $\widetilde{V}(E_{S_r}) = W_r$ for all r satisfies $\widetilde{V} = V\Lambda$ for some $\Lambda \in \mathcal{D}_K$, and

$$\mathrm{span}(v_i) = \bigcap_{r \in \mathcal{I}_i} W_r, \qquad i \in [K].$$

Remark (What the observation design determines). Equation (4) identifies every allowed mixture. If no group contains i without j, it permits nonzero B_{ji}: direction j can be added to direction i without changing any observed subspace. Repeating these groups cannot remove this ambiguity; adding a group containing i but excluding j does. Equation (5) requires such a contrast for every ordered pair. The three pair groups above satisfy this condition without pure interventions. Thus the result specifies which observation designs identify the lines and which additional groups remove the remaining mixtures (Appendix A.3).

Coverage implies that any competing dictionary is VB for an invertible B. A group containing i but excluding j forces $B_{ji} = 0$; invertibility turns subspace containment into equality. Injectivity transfers intersections to the observed fields. The tools are established: structural matrix algebras (Akkurt et al., 2005) and dictionary recovery through subspace intersections (Novikov & White, 2023). Here they characterize the full ambiguity for a prescribed observation design, including designs that do not identify the directions. Appendix A.3 gives the proof, prior-work comparison, and perturbation bounds.

5.1 Testing the identification premises

The synthetic tests distinguish support separation, excitation rank, and finite-duration error. Separating, rank-two groups recover all 80 generator lines across 20 seeds. Shortening duration from 0.2 to 0.025 lowers mean line angle from $0.1069°$ to $0.01293°$. Removing separation leaves a shear that changes one line by sine 0.7051 while moving every observed projector by at most 2.11×10^{-15}: the ambiguity permitted by Eq. (4) persists numerically. Hiding support masks tests whether intersections also recover group membership. All lines and memberships are recovered under the separating, full-rank design; line recall falls to 0.75 in the rank-deficient control (Appendix A.8).

Noisy-endpoint sweeps test the bias, noise, and excitation terms in the recovery bound (Appendix A.9). More samples improve spectral recovery, while neural field error levels off (Appendix A.10); better subspace estimates alone do not ensure better neural recovery.

6 Limitations

The empirical conclusions depend on the chosen representations and tasks. Shapes3D is controlled and factorized, and its frozen autoencoder states nearly determine the evaluation factors. Norman uses a 48-dimensional projection retaining about 30% of selected-gene variance. Its five overlapping test sets cover 35 combinations within one screen, not transfer to another cell line. Decoder matching measures output smoothing, while gene-space and latent-space training remain different statistical problems.

Optimization limits the attribution of the empirical differences. The FE-NODE and NCF adapter study tests nearby rates through warm-started continuations. Autonomous controls select the largest tested rate on Shapes3D and Burgers, usually at the final checkpoint; the earlier Shapes3D MPT runs use a different rate-selection and clipping protocol. Norman mean matching costs more per update and is compared with saved random-pair fits. Larger budgets or fully matched optimization may change the rankings. Cost measurements cover four earlier methods, not autonomous-odd.

Identification assumes grouped short-time observations in fixed coordinates: every direction is activated, and coefficients span each group's support. The finite-sample extension also assumes Gaussian excitation and endpoint noise. External endpoint data do not supply these premises. Node potentials exclude nonzero code circulation, and deterministic autonomous flows cannot represent a many-to-one or stochastic transition law exactly. Burgers tests terminal shifts rather than forecasting. The neural certificate requires bounds on field approximation, probe coverage, code error, stability, and numerical integration; 42.9% of audited settings give a vacuous bound.

7 Conclusion

A known condition difference specifies a query, but not how a predictor should extend the observed changes. Shared-generator flows impose reversal and same-direction composition; the controls separate their value. Under the tested procedures, linear control gives lower direct Burgers error, reversal without linear scaling gives better Shapes3D endpoints, and Zero-NODE gives better Norman mean responses. The identification theorem establishes when additional grouped short-time observations also determine the shared directions. The practical lesson is to choose relations between predictions by their heldout endpoint accuracy, rather than by internal agreement alone. How to make that choice reliably from a new task's training data remains open.

AI USE STATEMENT

Generative AI tools assisted with manuscript organization, drafting and editing, literature discovery, and the development and review of experimental code. We also used these tools to discuss theoretical claims, counterexample constructions, and proof steps, and to audit assumptions and derivations. For the experiments, assistance included designing diagnostic and sensitivity checks, preparing data-generation and evaluation scripts, and interpreting results. Synthetic observations were sampled from the structural models specified in the paper using numerical code. The authors independently checked and revised AI-assisted text and suggestions, verified references against the cited sources, examined the mathematical arguments, and tested the code by rerunning the relevant computations. The authors take full responsibility for the mathematical claims, experiments, code, and final text.

REPRODUCIBILITY STATEMENT

Appendix A provides the assumptions and proofs; Appendix B specifies the data splits, model selection, and evaluation protocols. Appendix C records the verification scope. The accompanying anonymous code archive includes an English README, environment requirements, experiment configurations, tests, and runners for the reported studies. It also provides data-preparation instructions and compact reference results; raw third-party datasets and large historical checkpoints are not bundled.

ETHICS STATEMENT

The single-cell study reanalyzes the publicly released Norman perturbation screen through the GEARS-preprocessed data (Norman et al., 2019; Roohani et al., 2024); it involves no new participant recruitment or biological experiments. We evaluate condition-level expression responses, not patient outcomes or clinical decisions. These predictions do not establish individual-cell counterfactuals or the safety or efficacy of an intervention. The code archive identifies the data sources and applicable upstream terms without redistributing the raw single-cell dataset.

REFERENCES

Mustafa Akkurt, George Phillip Barker, and Marcel Wild. Structural matrix algebras and their lattices of invariant subspaces. *Linear Algebra and its Applications*, 394:25–38, 2005. doi: 10.1016/j.laa.2004.05.009. URL https://doi.org/10.1016/j.laa.2004.05.009.

Simon Bing, Urmi Ninad, Jonas Wahl, and Jakob Runge. Identifying linearly-mixed causal representations from multi-node interventions. In *Proceedings of the Third Conference on Causal Learning and Reasoning*, volume 236 of *Proceedings of Machine Learning Research*, pp. 843–867, 2024. URL https://proceedings.mlr.press/v236/bing24a.html.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, volume 26, 2013. URL https://papers.nips.cc/paper/5071-translating-embeddings-for-modeling-multi-relational-data.

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

Ho Yin Chau, Frank Qiu, Yubei Chen, and Bruno Olshausen. Disentangling images with Lie group transformations and sparse coding. In *Proceedings of the 1st NeurIPS Workshop on Symmetry and Geometry in Neural Representations*, volume 197 of *Proceedings of Machine Learning Research*, pp. 22–47, 2023. URL https://proceedings.mlr.press/v197/chau23a.html.

Junfeng Chen and Kailiang Wu. Deep-OSG: Deep learning of operators in semigroup. *Journal of Computational Physics*, 493:112498, 2023. doi: 10.1016/j.jcp.2023.112498. URL https://doi.org/10.1016/j.jcp.2023.112498.

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

Jeremy E. Cohen and Nicolas Gillis. Identifiability of complete dictionary learning. *SIAM Journal on Mathematics of Data Science*, 1(3):518–536, 2019. doi: 10.1137/18M1233339. URL https://doi.org/10.1137/18M1233339.

Taco Cohen and Max Welling. Learning the irreducible representations of commutative Lie groups. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32(2) of *Proceedings of Machine Learning Research*, pp. 1755–1763, 2014. URL https://proceedings.mlr.press/v32/cohen14.html.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 23(226):1–61, 2022. URL https://www.jmlr.org/papers/v23/20-1335.html.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012. URL https://www.jmlr.org/papers/v13/gretton12a.html.

In Huh, Eunho Yang, Sung Ju Hwang, and Jinwoo Shin. Time-reversal symmetric ODE network. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://papers.nips.cc/paper/2020/hash/db8419f41d890df802dca330e6284952-Abstract.html.

Tyler Ingebrand, Adam J. Thorpe, and Ufuk Topcu. Zero-shot transfer of neural ODEs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://papers.nips.cc/paper_files/paper/2024/hash/7ce9df1d14adc2806df18c8e168d7ef9-Abstract-Conference.html.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural controlled differential equations for irregular time series. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6696–6707, 2020. URL https://papers.neurips.cc/paper_files/paper/2020/hash/4a5876b450b45371f6cfe5047ac8cd45-Abstract.html.

Sébastien Lachapelle, Pau Rodriguez, Yash Sharma, Katie E. Everett, Rémi Le Priol, Alexandre Lacoste, and Simon Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 428–484, 2022. URL https://proceedings.mlr.press/v177/lachapelle22a.html.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=c8P9NQVtmnO.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124, 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

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

Loïc Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dSprites: Disentanglement testing sprites dataset. https://github.com/google-deepmind/dsprites-dataset, 2017.

Thomas M. Norman, Max A. Horlbeck, Joseph M. Replogle, Alex Y. Ge, Albert Xu, Marco Jost, Luke A. Gilbert, and Jonathan S. Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019. doi: 10.1126/science.aax4438. URL https://doi.org/10.1126/science.aax4438.

Alexei Novikov and Stephen White. Dictionary learning for the almost-linear sparsity regime. In *Proceedings of the 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pp. 1516–1554, 2023. URL https://proceedings.mlr.press/v201/novikov23a.html.

Roussel Desmond Nzoyem, David A. W. Barton, and Tom Deakin. Neural context flows for meta-learning of dynamical systems. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8vzMLo8LDN.

Katharina Ott, Prateek Katiyar, Philipp Hennig, and Michael Tiemann. ResNet after all: Neural ODEs and their numerical solution. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=HxzSxSxLOJZ.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42:927–935, 2024. doi: 10.1038/s41587-023-01905-6. URL https://doi.org/10.1038/s41587-023-01905-6.

Meera Sitharam, Mohamad Tarifi, and Menghan Wang. An incidence geometry approach to dictionary learning. In *Proceedings of the 26th Canadian Conference on Computational Geometry*, pp. 419–430, 2014. URL https://cccg.ca/proceedings/2014/papers/paper62.pdf.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Daniel MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An extensive benchmark for scientific machine learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/0a9747136d411fb83f0cf81820d44afb-Abstract-Datasets_and_Benchmarks.html. Datasets and Benchmarks Track.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Yonina C. Eldar and Gitta Kutyniok (eds.), *Compressed Sensing: Theory and Applications*, pp. 210–268. Cambridge University Press, 2012. doi: 10.1017/CBO9780511794308.006. URL https://arxiv.org/abs/1011.3027.

Danru Xu, Dingling Yao, Sebastien Lachapelle, Perouz Taslakian, Julius von Kügelgen, Francesco Locatello, and Sara Magliacane. A sparsity principle for partially observable causal representation learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 55389–55433, 2024. URL https://proceedings.mlr.press/v235/xu24ac.html.

Yi Yu, Tengyao Wang, and Richard J. Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015. doi: 10.1093/biomet/asv008. URL https://doi.org/10.1093/biomet/asv008.

Xinqi Zhu, Chang Xu, and Dacheng Tao. Commutative Lie group VAE for disentanglement learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12924–12934, 2021. URL https://proceedings.mlr.press/v139/zhu21f.html.

A PROOFS AND IDENTIFICATION BOUNDARIES

We first establish what endpoint observations leave ambiguous, then prove how grouped local-time observations restrict that ambiguity. The recovery bounds track errors from finite duration, noise, and neural approximation. Separate map-error bounds explain why route agreement need not imply accurate prediction. Appendix B gives the benchmark protocols and additional experiments.

Table 5: Closest methodological comparisons. "Identification" denotes the estimand in the cited formulation, not informal interpretability.

Approach	Shared object and variation	Native test-time query	Difference from this setting
TransE (Bordes et al., 2013)	Entity embeddings and relation translations are shared across triples.	Link prediction by scoring $h + r \approx t$.	Establishes the translational precedent; it does not transport observations with a state-dependent field.
Neural CDE (Kidger et al., 2020)	A vector-field matrix is driven by a control path.	Supervised sequence modeling from an observed control path.	MPT uses the linear control path $X(t) = t\alpha$; the graph supplies α on an unseen edge.
FE-NODE (Ingebrand et al., 2024)	Basis functions span system dynamics; new observations determine coefficients.	New-system identification, prediction, and control with task support.	Our no-support adapter maps a known graph difference to coefficients using training edges only.
Neural Context Flows (Nzoyem et al., 2025)	A context-modulated field adapts across environments.	Few-shot context fitting; parameter-induced zero-shot adaptation is also discussed.	We fit an affine graph adapter on training edges and report support-based results separately.
GEARS (Roohani et al., 2024)	Gene and perturbation embeddings predict a perturbation set's transcriptome.	Heldout single- and multigene perturbations in gene space.	Norman uses the same heldout combinations, with native and decoder-matched scoring.
Disentanglement and causal representations (Locatello et al., 2020; Lachapelle et al., 2022; Xu et al., 2024; Bing et al., 2024)	Representations are constrained by paired changes, interventions, or sparsity.	Factor or causal-variable recovery under model-specific observation assumptions.	Our state representation is fixed; Theorem 1 identifies only generator lines under grouped local-time premises.

A.1 TRANSPORT NOTATION AND DIAGNOSTICS

The recurring objects have the following types. Symbols used only in a bound are defined with that bound.

Symbol	Meaning
$D \subseteq \mathbb{R}^d$, $x \in D$	State domain and source state.
$\beta_e, \alpha \in \mathbb{R}^K$	Condition code and difference $\alpha = \beta_{e'} - \beta_e$.
$v_i : D \to \mathbb{R}^d$, $V(x) \in \mathbb{R}^{d\times K}$	Shared fields and their values at x.
$V : \mathbb{R}^K \to \mathcal{F}(D)$	Linear map from coefficients to the whole field $Va = \sum_i a_i v_i$.
T_α, $\widehat{T}_\alpha$	Exact flow and computed prediction; $\widehat{T}_{e\to e'} = \widehat{T}_{\beta_{e'}-\beta_e}$.
S_r, W_r	Active index set and exact action subspace $V(E_{S_r})$.
V_X, U_i, W_r^X	Dictionary evaluated at probes X_M, its column $U_i = V_X e_i$, and the probe subspace $V_X(E_{S_r})$.
P_r, $\widehat{P}_r$	Exact and estimated orthogonal projectors in the finite-dimensional probe space.

For vectors, an unsubscripted norm is Euclidean; $\|\cdot\|_{\mathrm{op}}$ and $\|\cdot\|_F$ denote matrix operator and Frobenius norms. The vectors e_i form the standard coordinate basis, and I_m and Id denote the identity matrix and map. A hat denotes an estimate or fitted predictor; a tilde denotes an equivalent representation or intermediate estimate as specified locally. For a field g, $\phi_t^g(x)$ is the solution at time t of $\dot{z} = g(z)$ starting at x. Thus the combined field and its time-one flow are

$$g_{e,e'}(x) = V(x)(\beta_{e'} - \beta_e) = \sum_{i=1}^{K} (\beta_{e',i} - \beta_{e,i}) v_i(x), \qquad T_{e \to e'} = \phi_1^{g_{e,e'}}. \tag{6}$$

Writing $\alpha^{e,e'} = \beta_{e'} - \beta_e$, subtraction of node codes gives the algebraic identities

$$\alpha^{e,e''} = \alpha^{e,e'} + \alpha^{e',e''}, \qquad \alpha^{e',e} = -\alpha^{e,e'}. \tag{7}$$

The codes sum to zero around any closed graph path: they form the exact graph 1-cochain induced by node potentials. Edge effects with nonzero code circulation require a richer representation. The additive code relation does not by itself make finite actions compose along distinct control directions. Equation (1) is also a neural CDE with a linear control path: setting $X(t) = t\alpha$ in $dz = V_\theta(z)dX$ gives $\dot{z} = V_\theta(z)\alpha$.

Let $\mathcal{D}_{\mathrm{train}}$ collect the observations supplied on training edges, and $\mathcal{P}_{\mathrm{train}}$ their edge labels. For query $e \to e'$, the information available under Section 3.1 is

$$\mathcal{I}_{\mathrm{test}} = \{x, e, e', \beta_e, \beta_{e'}, \mathcal{P}_{\mathrm{train}}, \mathcal{D}_{\mathrm{train}}\}. \tag{8}$$

It includes node metadata but not target-edge outcomes, pairs, or coefficients fitted from that edge. For a designated invariant feature $c : D \to \mathbb{R}$, leakage is $|c(\widehat{T}(x)) - c(x)|$. We define the direct–composed path diagnostic by

$$\mathcal{E}_{\mathrm{path}} = \left\| \widehat{T}_{e \to e''}(x) - \widehat{T}_{e' \to e''}(\widehat{T}_{e \to e'}(x)) \right\|_2^2. \tag{9}$$

Neither quantity alone tests whether a predicted intervention reaches its target.

Write $Dv_i(x)$ for the Jacobian of v_i with respect to the state. For C^2 fields with bounded derivatives near the small-control trajectories, Taylor expansion of the exact flow gives

$$T_\alpha(x) = x + \sum_i \alpha_i v_i(x) + \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j Dv_i(x) v_j(x) + O(\|\alpha\|^3). \tag{10}$$

The quadratic term shows how state dependence produces nonlinear control interactions in the endpoint even though the generator is linear in the control. The derivation follows Proposition 2 below.

A.2 Gauge invariance and finite-time non-identifiability

Two ambiguities must be distinguished: different bases can define the same vector field, and different vector fields can define the same fixed-duration map. The first is the linear gauge below.

Proposition 1 (Linear gauge invariance)**.** *Let $A \in GL(K)$, $b \in \mathbb{R}^K$, $\widetilde{V}(x) = V(x)A$, and $\widetilde{\beta}_e = A^{-1}\beta_e + b$. For every e, e' and $x \in D$,*

$$\widetilde{V}(x)(\widetilde{\beta}_{e'} - \widetilde{\beta}_e) = V(x)(\beta_{e'} - \beta_e).$$

Where these fields have unique flows, $\widetilde{T}_{e \to e'} = T_{e \to e'}$ on their common domain.

Proof. The translation cancels from code differences and $\widetilde{V}(x)(\widetilde{\beta}_{e'} - \widetilde{\beta}_e) = V(x)AA^{-1}(\beta_{e'} - \beta_e)$. The induced ODEs have the same right-hand side, hence the same flow wherever uniqueness holds. □

The proposition explains why accurate transport alone cannot validate a learned mechanism basis: every permitted change of basis preserves the entire combined action. A basis-recovery claim needs information that distinguishes these representations, such as the grouped supports in Theorem 1.

For the second ambiguity, let $J = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$ and $g_m(x) = 2\pi m J x$ for $m \in \mathbb{Z}$. These distinct fields all have the identity as their time-one flow. Fixed-duration endpoints therefore need not determine even the combined infinitesimal generator.

Short-time observations remove this duration ambiguity: for nonzero $t_n \to 0$, $g(x) = \lim_n (\phi^g_{t_n}(x) - x)/t_n$. Agreement along such a sequence determines the generator at x. Grouped observations of these velocities supply the action subspaces in Section 5. By contrast, the following flow identities hold regardless of whether the basis is identified.

Proposition 2 (Autonomous flow algebra)**.** *Let $g_\alpha = V\alpha$ be locally Lipschitz on D, with $T_\alpha = \phi_1^{g_\alpha}$. On any common domain where the required trajectories exist, $T_0 = \mathrm{Id}$, $T_{-\alpha} = T_\alpha^{-1}$, and $T_{b\alpha} \circ T_{a\alpha} = T_{(a+b)\alpha}$ for $a, b \in \mathbb{R}$.*

Remark (Architectural relations and learned accuracy). These classical flow identities separate two choices tested in the experiments. An autonomous field with $g_{-\alpha} = -g_\alpha$ already enforces reversal; linear control scaling also ties control magnitude to elapsed flow time and gives same-direction composition. Autonomous-odd retains the first relation without imposing the second.

Proof. Since $g_0 = 0$, its flow is the identity. Linearity gives $g_{a\alpha} = a g_\alpha$ and hence $T_{a\alpha} = \phi_a^{g_\alpha}$. Uniqueness of the autonomous flow gives $\phi_s^g \circ \phi_t^g = \phi_{s+t}^g$, which yields reversal and collinear composition. □

For two non-collinear controls, set $v = g_\alpha$ and $w = g_\gamma$, and assume these fields are C^2 on an open neighborhood of the trajectories. On a common local flow domain, the identity $\phi_t^w \circ \phi_s^v = \phi_1^{sv+tw}$ holds for all sufficiently small s, t if and only if $[v, w] = Dw\, v - Dv\, w = 0$ there. The necessity concerns this local family, not equality at one fixed time. At a fixed state x, the difference $T_{\varepsilon\gamma}(T_{\varepsilon\alpha}(x)) - T_{\varepsilon(\alpha+\gamma)}(x)$ is $\frac{1}{2}\varepsilon^2[v, w](x) + O(\varepsilon^3)$. Thus code addition alone does not imply finite-flow composition. Fixed-step RK4 introduces a further numerical defect even in the collinear case.

Equation (10) follows by Taylor expansion in time. For $g_\alpha = \sum_i \alpha_i v_i$, the trajectory satisfies $z'(0) = g_\alpha(x)$ and $z''(0) = Dg_\alpha(x) g_\alpha(x)$. Its third derivative is $D^2 g_\alpha(z)[g_\alpha(z), g_\alpha(z)] + Dg_\alpha(z) Dg_\alpha(z) g_\alpha(z)$. Bounded fields and first two derivatives near the small-control trajectories bound this remainder by $O(\|\alpha\|^3)$ uniformly over $0 \le t \le 1$. Substituting $t = 1$ yields the stated quadratic interaction. The expansion concerns the exact flow near zero control, not arbitrary finite codes or exact RK4 output.

A.3 Proof and stability of support-incidence identification

Exact residual transformations. Let $E_S = \{a \in \mathbb{R}^K : a_j = 0 \text{ for } j \notin S\}$, define the incidence signature $\mathcal{I}_i = \{r : i \in S_r\}$, and set

$$C_i = \bigcap_{r : i \in S_r} S_r = \{j : \mathcal{I}_i \subseteq \mathcal{I}_j\},$$

with an empty intersection interpreted as $[K]$. The relation $i \preceq j$ iff $j \in C_i$ is a preorder.

We first prove the exact gauge characterization. If $B(E_{S_r}) = E_{S_r}$ and $i \in S_r$, then $Be_i \in E_{S_r}$. Hence $B_{ji} = 0$ whenever at least one support contains i and excludes j, equivalently whenever $j \notin C_i$. Conversely, suppose this zero pattern holds. For $i \in S_r$, the definition gives $C_i \subseteq S_r$, so $Be_i \in E_{S_r}$. Expanding $x \in E_{S_r}$ in coordinate axes yields $Bx \in E_{S_r}$. Thus $B(E_{S_r}) \subseteq E_{S_r}$; invertibility and equal finite dimension imply equality. This proves Eq. (4).

Remark (Admissible blocks and shears). Let $\mathcal{Q}$ be the partition of $[K]$ into classes with equal incidence signatures, and write $\mathcal{I}_Q$ for the signature of a class Q. In coordinates ordered by these classes,

$$\mathcal{G}_\mathcal{S} = \left\{ B : \begin{array}{ll} B_{Q,Q} \in GL(|Q|) & \text{for every } Q \in \mathcal{Q}, \\ B_{Q',Q} = 0 & \text{if } \mathcal{I}_Q \not\subseteq \mathcal{I}_{Q'} \end{array} \right\}. \tag{11}$$

All remaining off-diagonal blocks are unrestricted. Ordering the distinct signatures by inclusion makes this a block-triangular matrix; invertibility is equivalent to invertibility of its diagonal blocks. Thus equal signatures allow arbitrary invertible mixing within their class. For $i \neq j$ and $t \neq 0$,

$$I_K + t e_j e_i^\top \in \mathcal{G}_\mathcal{S} \quad \Longleftrightarrow \quad \mathcal{I}_i \subseteq \mathcal{I}_j. \tag{12}$$

Strict inclusion allows this shear but excludes the reverse shear $I_K + te_i e_j^\top$. Incomparable signatures give $B_{ji} = B_{ij} = 0$ for every $B \in \mathcal{G}_\mathcal{S}$, so neither direction can mix into the other. The residual group is the unit group of a structural matrix algebra (Akkurt et al., 2005).

For the line-identification conclusion, coverage $\cup_r S_r = [K]$ gives $\sum_r W_r = \operatorname{im} V$. Any competing injective dictionary with the same grouped subspaces therefore has the same image and differs by an invertible B, to which the preceding characterization applies. Coverage also ensures that each $\mathcal{I}_i$ is nonempty. It is not needed for the zero-pattern identity itself: with $K = 1$ and only empty supports, that identity holds but the observed zero subspaces reveal no generator line. Because V is injective,

$$V(E_S) \cap V(E_T) = V(E_{S \cap T}), \tag{13}$$

and the identity extends to finite intersections. The gauge is diagonal exactly when $C_i = \{i\}$ for every i, or equivalently when every ordered pair $i \neq j$ is separated by a support containing i and excluding j. In this case

$$\bigcap_{r : i \in S_r} W_r = V(E_{C_i}) = \operatorname{span}(v_i).$$

Under separation, any competing dictionary with the same labeled supports is therefore $\widetilde{V} = V\Lambda$ with $\Lambda \in \mathcal{D}_K$. If labels may change, let $P_\pi e_i = e_{\pi(i)}$. The equivalent representation is

$$\widetilde{V} = V P_\pi \Lambda, \qquad \widetilde{S}_r = \pi^{-1}(S_r), \qquad \widetilde{V}(E_{\widetilde{S}_r}) = W_r. \tag{14}$$

Here π is one common permutation across all groups, and $\Lambda \in \mathcal{D}_K$. A permutation that changes fixed support labels is not an additional element of $\mathcal{G}_\mathcal{S}$. If separation fails, Eq. (12) constructs a non-diagonal equivalent dictionary, proving the necessity of Eq. (5).

Approximate subspace preservation. The exact argument also bounds how much an approximately preserving transform can mix forbidden coordinates. Let P_S project onto E_S, let $S^c = [K] \setminus S$, and suppose

$$\left\| P_{S_r^c} B P_{S_r} \right\|_{\mathrm{op}} \leq \varepsilon_r.$$

For $j \notin C_i$, choose any witnessing support with $i \in S_r$ and $j \notin S_r$. Then

$$|B_{ji}| = |e_j^\top P_{S_r^c} B P_{S_r} e_i| \leq \varepsilon_r.$$

Minimizing over witnesses gives $|B_{ji}| \leq \varepsilon_{ji}$, where $\varepsilon_{ji} = \min_{r : i \in S_r, j \notin S_r} \varepsilon_r$. If $\Pi_\mathcal{S}$ zeros all forbidden entries, then

$$\|B - \Pi_\mathcal{S} B\|_F \leq \left(\sum_{i=1}^{K} \sum_{j \notin C_i} \varepsilon_{ji}^2 \right)^{1/2}. \tag{15}$$

Under separation and a uniform $\varepsilon_r \leq \varepsilon$, this becomes $\operatorname{dist}_F(B, \mathrm{Diag}) \leq \sqrt{K(K-1)}\varepsilon$. Here Diag contains all real diagonal $K \times K$ matrices, and $\operatorname{dist}_F(B, \mathrm{Diag}) = \inf_{A \in \mathrm{Diag}} \|B - A\|_F$. The bound controls distance to the permitted zero-pattern algebra. It assumes known supports and gives no condition-number bound on the projected matrix. Mechanisms with identical incidence signatures remain indistinguishable.

Pure interventions are sufficient but not necessary. With supports $\{1, 2\}$, $\{1, 3\}$, and $\{2, 3\}$, every closure is a singleton even though each intervention activates two mechanisms. By contrast, if mechanism 1 never appears without mechanism 2, its closure contains 2 and the corresponding shear cannot be removed from these observations.

Relation to dictionary identification. Recovering directions from mixtures is an established dictionary-learning problem. Cohen & Gillis (2019) give deterministic sample conditions for uniqueness of complete and undercomplete dictionaries. Sitharam et al. (2014) use incidence rigidity to characterize generic finite identifiability. Novikov & White (2023) recover support subspaces and intersect them to recover atoms under a random overcomplete model. Theorem 1 starts from known grouped subspaces and an injective field dictionary. It characterizes all transformations compatible with a prescribed support family, rather than improving the sample complexity of those recovery results. When separation fails, the characterization identifies the remaining mixtures and the additional support contrasts that exclude them. The proof applies the zero pattern of a structural matrix algebra (Akkurt et al., 2005) to the grouped intervention observations.

Unknown support masks. The preceding characterization assumes known supports. When groups are known but their masks are not, one-dimensional subspace intersections can recover both lines and memberships. Let $\mathcal{W} = (W_r)_{r=1}^R$. For each nonempty class set $J \subseteq [R]$, define

$$W_J = \bigcap_{r \in J} W_r, \qquad \mathfrak{A}(\mathcal{W}) = \{W_J : \dim W_J = 1\}.$$

Injectivity of V and Eq. (13) give

$$W_J = V(E_{\cap_{r \in J} S_r}). \tag{16}$$

Every one-dimensional meet is a true dictionary line. Under coverage and separation, choosing $J = \mathcal{I}_i$ makes the support intersection $C_i = \{i\}$, so every true line occurs. Conversely, if $\operatorname{span}(v_i) = W_J$, then $\cap_{r \in J} S_r = \{i\}$; for every $j \neq i$, some $r \in J$ contains i and excludes j. Hence, for a covered family,

$$\mathfrak{A}(\mathcal{W}) = \{\operatorname{span}(v_i) : i \in [K]\} \quad \iff \quad \mathcal{S} \text{ is separating.} \tag{17}$$

Once these unlabeled lines have been enumerated, linear independence gives the mask readout

$$S_r = \{i : \operatorname{span}(v_i) \subseteq W_r\}. \tag{18}$$

This estimator uses the supplied grouping, without requiring its masks or cardinalities. It does not infer classes from pooled transitions. Exhaustive enumeration uses up to $2^R - 1$ intersections and is practical for the small numbers of classes used here. Restricting the intersection order recovers every line only if each line has a singleton witness of that order.

We use the acute angle between unoriented lines: for nonzero u, v, its cosine is $|u^\top v|/(\|u\| \, \|v\|)$ and its sine is $\sqrt{1 - \cos^2 \angle(u, v)}$. Equivalently, $\sin \angle(u, \mathcal{L}) = \|(I - P_{\mathcal{L}})u\| \, / \, \|u\|$, where $P_{\mathcal{L}}$ projects onto the line $\mathcal{L}$.

For stability, let P_r and $\widehat{P}_r$ be exact and estimated class projectors and put

$$Q_J = I - \frac{1}{|J|} \sum_{r \in J} P_r, \qquad \widehat{Q}_J = I - \frac{1}{|J|} \sum_{r \in J} \widehat{P}_r, \qquad \epsilon_J = \frac{1}{|J|} \sum_{r \in J} \left\| \widehat{P}_r - P_r \right\|_{\mathrm{op}}.$$

Then $\ker Q_J = W_J$ and $\left\| \widehat{Q}_J - Q_J \right\|_{\mathrm{op}} \leq \epsilon_J$. Suppose every positive eigenvalue among the finite family $\{Q_J\}$ is at least $\gamma > 0$ and $\max_J \epsilon_J < \gamma/3$. Weyl's inequality separates the cases $\dim W_J = 0$, 1, and at least 2: respectively, the smallest estimated eigenvalue exceeds $2\gamma/3$; exactly one lies below $\gamma/3$; or at least two lie below $\gamma/3$. For a selected singleton, the residual argument used in Theorem 2 yields line sine at most $2\epsilon_J/\gamma$.

Membership recovery needs a margin as well. Let u_i and $\widehat{u}_i$ be unit true and estimated directions after matching and orientation, and set

$$\varepsilon_{\mathrm{mem}} = \max_r \left\| \widehat{P}_r - P_r \right\|_{\mathrm{op}} + \max_i \|\widehat{u}_i - u_i\|, \qquad \kappa = \min_{r, i : i \notin S_r} \|(I - P_r)u_i\|.$$

Assume the set of nonmembers is nonempty. The residual $\left\|(I - \widehat{P}_r)\widehat{u}_i\right\|$ differs from $\|(I - P_r)u_i\|$ by at most $\varepsilon_{\mathrm{mem}}$. If $\kappa > 2\varepsilon_{\mathrm{mem}}$, thresholding at $\kappa/2$ distinguishes members from nonmembers. This is a deterministic sufficient condition; the experimental discovery procedure uses fixed thresholds, not an estimated certificate for κ. The finite-sample theorem below supplies projector errors under a specified sampling model.

Choosing support groups. For $K \geq 2$, the signatures determine how many groups can suffice for separation. It requires $\mathcal{I}_i \not\subseteq \mathcal{I}_j$ for every $i \neq j$, so the K signatures form an antichain in $2^{[R]}$. Sperner's theorem gives

$$K \leq \binom{R}{\lfloor R/2 \rfloor}, \qquad R_{\min}(K) = \min \left\{ R : \binom{R}{\lfloor R/2 \rfloor} \geq K \right\}. \tag{19}$$

The minimum is over positive integers R. It is attained by selecting distinct signatures of cardinality $\lfloor R/2 \rfloor$ and defining $S_r = \{i : r \in \mathcal{I}_i\}$. For $K = 1$, one nonempty group suffices. This is an incidence-level count; every class still needs enough coefficient excitation and samples.

The number of groups alone does not measure stability. The separation margin also depends on how often each pair of mechanisms is distinguished. In coefficient space let

$$Q_i^{\mathrm{coef}} = \sum_{r \in \mathcal{I}_i} (I - P_{S_r}), \qquad q_{ij} = |\mathcal{I}_i \setminus \mathcal{I}_j|.$$

Direct evaluation on the coordinate axes gives

$$Q_i^{\mathrm{coef}} e_j = q_{ij} e_j, \quad \ker Q_i^{\mathrm{coef}} = E_{C_i}, \quad \gamma_i^{\mathrm{coef}} = \min_{j \notin C_i} q_{ij}. \tag{20}$$

The expression for γ_i^{coef} applies when $C_i \neq [K]$; otherwise $Q_i^{\mathrm{coef}} = 0$ has no positive eigenvalue. For equal-cardinality signatures, $q_{ij} = |\mathcal{I}_i \triangle \mathcal{I}_j|/2$. Designing a separated support family with a large ideal incidence gap is therefore the same combinatorial objective as constructing a constant-weight code with large minimum Hamming distance. The antichain capacity statement is machine-checked by invoking Mathlib's formal Sperner theorem.

Perturbed observed subspaces. In observation coordinates, a spectral gap measures how sharply the subspaces isolate a line. The next result converts projector error into line error.

Theorem 2 (Robust support-intersection recovery)**.** *Let $\mathcal{R}_i \subseteq [R]$ be nonempty. In a finite-dimensional real Hilbert space, let P_r be the orthogonal projector onto W_r, and assume $\bigcap_{r \in \mathcal{R}_i} W_r = \mathcal{L}_i = \mathrm{span}(v_i)$ for a nonzero v_i. Let $\widehat{P}_r$ be estimated orthogonal projectors and define*

$$Q_i = \sum_{r \in \mathcal{R}_i} (I - P_r), \qquad \widehat{Q}_i = \sum_{r \in \mathcal{R}_i} (I - \widehat{P}_r).$$

Assume Q_i has spectral gap $\gamma_i > 0$ on $\mathcal{L}_i^{\perp}$, and let $\Delta_i = \sum_{r \in \mathcal{R}_i} \left\| \widehat{P}_r - P_r \right\|_{\mathrm{op}}$. If $\widehat{v}_i$ is a unit bottom eigenvector of $\widehat{Q}_i$, then

$$\sin \angle(\widehat{v}_i, \mathcal{L}_i) \leq 2\Delta_i / \gamma_i.$$

Remark (Identification versus stable recovery). The numerator Δ_i measures error in the estimated group subspaces; the denominator γ_i measures how sharply their intersection isolates the line. At a fixed positive gap, more accurate group estimates tighten the line bound. When $2\Delta_i/\gamma_i \geq 1$, the bound gives no restriction beyond the range of a line sine. The gap is measured in observation coordinates, so the exact support-separation criterion alone does not ensure stable recovery. Choosing intervention groups or probe states must also account for this geometry.

Proof. The kernel of Q_i is $\mathcal{L}_i$ and $\|E_i\|_{\mathrm{op}} \leq \Delta_i$ for $E_i = \widehat{Q}_i - Q_i$. If $\widehat{\lambda}_i$ is the bottom eigenvalue of $\widehat{Q}_i$, comparison with a unit vector in $\mathcal{L}_i$ gives $0 \leq \widehat{\lambda}_i \leq \Delta_i$. The eigenvector equation therefore gives $Q_i \widehat{v}_i = \widehat{\lambda}_i \widehat{v}_i - E_i \widehat{v}_i$ and $\|Q_i \widehat{v}_i\| \leq 2\Delta_i$. The spectral gap on $\mathcal{L}_i^{\perp}$ gives $\|Q_i \widehat{v}_i\| \geq \gamma_i \sin \angle(\widehat{v}_i, \mathcal{L}_i)$. □

Line recovery leaves scale and orientation unresolved. Recovering normalized fields and environment codes also requires local-time generator information, connected graph coverage, an orientation convention, and a reference environment, all in fixed coordinates.

Finite duration and endpoint noise. The deterministic projector bound becomes statistical once an observation model is specified. We first distinguish whole fields from their measured values. For fixed probes $X_M = \{x_1, \ldots, x_M\} \subset D$, define

$$\mathrm{Ev}_{X_M}(g) = (g(x_1), \ldots, g(x_M)) \in \mathbb{R}^H, \qquad V_X = \mathrm{Ev}_{X_M} \circ V, \qquad H = Md.$$

Write $U_i = V_X e_i$ and $W_r^X = V_X(E_{S_r})$. If $\mathrm{rank}(V_X) = K$, evaluation is injective on the dictionary span and, for every nonempty $J \subseteq [R]$,

$$\bigcap_{r \in J} W_r^X = V_X \left(E_{\bigcap_{r \in J} S_r} \right).$$

Theorem 1 then applies to the stacked directions U_i. A separating support family isolates each stacked line, and Theorem 2 applies with W_r^X and U_i in place of W_r and v_i.

This global rank condition is stronger than full column rank of each class matrix $V_r = (U_i)_{i \in S_r}$. Classwise rank suffices for the projector bound under the sampling model below, but need not preserve intersections: distinct whole-field subspaces can coincide at the probes. The finite-sample theorem states its line bound separately under the probe-space intersection and gap conditions. It does not infer those conditions from classwise rank. We next bound finite-difference bias at observation duration t, then add Gaussian excitation and endpoint noise.

Proposition 3 (Short-time bias). *Let $t > 0$ and $L, G \geq 0$. Suppose $g : D \to \mathbb{R}^d$ is L-Lipschitz and its trajectory from $x \in D$ exists in D over $[0, t]$, with $\|g(\phi_s^g(x))\| \leq G$ for every $s \in [0, t]$. Then*

$$\|\{\phi_t^g(x) - x\}/t - g(x)\| \leq LGt/2.$$

Proof. Write the difference as $t^{-1} \int_0^t \{g(\phi_s^g(x)) - g(x)\}\, ds$ and use $\|\phi_s^g(x) - x\| \leq Gs$. □

The finite difference measures average motion over an interval, whereas identification uses the initial velocity. This standard bound makes the resulting bias explicit: increasing sample count at a fixed duration does not remove the discrepancy. The next result shows why shortening the interval is not cost-free either.

Theorem 3 (Noisy finite-sample support recovery). *For each class $r \in [R]$, let $s_r = |S_r| \in \{1, \dots, H\}$, $n_r \geq 1$, and fix a field-stack matrix $V_r \in \mathbb{R}^{H \times s_r}$ satisfying $0 < \nu_r \leq \sigma_{\min}(V_r)$ and $\|V_r\|_{\mathrm{op}} \leq M_r$. Observe $Y_r = V_r A_r + D_r + N_r \in \mathbb{R}^{H \times n_r}$, where $A_r = \Sigma_r^{1/2} G_r \in \mathbb{R}^{s_r \times n_r}$, G_r has independent standard-normal entries, and the fixed covariance satisfies $0 < \mu_r I_{s_r} \preceq \Sigma_r \preceq \Lambda_r I_{s_r}$. Let $N_r \in \mathbb{R}^{H \times n_r}$ have independent $\mathcal{N}(0, \sigma_t^2)$ entries, independently of (A_r, D_r). Choose a nonnegative bound $d_r(t)$ on $\|D_r\|_{\mathrm{op}} / \sqrt{n_r}$; it may depend on (A_r, D_r). Let P_r project onto $\mathrm{col}(V_r)$ and $\widehat{P}_r$ onto the leading s_r left singular vectors of Y_r. For failure probability $\eta \in (0, 1)$, set*

$$u = \sqrt{2 \log(6R/\eta)}, \quad q_{s,r} = \frac{\sqrt{s_r} + u}{\sqrt{n_r}}, \quad q_{H,r} = \frac{\sqrt{H} + u}{\sqrt{n_r}}, \quad q_{\times,r} = \frac{2\sqrt{H} + u}{\sqrt{n_r}},$$

suppose $q_{s,r} < 1$ and define

$$a_r = M_r \sqrt{\Lambda_r}(1 + q_{s,r}), \qquad g_r = \nu_r^2 \mu_r (1 - q_{s,r})^2,$$
$$\epsilon_r = 2a_r d_r(t) + d_r(t)^2 + \sigma_t^2(2q_{H,r} + q_{H,r}^2) + 2\sigma_t\{a_r + d_r(t)\}q_{\times,r}, \quad \bar{\delta}_r = \min\{1, 2\epsilon_r / g_r\}.$$

With probability at least $1 - \eta$ over the coefficient and noise draws, simultaneously for all $r \in [R]$,

$$\left\| \widehat{P}_r - P_r \right\|_{\mathrm{op}} \leq \bar{\delta}_r.$$

For every line satisfying the intersection and gap assumptions of Theorem 2, its bottom-eigenvector estimate obeys on the same event

$$\sin \angle(\widehat{v}_i, \mathcal{L}_i) \leq \frac{2}{\gamma_i} \sum_{r \in \mathcal{R}_i} \bar{\delta}_r.$$

Remark (Choosing observation duration and excitation). In ϵ_r, the first two terms account for finite-duration bias; the remaining terms account for noise and its interaction with the signal and bias. When $q_{s,r} \leq 1/2$ and $d_r(t) = O(t)$, the stochastic term is $O(n_r^{-1/2})$ at fixed t, but finite-duration bias remains. Independent source and target noise of standard deviation σ_x gives $\sigma_t = \sqrt{2}\sigma_x / t$, so shorter observations amplify noise. Weak excitation or nearly dependent field stacks reduce g_r, increasing the projector-error bound. Sample-size, duration, and excitation sweeps test these different dependencies; no one sweep replaces the others. The proof combines Gaussian concentration and Davis–Kahan perturbation to carry covariance error into subspace error.

Proof. Gaussian singular-value concentration bounds $\sigma_{\min}(A_r)/\sqrt{n_r}$ below by $\sqrt{\mu_r}(1 - q_{s,r})$ and $\|A_r\|_{\mathrm{op}} / \sqrt{n_r}$ above by $\sqrt{\Lambda_r}(1 + q_{s,r})$ (Vershynin, 2012, Corollary 5.35). Write $S_r^{\mathrm{sig}} = V_r A_r$ and $F_r = S_r^{\mathrm{sig}} + D_r$. On this event, $C_r = S_r^{\mathrm{sig}}(S_r^{\mathrm{sig}})^\top / n_r$ has rank s_r, eigengap at least g_r, and $\|S_r^{\mathrm{sig}}\|_{\mathrm{op}} / \sqrt{n_r} \leq a_r$. Expanding the empirical covariance gives

$$\frac{Y_r Y_r^\top}{n_r} - (C_r + \sigma_t^2 I_H) = \frac{F_r F_r^\top - S_r^{\mathrm{sig}}(S_r^{\mathrm{sig}})^\top}{n_r} + \left(\frac{N_r N_r^\top}{n_r} - \sigma_t^2 I_H \right) + \frac{F_r N_r^\top + N_r F_r^\top}{n_r}.$$

The first term has norm at most $2a_r d_r(t) + d_r(t)^2$. Gaussian singular-value bounds control the second by $\sigma_t^2(2q_{H,r} + q_{H,r}^2)$. Conditional on F_r, independence makes $F_r N_r^\top$ distributed as $\sigma_t(F_r F_r^\top)^{1/2} Z_r$, where Z_r is an $H \times H$ standard Gaussian matrix. Its norm is at most $\sigma_t \|F_r\|_{\mathrm{op}} (2\sqrt{H} + u)$ except on a Gaussian tail event. The third term is therefore at most $2\sigma_t\{a_r + d_r(t)\}q_{\times,r}$.

At most six one-sided Gaussian tail events per class are used. Their union has probability at most $6R\exp(-u^2/2) = \eta$, so the covariance perturbation is bounded by ϵ_r simultaneously. No independence between classes is needed for this union bound. Adding $\sigma_t^2 I_H$ to C_r leaves its eigenspaces and gap unchanged. When $\epsilon_r < g_r/2$, Weyl's inequality separates its leading eigenvalues from the remaining empirical eigenvalues by at least $g_r - \epsilon_r$. The operator-norm form of Davis–Kahan (Yu et al., 2015, Theorem 1 and the following remark) then gives

$$\left\|\widehat{P}_r - P_r\right\|_{\mathrm{op}} \le \frac{\epsilon_r}{g_r - \epsilon_r} \le \frac{2\epsilon_r}{g_r}.$$

If $\epsilon_r \ge g_r/2$, the projector bound $\left\|\widehat{P}_r - P_r\right\|_{\mathrm{op}} \le 1$ suffices. These cases yield $\bar{\delta}_r$; when $s_r = H$, both projectors are the identity and the bound is immediate. Theorem 2 gives the line bound. □

The bound permits a realized $d_r(t)$, which the synthetic audit computes using the known generator; an observable certificate would need its own valid bias bound. The result concerns labeled support groups in fixed coordinates.

From field estimates to neural transport. The spectral estimator recovers field values at probe states. A neural predictor must also approximate those values, extend them between probes, and integrate them accurately. The following certificate keeps these errors separate.

Theorem 4 (Spectral-to-neural transport certificate)**.** *Let $X_M = \{x_1, \dots, x_M\} \subset D$ be an h-net: every $x \in D \subset \mathbb{R}^d$ is within distance $h \ge 0$ of a probe. Choose one aligned permutation, orientation, and scale gauge, with the true stack $U_i = (v_i(x_1), \dots, v_i(x_M)) \in \mathbb{R}^{Md}$ of unit norm. Let $\widetilde{U}_i$ be a unit spectral estimate oriented so that $\langle \widetilde{U}_i, U_i \rangle \ge 0$, and set $\widehat{U}_i = (\widehat{v}_i(x_1), \dots, \widehat{v}_i(x_M))$. Define*

$$\delta_i = \sin \angle(\widetilde{U}_i, U_i), \qquad \rho_i = \left\|\widehat{U}_i - \widetilde{U}_i\right\|_2.$$

If v_i and $\widehat{v}_i$ are L_i- and $\widehat{L}_i$-Lipschitz on D, then

$$\sup_{x \in D} \|\widehat{v}_i(x) - v_i(x)\|_2 \le e_i := \rho_i + \sqrt{2}\delta_i + (L_i + \widehat{L}_i)h. \tag{21}$$

For aligned true and learned codes α and $\widehat{\alpha}$, set

$$\widehat{M}_i = \sup_D \|\widehat{v}_i\|, \quad E_\alpha = \sum_i |\alpha_i| e_i + \sum_i |\widehat{\alpha}_i - \alpha_i| \widehat{M}_i, \quad \widehat{L}_\alpha = \sum_i |\widehat{\alpha}_i| \widehat{L}_i.$$

Suppose the true and learned trajectories from $x \in D$ stay in D over $[0, T]$, $T \ge 0$, and the displayed supremum bounds are finite. Assume $\left\|\widehat{T}_{\mathrm{num}}(x) - \phi_T^{\widehat{V}\widehat{\alpha}}(x)\right\|_2 \le \zeta_{\mathrm{num}}$. Then

$$\left\|\widehat{T}_{\mathrm{num}}(x) - \phi_T^{V\alpha}(x)\right\|_2 \le \zeta_{\mathrm{num}} + E_\alpha \Psi(\widehat{L}_\alpha, T), \tag{22}$$

where $\Psi(0, T) = T$ and $\Psi(L, T) = (e^{LT} - 1)/L$ for $L > 0$. If the spectral estimates and their observation model satisfy Theorem 3, its simultaneous event permits the upper bound

$$\delta_i \le b_i = \frac{2}{\gamma_i} \sum_{r \in \mathcal{R}_i} \bar{\delta}_r, \tag{23}$$

The endpoint certificate then holds on this event whenever the remaining deterministic bounds hold.

Remark (What recovered directions imply for prediction). Equation (21) separates neural fitting error ρ_i, spectral line error δ_i, and the probe-coverage term involving h. Equation (22) then propagates field and code error and adds numerical integration error. For fixed R, H and bounded conditioning,

let $n = \min_r n_r$. With $d_r(t) = O(t)$, the spectral contribution is $O(t) + O_{\mathbb{P}}(C(t)n^{-1/2})$, where $C(t)$ collects duration-dependent factors. Reducing this contribution alone need not improve endpoint accuracy when fitting, coverage, or code error dominates. The proof uses interpolation and Gronwall's inequality; it assumes neural fitting and regularity bounds rather than proving optimization convergence. The solver term accounts for discretization dependence (Ott et al., 2021).

Proof. After orientation, the line angle θ_i lies in $[0, \pi/2]$, so $\left\|\widetilde{U}_i - U_i\right\| = 2\sin(\theta_i/2) \le \sqrt{2}\sin\theta_i$. The triangle inequality bounds the complete probe stack by $\rho_i + \sqrt{2}\delta_i$, hence bounds each probe point by the same quantity. For any $x \in D$, insert a probe within distance h and apply the two Lipschitz inequalities to obtain Eq. (21).

Adding and subtracting $\sum_i \alpha_i \widehat{v}_i(x)$ bounds the combined-generator discrepancy by E_α. The learned combined field is $\widehat{L}_\alpha$-Lipschitz. Gronwall's inequality gives the exact-flow term in Eq. (22); the numerical error is added by the triangle inequality. Finally, Eq. (23) is the simultaneous finite-sample line bound above. The scalar composition of these inequalities is machine-checked in Lean. □

The coverage assumption cannot be replaced by accurate fit at finitely many probes. Given a finite X_M in an open domain, choose a ball around $x_0 \notin X_M$ that avoids all probes and a smooth bump ψ equal to one on a smaller concentric ball. For a nonzero $a \in \mathbb{R}^d$, the zero field and $w(x) = \psi(x)a$ agree at every probe, but their flows from x_0 are x_0 and $x_0 + ta$ for sufficiently small $t > 0$. Probe fit alone cannot bound error uniformly on D. If M radius-h balls cover $[0,1]^d$ and ω_d is the unit-ball volume, then $M\omega_d h^d \ge 1$, so $h \ge (M\omega_d)^{-1/d}$. Coverage carries a dimension-dependent cost.

A.4 Cold-start and observation-coordinate boundaries

The recovery statements above condition on an estimator's output. They do not ensure that joint field–code optimization reaches that output. For nonzero y, consider

$$\ell(u, a; y) = \frac{1}{2}(ua - y)^2.$$

Its derivatives are $(ua - y)a$ and $(ua - y)u$, so $(u, a) = (0, 0)$ is stationary with loss $y^2/2$, whereas $(1, y)$ has zero loss. The unregularized field–code objective is not globally gradient dominated, and ordinary gradient descent started at the origin remains there for every step size. This rules out an every-initialization convergence guarantee for that method, not for algorithms that can escape a stationary point. The observed 39/40 success rate addresses a different question: behavior under the tested random starts.

Unknown observation coordinates introduce another ambiguity. Let $f : \mathcal{Z} \to \mathcal{X}$ be injective, let ϕ_a denote the latent action, and let $h : \mathcal{Z} \to \mathcal{Z}$ be any diffeomorphism. Define

$$\widetilde{f} = f \circ h^{-1}, \qquad \widetilde{\phi}_a = h \circ \phi_a \circ h^{-1}.$$

Substitution gives

$$\widetilde{f}\big(\widetilde{\phi}_a(h(z))\big) = f(\phi_a(z)) \tag{24}$$

for every z and a. Reconstruction and pixel transport leave an arbitrary nonlinear coordinate conjugacy in addition to the linear basis gauge within fixed coordinates.

Labeled equivariance can restrict this coordinate ambiguity. Suppose an additive action Φ_a is transitive on an orbit $\mathcal{O}$, and encoders $E, E' : \mathcal{O} \to \mathbb{R}^K$ satisfy

$$E(\Phi_a(x)) = E(x) + a, \qquad E'(\Phi_a(x)) = E'(x) + a.$$

Choose $x_0 \in \mathcal{O}$. Every $x \in \mathcal{O}$ equals $\Phi_a(x_0)$ for some a, and hence

$$E'(x) - E(x) = E'(x_0) + a - E(x_0) - a = E'(x_0) - E(x_0).$$

The two coordinates differ only by an origin translation on that orbit. Without transitivity, the translation and transverse coordinates may vary between orbits. This uniqueness statement relies on labeled equivariance, which the reconstruction-only image encoder used here is not constrained to satisfy.

For pure interventions, the remaining question is whether each observed direction also uses a single learned coordinate. The following conditional matrix bound quantifies this restriction.

Theorem 5 (Approximate anchor reduction). *Let $K \geq 1$, $\varepsilon \geq 0$, $B \in GL(K)$, and $b_i = Be_i$. Suppose $p : [K] \to [K]$ and $d_i \in \mathbb{R}$ satisfy $\left\|b_i - d_i e_{p(i)}\right\|_2 \leq \varepsilon$ for every i. If $\sigma_{\min}(B) > \sqrt{2}\varepsilon$, then p is a permutation and, for $Pe_i = e_{p(i)}$ and $\Lambda = \operatorname{diag}(d_1, \ldots, d_K)$, $\|B - P\Lambda\|_F \leq \sqrt{K}\varepsilon$.*

Remark (Purity needs conditioning). Nearly one-sparse columns can assign several coordinate directions to the same axis. The smallest-singular-value condition prevents this collapse and makes the assignment one-to-one. The result quantifies a conditional reduction to permutation and scaling. Applying it to intervention observations requires a code-purity condition as well as a pure true intervention, as made explicit below. Theorem 1 provides a different route through separating mixed interventions.

Proof. Set $r_i = b_i - d_i e_{p(i)}$. Each d_i is nonzero: otherwise $\|Be_i\| \leq \varepsilon < \sigma_{\min}(B)$. If columns $i \neq j$ choose the same row, the nonzero vector $c = d_j e_i - d_i e_j$ satisfies $Bc = d_j r_i - d_i r_j$. Hence $\|Bc\| \leq \sqrt{2}\varepsilon \|c\|$, contradicting the smallest-singular-value premise. Thus p is a permutation. Summing the squared column norms of $B - P\Lambda$ gives the Frobenius bound. □

To use the field-basis convention $\widetilde{V} = VB$, set $C = B^{-1}$. A true anchor with coefficient $a^{(i)} = s_i e_i$, $s_i \neq 0$, and an exactly matched learned generator satisfies

$$\widetilde{V}\widetilde{a}^{(i)} = Va^{(i)} \quad \Longrightarrow \quad \widetilde{a}^{(i)}/s_i = Ce_i,$$

where injectivity of V gives the implication. Thus a bound $\left\|\widetilde{a}^{(i)}/s_i - d_i e_{p(i)}\right\|_2 \leq \varepsilon$ supplies the column premise for C, not for B. If $\sigma_{\min}(C) > \sqrt{2}\varepsilon$, the theorem gives $\|C - P\Lambda\|_F \leq \sqrt{K}\varepsilon$. The inverse identity $B - (P\Lambda)^{-1} = B(P\Lambda - C)(P\Lambda)^{-1}$ then yields

$$\left\|B - (P\Lambda)^{-1}\right\|_F \leq \sqrt{K}\varepsilon \|B\|_{\mathrm{op}} \left\|(P\Lambda)^{-1}\right\|_{\mathrm{op}}.$$

This last factor records the conditioning needed to translate a code-space bound into a field-basis bound. Approximate generator matching requires its own contribution to ε; neither endpoint fit nor an ℓ_1 penalty alone establishes these code-purity and conditioning assumptions.

In the synthetic 10,000-update anchor experiment, basis cosine is 0.9999992 and aligned-beta error is 5.86×10^{-4} in 10/10 seeds. Basis cosine is the mean absolute cosine after the best one-to-one matching to the true fields. Aligned-beta error is the coordinate RMSE after subtracting each code set's reference-node code and fitting a least-squares linear map from learned to true codes. This post-hoc fit uses ground truth for evaluation, not training. The image experiments do not verify the anchor assumptions.

A.5 Generator and flow bounds

Prediction error can be bounded without identifying individual fields. We use two levels of information. A uniform combined-field error bound extends to controls in the observed span (Proposition 4). When only finite maps are considered, endpoint error depends on route disagreement, accumulated map error, and any difference between the true route and direct action (Theorem 6). Neither starts from empirical endpoint loss alone.

Proposition 4 (Transition-span bound). *Write $G_x(\alpha) = V(x)\alpha$ and $\widehat{G}_x(\alpha) = \widehat{V}(x)\alpha$. If $\alpha^\star = \sum_{j=1}^m c_j \alpha^{(j)}$ and $\sup_{x \in D} \left\|\widehat{G}_x(\alpha^{(j)}) - G_x(\alpha^{(j)})\right\| \leq \epsilon_j$, then*

$$\sup_{x \in D} \left\|\widehat{G}_x(\alpha^\star) - G_x(\alpha^\star)\right\| \leq \sum_{j=1}^m |c_j| \epsilon_j. \tag{25}$$

If $\alpha^\star$ lies outside the observed span, two linear generator maps can agree on every observed code and differ arbitrarily at $\alpha^\star$.

Equation (25) distinguishes code coverage from estimation accuracy. Inside the observed span, the coefficients needed to express the query amplify generator error. Under a common error bound, the

smallest amplification factor is $\inf\{\|c\|_1 : Ac = \alpha^\star\}$, where the columns of A are the observed codes. Outside that span, the second conclusion permits arbitrary disagreement even with perfect agreement at every observed code. More examples at the same codes cannot resolve this missing direction. The in-span bound requires uniform combined-generator accuracy; endpoint empirical risk minimization (ERM) does not supply it automatically.

Theorem 6 (Grounded missing-edge decomposition)**.** *Let $(\mathcal{X}, \mathrm{d})$ be a metric space, $m \geq 1$, and $F_j, \widehat{F}_j, H^\star, \widehat{H} : \mathcal{X} \to \mathcal{X}$. For $L_j, \epsilon_j \geq 0$, assume, for all $z, z' \in \mathcal{X}$,*

$$\mathrm{d}(F_j(z), F_j(z')) \leq L_j \mathrm{d}(z, z'), \qquad \mathrm{d}(\widehat{F}_j(z), F_j(z)) \leq \epsilon_j.$$

Define $F_{1:m} = F_m \circ \cdots \circ F_1$ and $\widehat{F}_{1:m} = \widehat{F}_m \circ \cdots \circ \widehat{F}_1$. For each $x \in \mathcal{X}$, set

$$\delta_{\mathrm{route}}(x) = \mathrm{d}(\widehat{H}(x), \widehat{F}_{1:m}(x)), \qquad \delta_\star(x) = \mathrm{d}(F_{1:m}(x), H^\star(x)).$$

Then, with an empty product equal to one,

$$\mathrm{d}(\widehat{H}(x), H^\star(x)) \leq \delta_{\mathrm{route}}(x) + \sum_{j=1}^{m} \epsilon_j \prod_{\ell=j+1}^{m} L_\ell + \delta_\star(x). \tag{26}$$

Remark (When route agreement is evidence). Here F_j are true edge maps and $H^\star$ is the true direct map; their hatted counterparts are predictions. In Eq. (26), δ_{route} compares two predictions, the middle term propagates local prediction error, and $\delta_\star$ compares the true route with the true direct action. Small route disagreement yields a useful endpoint bound when the other two terms are also small. For noncommuting interventions, the true-action defect can remain even with perfect edge predictors. This decomposition explains why the experiments evaluate routed endpoint accuracy as well as internal agreement; its proof uses triangle inequalities and Lipschitz propagation.

The proof only needs the assumptions on a common domain containing all true and learned intermediate states. Appendix A.7 tests each term and replaces uniform errors by a finite-cover estimate under additional regularity and coverage assumptions. The bound uses a metric, not squared error. For Euclidean states, squaring the full right-hand side and dividing by d bounds the coordinate MSE; squared error itself cannot be substituted for d in the triangle inequality.

For Proposition 4, linearity gives

$$\widehat{G}_x(\alpha^\star) - G_x(\alpha^\star) = \sum_j c_j \{\widehat{G}_x(\alpha^{(j)}) - G_x(\alpha^{(j)})\},$$

and the triangle inequality gives Eq. (25). If $\alpha^\star$ is outside the observed span, choose a linear functional ℓ that vanishes on that span and has $\ell(\alpha^\star) = 1$. For any unit output vector w, $\widehat{G}(\alpha) = G(\alpha) + M\ell(\alpha)w$ agrees on observed codes and differs by M at $\alpha^\star$.

If g is L-Lipschitz and $\sup_x \|g(x) - g^\star(x)\| \leq \epsilon$ on a common domain containing both trajectories from x over $[0, T]$, their distance $r(t)$ satisfies $r(t) \leq L \int_0^t r(s)\, ds + \epsilon t$. Gronwall's inequality yields

$$\left\| \phi_T^g(x) - \phi_T^{g^\star}(x) \right\| \leq \begin{cases} \epsilon T, & L = 0, \\ \epsilon(e^{LT} - 1)/L, & L > 0. \end{cases}$$

The inequality carries transition-span generator error to endpoint error; uniform control of the generator remains an assumption.

A.6 Proof of Theorem 6

Starting from $x_0 = \widehat{x}_0 = x$, define $x_j = F_j(x_{j-1})$ and $\widehat{x}_j = \widehat{F}_j(\widehat{x}_{j-1})$. Induction gives

$$\begin{aligned} \mathrm{d}(\widehat{x}_j, x_j) &\leq \mathrm{d}(\widehat{F}_j(\widehat{x}_{j-1}), F_j(\widehat{x}_{j-1})) + \mathrm{d}(F_j(\widehat{x}_{j-1}), F_j(x_{j-1})) \\ &\leq \epsilon_j + L_j \mathrm{d}(\widehat{x}_{j-1}, x_{j-1}) \\ &\leq \sum_{r=1}^{j} \epsilon_r \prod_{\ell=r+1}^{j} L_\ell. \end{aligned}$$

Inserting the learned and true route endpoints between $\widehat{H}(x)$ and $H^\star(x)$ gives Eq. (26). If all $L_j \leq L$ and $\epsilon_j \leq \epsilon$, the propagated term is at most $\epsilon \sum_{r=0}^{m-1} L^r$.

Scope of formal verification. The co-occurrence preorder, the necessary-and-sufficient support pattern, separation, singleton meets, hidden-support readout, the nonseparating shear, the forbidden-entry implication, and the scalar residual-to-angle step are formalized in `formalization/MPTIdentifiability.lean`. The metric-space recursion, zero-error and exact-action cases, transition-span inequality, scalar spectral-to-neural budget, bilinear stationary counterexample, observation conjugacy, and orbitwise equivariant-coordinate uniqueness are formalized in Lean 4.19 with Mathlib 4.19 in `formalization/MPTTheory.lean`. Neither file contains `sorry`, `admit`, or a custom axiom. The formalization does not cover neural optimization beyond the scalar counterexample, noisy meet selection, local-flow differentiation, the matrix projector perturbation argument, Gaussian concentration, Davis–Kahan, the smooth bump construction, or continuous-time Gronwall.

A.7 Finite-cover diagnostic

This diagnostic asks whether each term in the finite-map bound is needed and whether a finite-cover estimate is numerically useful. The full inequality has no violation in 13,824 heldout path–context cases. Removing propagated edge error breaks it on every constant case and 94.8% of nonlinear cases; removing route disagreement fails on 5.4% of nonlinear cases; removing true-action defect fails on every noncommuting oracle-route case. Each term is needed in these tests. The finite-cover bound, however, exceeds observed error by factors of 232–1,244.

Table 6: Heldout-context audit of Theorem 6. Values are means over 18 route–seed results per world; "cover/error" measures the numerical looseness of the finite-cover diagnostic.

World	Direct error	Route term	Edge term	Action term	Violations	Cover/error
Constant commuting	0.00397	2.80×10^{-8}	0.00541	0	0	512
Nonlinear commuting	0.01646	0.00419	0.02544	0	0	232
Noncommuting	0.00249	0.1246	0.00404	0.1250	0	1,244

Let S_j be an r_j-net of an edge domain, and suppose F_j and $\widehat{F}_j$ are L_j- and $\widehat{L}_j$-Lipschitz. If the maximum map discrepancy on S_j is $\bar{\epsilon}_j$, then inserting the nearest $s \in S_j$ yields

$$\mathrm{d}(\widehat{F}_j(z), F_j(z)) \le \bar{\epsilon}_j + (\widehat{L}_j + L_j) r_j.$$

Substitution into Theorem 6 gives a finite-cover bound. No evaluated point violates it, but its mean is 512, 232, and 1,244 times the observed error in the constant-commuting, nonlinear-commuting, and noncommuting worlds. Its Spearman correlation with path error is -0.205 in the nonlinear world. At this cover density, the bound is too loose to estimate prediction error usefully.

A.8 Support-incidence audit protocol

The audit tests three separate requirements: the supports must separate directions, coefficients must span each support, and observations must resolve local velocities. Positive and premise-removal cases use the same field family.

This audit follows the code's zero-based mechanism labels $0, 1, 2, 3$, corresponding to theorem indices $1, 2, 3, 4$. For seed $s \in \{0, \dots, 19\}$, it draws an orthogonal matrix $R_s \in O(4)$ and 32 standard-normal states $x \in \mathbb{R}^4$, then evaluates

$$y = R_s x, \qquad v_i(x) = R_s^\top e_i\big(1 + 0.25 \tanh(y_i)\big).$$

These fields commute because each transformed coordinate evolves independently. The positive family is $\{\{i, j\} : i < j\}$; each support receives twelve coefficient vectors with both active entries nonzero and active coefficient rank two. Stacking field evaluations over states embeds the generator dictionary in a 128-dimensional Euclidean space. For a support S_r, the top $|S_r|$ left singular vectors of its generator matrix define $\widehat{P}_r$. Mechanism i is estimated by the smallest eigenvector of $\widehat{Q}_i = \sum_{r: i \in S_r} (I - \widehat{P}_r)$.

With exact generators, the minimum line cosine over all 80 seed–mechanism cases is $1 - 4.0 \times 10^{-16}$, and every Q_i has a one-dimensional kernel with incidence gap 2. We then replace the generators

by RK4 finite differences $(\phi_t(x) - x)/t$ at durations $0.2, 0.1, 0.05, 0.025$. Mean relative generator error falls from 0.01384 to 0.001734 between the two endpoint durations, while the corresponding line angle falls from $0.1069°$ to $0.01293°$. None of the 320 cases violates $\sin\angle(\widehat{v}_i, v_i) \le 2\Delta_i/\gamma_i$. Maximum bound utilization is 0.2940: tighter than the square-root alternative, though still conservative.

Line angles compare flattened field values over the fixed probe states, not individual pointwise vectors. Relative generator error is the norm of the finite-difference error divided by the true generator norm, stacking all groups, coefficients, and probe states. Bound utilization is the observed line sine divided by its stated upper bound. Projector changes use the operator norm.

The nonseparating family is $\{01, 012, 013, 12, 13, 23\}$, with 01 abbreviating $\{0, 1\}$, and likewise for the other sets. Mechanism 0 never occurs without mechanism 1, and Q_0 has a two-dimensional kernel in every seed. The shear $I + e_1 e_0^\top$ changes the first line by mean sine 0.7051 while changing the observed projectors by at most 2.11×10^{-15}. The separating family detects the same shear, with minimum projector change 0.6789. Replacing each rank-two coefficient class by a rank-one ray makes all four prescribed intersections empty in every seed; forcing the algorithm to return a minimum eigenvector gives mean cosine 0.8628, but that vector is not identified by the data. A separate two-dimensional control compares the zero field with $g(x) = 2\pi Jx$, using the rotation matrix J from Appendix A.2. Their time-one endpoint MSE is at most 6.99×10^{-32}, even though finite-difference generator error ranges from 0.6012 at $t = 0.2$ to 0.07849 at $t = 0.025$.

For the hidden-mask test in Table 7, estimated lines are matched injectively to true lines to maximize total absolute cosine. Line recall is the fraction of true lines with a matched cosine at least 0.98; mean line sine averages the matched pairs, including any below that threshold. Membership is inferred when $\left\|(I - \widehat{P}_r)\widehat{u}_i\right\| \le 0.12$ for a unit estimated line $\widehat{u}_i$. Support accuracy is the fraction of correct member/nonmember decisions over all RK group–mechanism pairs after retaining only reliable matches; an unrecovered mechanism is predicted absent. It is not the fraction of groups recovered exactly.

Table 7: Direct identification tests over ten seeds. Controls remove one theorem premise; line sine is computed over matched recovered lines.

Setting	Line recall	Support accuracy	Mean line sine
Hidden masks, 64/class	1.00	1.00	0.00797
Hidden masks, 512/class	1.00	1.00	0.00283
Nonseparating, 512/class	0.75	0.875	0.00277
Rank deficient, 512/class	0.75	0.750	0.00295

A.9 Noisy-endpoint support audit

The noisy experiment tests the tradeoff predicted by Theorem 3: smaller duration reduces bias but amplifies endpoint noise, while weak excitation makes recovery harder. It uses the same six pair supports and twenty rotations, evaluated at eight states so that $H = 32$. Coefficients are Gaussian, with one active variance fixed at one and the other in $\{0.03, 0.1, 0.3, 1\}$. Independent endpoint errors give difference noise of scale $\sqrt{2}(0.003)/t$. The estimator receives noisy finite differences and support labels; true generators are used only to evaluate recovery and the bound's constants.

Table 8: Noisy finite-sample support recovery, 20-seed mean $\pm$ standard deviation. Duration is 0.1 and minimum coefficient variance is one.

n_r	Mean line sine	Standard deviation	Mean theorem bound
128	0.004448	0.000298	2.330
256	0.003182	0.000218	1.172
512	0.002326	0.000130	0.724
1,024	0.001682	0.000111	0.521
2,048	0.001310	0.000126	0.411

The sample-size log–log slope is -0.445. At $n_r = 1{,}024$, durations $t \in \{0.025, 0.05, 0.1, 0.2, 0.4\}$ give mean line sines 0.00613, 0.00296, 0.00168, 0.00272, and 0.00949. Error is larger at both short and long durations, consistent with the tradeoff between noise amplification and finite-time bias. At $t = 0.1$, minimum variances 0.03, 0.1, 0.3, and 1 yield 0.00569, 0.00336, 0.00226, and 0.00168. All sampled random-matrix events satisfy the stated inequalities, and neither the oracle nor finite-sample bound is violated. In six of fourteen settings (42.9%), however, the line bound is at least one and gives no useful restriction on the angle.

A.10 Spectral-to-neural bridge

Does better recovery at probe states translate into better neural fields away from those states? A ten-seed experiment gives the spectral estimator and neural MPT the same noisy duration-0.1 endpoints. It uses four randomly rotated nonlinear commuting fields, 16 probes, and all six pair supports. The learner sees support labels but not realized Gaussian coefficients. Each recovered line is normalized to unit root-mean-square field norm and fitted by a three-layer width-64 tanh MLP. Support-masked codes are initialized by least squares on the local-time columns. Fields and codes then receive 500 Adam updates through four-step RK4 predictions, with a penalty retaining the recovered probe values. Evaluation uses 128 new states and 128 new coefficient vectors.

Table 9: Same-observation spectral and neural recovery, ten-seed mean $\pm$ standard deviation.

n_r	Spectral line sine	Neural probe sine	Neural new-state sine
64	0.00645 ± 0.00030	0.01541 ± 0.00129	0.04417 ± 0.00729
128	0.00453 ± 0.00018	0.01527 ± 0.00120	0.04536 ± 0.00719
256	0.00326 ± 0.00009	0.01382 ± 0.00181	0.04303 ± 0.00490
512	0.00229 ± 0.00012	0.01468 ± 0.00150	0.04187 ± 0.00739

At $n_r = 512$, the heldout combined-generator cosine is 0.99899 ± 0.00036 and endpoint MSE is $(2.04 \pm 0.76) \times 10^{-5}$, compared with 0.00968 ± 0.00065 for the identity map. The cosine compares the true and learned combined fields stacked over new states, averaged over evaluation coefficients after scale alignment; unlike line cosine, its sign is retained. The mean of the per-seed identity-to-model endpoint-MSE ratios is 533; this comparison excludes a zero-field explanation on these transitions. The line-error log–log slope is -0.495 for the spectral estimator but only -0.035 for the fitted neural fields. Increasing the sample size improves spectral recovery, while neural approximation error remains nearly unchanged.

Table 10: Assumption removals in the neural bridge, ten-seed means. Fit is endpoint loss normalized by duration squared.

Setting	Spectral sine	Neural sine	Fit loss	Shear mask ratio
Separated, rich	0.00229	0.01468	0.00199	0.363
Nonseparating	0.179	0.193	0.00689	0
One mechanism unexcited	0.502	0.349	0.00508	0.491

The shear mask ratio is the RMS of transformed codes on forbidden entries, with zeros elsewhere, divided by the RMS of all transformed code entries. Both averages use all transitions and coordinates. For the nonseparating family, applying $I + e_1 e_0^\top$ to each trained dictionary and the inverse map to its codes preserves every mask and changes integrated endpoints by only 2.56×10^{-7} on average. The same transform violates the masks in the separating family. In the missing-excitation control, only the first coordinate of each pair varies. One mechanism is absent from every noiseless observation; changing its field leaves training unchanged, and recovery fails. Both controls have larger and more variable fitting losses than the fully excited reference. Non-identifiability follows from the exact shear and unobserved-coordinate constructions, independently of those losses. The successful runs show that spectral estimates can initialize neural transport, but they still use spectral anchors and known supports in fixed coordinates.

A.11 Lie-derivative and bracket diagnostics

These diagnostics check classical flow identities used to interpret preservation and order effects. They are separate from the support-identification result.

For a differentiable feature c, $\frac{d}{dt}c(\phi_t^v(x)) = \nabla c(\phi_t^v(x))^\top v(\phi_t^v(x))$. If the absolute Lie derivative is bounded by η along a trajectory, feature drift is at most $T\eta$. If it vanishes for every basis field, it vanishes for every linear combination.

For C^2 fields with bounded derivatives near the short trajectories, use the convention $[v, w](x) = Dw(x)v(x) - Dv(x)w(x)$, as in the implementation. Taylor expansion gives

$$\phi_\varepsilon^w \circ \phi_\varepsilon^v(x) - \phi_\varepsilon^v \circ \phi_\varepsilon^w(x) = \varepsilon^2[v, w](x) + O(\varepsilon^3),$$

since the constant, linear, and same-field quadratic terms cancel, leaving $Dw\, v - Dv\, w$. In the synthetic noncommuting system, the norm ratio between the observed order effect and the bracket prediction is 1.00031, 1.00069, and 1.00108 at $\varepsilon = 0.02, 0.05, 0.1$, averaged over five checkpoints. The ratio checks magnitude and does not test the sign convention. It tests this local expansion on the synthetic family, not arbitrary learned real-world fields.

B Additional Experimental Results

The additional studies test distinct explanations for the main results. Premise-removal controls construct cases where an omitted identification condition leaves ambiguity. Prediction controls test whether lower disagreement or leakage reflects accurate transport rather than an unchanged output, decoder smoothing, or variance collapse. Table 11 pairs each question with the outcome that would contradict it.

Table 11: Theoretical premise tests and empirical hypotheses, with their falsifying outcomes. The prediction hypotheses need not hold on every benchmark.

Claim	Direct test	Falsifying outcome
Covered, rich separating groups identify generator lines	Local-time recovery and premise-removal controls	Ambiguous lines despite verified theorem premises
Linear control gives the best missing-edge endpoints	Autonomous odd-control comparison	Nonlinear odd control gives lower endpoint error
Each bound term has a witness in the tested families	Delete each measured term	No violation after deleting that term in those tests
Route agreement predicts grounded composition	Compare with the true route endpoint	Lower disagreement but higher endpoint error
Endpoint transport recovers the physical generator	Heldout spectral perturbation	Accurate endpoints but unstable off-manifold response
Preservation is not only decoder smoothing	Give GEARS the same frozen SVD decoder	The residual non-response gap disappears
Consistency improves perturbation response	Five-split Norman control	Lower path error but worse DE20 and non-DE error
Variance collapse explains the response ranking	Full-pool mean objective	Variance recovers but the ranking persists
A direct flow is always the best cost–accuracy choice	Common-query direct/routed timing	A routed alternative is faster and more accurate

The main comparisons lead to different choices: odd control improves Shapes3D endpoints (Table 3), mean matching leaves the Norman response ranking unchanged (Table 27), and observed routes can improve accuracy or cost (Table 32). The protocols below show what is held fixed in each comparison.

B.1 Datasets and query construction

PDE fields. Burgers uses the official PDEBench data (Takamoto et al., 2022): 10,000 initial conditions shared across five viscosities, with terminal fields sampled on 128 spatial points. Contexts

are split 70/10/20% before training. The task maps between terminal fields at different viscosities, rather than forecasting their time evolution. The advection control is generated for this study: 2,048 initial fields contain random Fourier modes 1–16, normalized to unit spatial root-mean-square (RMS) amplitude. On the periodic coordinate $s \in [0, 1)$ at duration $\tau = 0.2$, the exact solution is $u_c(s) = u_0(s - c\tau)$ for speeds $c \in \{0, 0.25, 0.5, 0.75, 1\}$, where u_0 is the initial field. It uses 128 spatial points and the same context-split proportions. Its data are not drawn from PDEBench. Appendix B.11 gives the parameter values and training, validation, and test edges.

Shapes3D. The public dataset contains 480,000 64×64 RGB images with six labeled factors (Burgess & Kim, 2018). We sample 4,096 base contexts per seed and construct 17 nodes by shifting orientation, scale, object hue, and wall hue, leaving shape and floor hue fixed within each context. A 70/10/20% context split is shared across models. A reconstruction-trained encoder maps images to 16-dimensional states and remains frozen during transport fitting. Factor probes are separate supervised classifiers used to read out factor labels, not to train transport. Each of the 30 heldout edges is scored on 128 fixed latent contexts and 32 decoded-image contexts. Appendix B.9 gives model selection and representation checks.

Norman Perturb-seq. We use the public combinatorial perturbation screen of Norman et al. (2019) through the GEARS preprocessing (Roohani et al., 2024). Each split includes 24 observed single perturbations and 32/11/11 training/validation/test doubles. Gene selection and the 48-dimensional TruncatedSVD use training conditions only. All model outputs are evaluated in the same 2,048-gene space. The observations are condition-level cell populations; no individual pre/post-intervention pairing is supplied. Appendix B.12 specifies the five splits and GEARS protocol.

B.2 Baseline definitions and provenance

Table 12 separates three sources: published methods, our adaptations of them, and controls defined for this study. Neural ODE supplies the continuous-dynamics formulation; the zero and odd subtractions are our controls. Each ODE is integrated from $z(0) = x$ to $z(1)$ using RK4; FE-NODE then combines the basis displacements. Latent models use dense networks; PDE models use periodic convolutions except for FNO's spectral blocks. Model families have separate learned parameters.

In the table, h_θ is a learned vector-valued network and $A_\theta(x)$ a learned state-dependent matrix. The adapter matrix A and bias b are fitted separately for each method; c is an FE-NODE coefficient vector and ξ an NCF task-context vector. NCF's context is a model parameter, unlike the paired data contexts in Section 3.1. The means $\bar{z}_0$ and $\bar{z}_i$ are training control and single-perturbation latent means. Shared MLP is also called SharedConditionalMLP in the earlier experiments.

FE-NODE retains episodic support–query fitting and Gram-corrected ridge coefficients; NCF uses first-order context modulation, alternating weight/context updates, and nearest-context pools. Their graph adapters are our additions (Appendix B.10). GEARS uses the released `cell-gears` package and graph preprocessing. FNO retains spectral and pointwise blocks, with width chosen to match the parameter budget. Earlier pairwise MLP controls fit $x + h_{e,e'}(x)$ separately on each observed edge and return identity on an untrained pair. That fallback tests the consequence of having no cross-edge prediction rule. SharedAffine is the state-independent control $x + B\alpha$.

B.3 Metric definitions and aggregation

All methods in a comparison use the same queries, contexts, and evaluation coordinates. We first average within each query, then across queries; uncertainty is computed from seed or split averages as specified below.

Paired endpoints. For query set $\mathcal{Q}$, n_q contexts per query, and d scalar coordinates, let y_{qn} and $\widehat{y}_{qn}$ be the observed and predicted target for context n on query q. The reported endpoint error is

$$\mathrm{MSE}_{\mathcal{Q}} = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{1}{n_q d} \sum_{n=1}^{n_q} \|\widehat{y}_{qn} - y_{qn}\|_2^2 . \tag{27}$$

Table 12: Sources and implemented prediction rules. "Here" identifies a control defined for this study, not a separate published method. The $+\alpha$ suffix denotes our adapter fitted only on training edges.

Model	Source	Implemented rule
MPT	This paper	$\dot{z} = V_\theta(z)\alpha$; Eq. (1).
Shared MLP / Shared CNN	Here: conditional residual control	$T_\alpha(x) = x + h_\theta(x, \alpha)$. The CNN is the PDE version.
SharedLowRank	Here: control-linear residual control	$T_\alpha(x) = x + A_\theta(x)\alpha$; no ODE integration.
Conditional ODE	Neural ODE (Chen et al., 2018); conditioning here	$\dot{z} = h_\theta(z, t, \alpha)$; endpoint fitting only.
Zero-NODE	Here: zero-anchored ODE control	$\dot{z} = h_\theta(z, t, \alpha) - h_\theta(z, t, 0)$.
Autonomous zero	Here: autonomous control	$\dot{z} = h_\theta(z, \alpha) - h_\theta(z, 0)$.
Autonomous odd	Here: odd-control comparison	$\dot{z} = \frac{1}{2}[h_\theta(z, \alpha) - h_\theta(z, -\alpha)]$.
FE-NODE $+\alpha$	Ingebrand et al. (2024); adapter here	Independent ODE-basis displacements with ridge-estimated coefficients; $c(\alpha) = A\alpha$.
NCF $+\alpha$	Nzoyem et al. (2025); adapter here	First-order context self-modulation; affine query context $\xi(\alpha) = A\alpha + b$.
FNO	Li et al. (2021); task adaptation here	Residual one-dimensional FNO conditioned on the source field, spatial grid, and broadcast α.
GEARS	Roohani et al. (2024); official package	Gene- and perturbation-graph predictor, trained in gene space on the matched condition split.
Identity	Here: no-change reference	$T_\alpha(x) = x$.
Additive singles	Here: empirical additive reference	$T_\alpha(x) = x + \sum_i \alpha_i(\bar{z}_i - \bar{z}_0)$, using training single/control latent means.

Shapes3D uses encoder outputs for latent MSE and RGB pixels in $[0, 1]$ for image MSE. PDE fields use $y = (u - \mu_{\text{train}})/\sigma_{\text{train}}$, with channel statistics pooled over training contexts and all parameter nodes. Thus the main field MSE is error in standardized coordinates. The separately reported relative L^2 error averages $\|\widehat{u} - u\|_2 \,/\, \|u\|_2$ on raw fields; these are different normalizations.

Image factor changes. Frozen probes produce predicted factor classes $\widehat{f}_j$. Let L_j be the number of classes, A_q the factors changed by query q, and A_q^c the remaining labeled factors. The function d_j is the absolute class-index distance, using the shortest circular distance for hue and orientation under the implemented convention. For each context,

$$\begin{aligned} \text{target error} &= \frac{1}{|A_q|} \sum_{j \in A_q} \frac{d_j(\widehat{f}_j, f_j^{\text{target}})}{L_j - 1}, \\ \text{leakage} &= \frac{1}{|A_q^c|} \sum_{j \notin A_q} \frac{d_j(\widehat{f}_j, f_j^{\text{source}})}{L_j - 1}. \end{aligned} \tag{28}$$

These are mean absolute class errors, not squared latent errors. Joint success requires exact target classes on active factors and exact source classes on all remaining factors. We average contexts and then queries. Decoded factor scores use the frozen image evaluator in place of the latent probe.

Routes and movement. Routed endpoint MSE applies Eq. (27) to the composed prediction and the observed endpoint. Direct–route disagreement instead compares the two predictions; reported path MSE divides Eq. (9) by the state dimension. PDE movement averages $\|\widehat{u} - u_{\text{source}}\|_2 \,/\, \|u_{\text{target}} - u_{\text{source}}\|_2$. For Shapes3D query q, the current evaluator computes the per-sample RMS displacement and then averages:

$$m_q = \frac{1}{n_q} \sum_{n=1}^{n_q} \frac{\left\|\widehat{z}_{qn} - z_{qn}^{\text{source}}\right\|_2}{\sqrt{d}}.$$

The earlier Zero-NODE study instead used the query-level RMS $r_q = \{(n_q d)^{-1} \sum_n \left\|\widehat{z}_{qn} - z_{qn}^{\text{source}}\right\|_2^2\}^{1/2}$. Both are averaged across queries, but m_q and r_q are different statistics. The main odd-control comparison uses m_q for both models; the historical Zero-NODE result below retains r_q and is not directly compared with it. Neither image statistic is normalized by the true target movement.

Condition-mean expression. For heldout condition q, let $\widehat{\mu}_q$, μ_q, and μ_0 be the predicted, observed target, and control gene means. Let D_q contain the indices of the twenty largest entries of $|\mu_q - \mu_0|$, and N_q the least-changing half of the 2,048 genes. For $G_q = D_q$, N_q, or the full gene set, respectively, DE20, non-DE, and gene MSE equal

$$\frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \frac{1}{|G_q|} \sum_{g \in G_q} (\widehat{\mu}_{qg} - \mu_{qg})^2. \tag{29}$$

Gene subsets are fixed from observed condition means for evaluation and shared across methods. This magnitude-based DE20 rule is specified by our evaluation protocol; it is not a significance test for differential expression. Each condition mean is computed before scoring, so individual-cell MSE is not the primary endpoint. Delta Pearson is the Pearson correlation, across all selected genes, between $\widehat{\mu}_q - \mu_0$ and $\mu_q - \mu_0$. Opposite direction is the fraction of $g \in D_q$ for which $(\widehat{\mu}_{qg} - \mu_{0g})(\mu_{qg} - \mu_{0g}) < 0$; zero changes are not counted as opposite. Both are averaged across conditions. Norman's movement ratio uses latent condition-mean changes. Uncertainty is over ten seed means for the main image/PDE comparisons and five split means for Norman, each of the latter averaging three training seeds.

B.4 Synthetic prediction controls

The synthetic tasks separate accurate movement from low consistency or leakage scores. Equal-update comparisons include an unconstrained shared predictor and, where relevant, an explicit invariant penalty. The no-beta experiment below separately tests whether every node has been observed.

Table 13: Equal-update synthetic results. Lower is better except generator cosine.

World/model	Train MSE	Heldout MSE	Leakage	Path error	Generator cosine
Commuting MPT	2.99×10^{-7}	0.0805	5.55×10^{-4}	2.47×10^{-14}	1.0000
Nonlinear MPT, fit only	9.68×10^{-5}	5.38×10^{-4}	–	7.90×10^{-6}	–
Nonlinear Shared MLP	2.36×10^{-3}	9.09×10^{-3}	–	3.79×10^{-4}	–
Spurious MPT, invariant	0.0531	0.115	1.01×10^{-3}	–	–
Spurious Shared MLP	1.98×10^{-5}	0.156	0.382	–	–

The constant commuting no-beta-initialization experiment learns the codes rather than initializing them at ground truth. It separates graph coverage from gauge ambiguity. When every environment node appears in training and only selected edges are withheld, heldout MSE is 4.84×10^{-7} and aligned transition-generator cosine is 0.9999996. When one node is absent altogether, heldout MSE rises to 0.378 despite accurate recovery on the observed subspace. Linear alignment cannot repair a code that the training graph never identifies.

B.5 Optimization and representation boundary audits

The cold-start test asks whether neural fitting still succeeds without spectral initialization. It uses the same grouped noisy local-time endpoints as the spectral-to-neural experiment, supplies support masks, and withholds directions and realized coefficients. Randomly initialized ERM passes the prespecified line, generator, and nontrivial-endpoint criteria in 19/20 runs; scale normalization gives 20/20. Their endpoint MSEs are 5.38×10^{-5} and 9.03×10^{-5}, compared with 2.39×10^{-5} under spectral initialization. Most sampled starts succeed, while spectral initialization gives the lowest endpoint error. The stationary-point example in Appendix A explains why these observations cannot yield a guarantee from every initialization.

The observation-space test maps three translations through a smooth injective nonlinear function into 64 densely mixed pixel coordinates. Grouped local differences recover every pushforward action line and class membership over ten seeds, with mean line sine 0.00714. Masks, latent states, coefficients, encoder, and decoder are withheld. A non-componentwise latent diffeomorphism with its conjugated decoder still preserves pixel endpoints to 2.22×10^{-16}. The test recovers action directions in this synthetic observation space, but the latent coordinates remain ambiguous.

B.6 Earlier dSprites development study and fit-matched boundary

The dSprites sweep (Matthey et al., 2017) asks whether a leakage advantage persists at comparable predictive accuracy. It preceded the independent Shapes3D study. The initial validation-selected comparison favored MPT on leakage, but matching validation predictive error removed evidence for that difference. We retain both comparisons to distinguish the original selection rule from the stricter fit-matched test.

The preservation sweeps penalized changes in known invariant coordinates, using the same selection rule for MPT, SharedConditionalMLP, and the conditional ODE. The Shapes3D model uses endpoint fit alone; adding path regularization did not improve routed endpoint error. Section 4.5 and Appendix B.7 compare fit-only and path-regularized models.

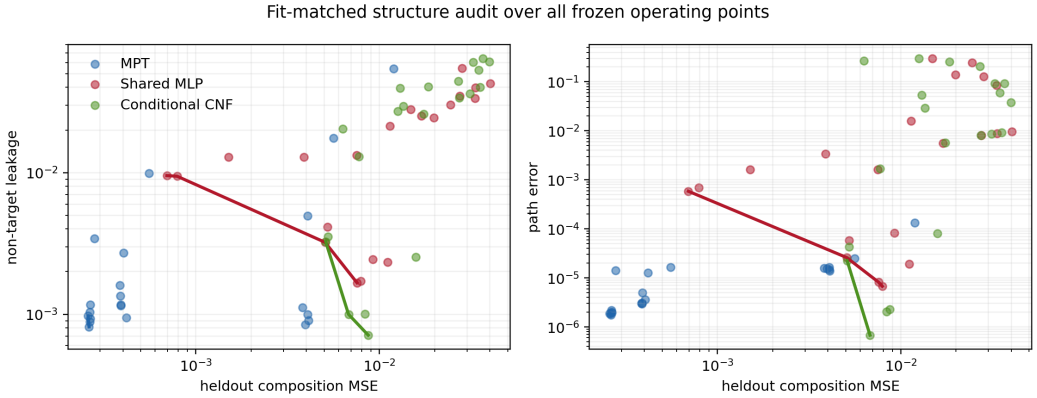

Figure 2: Complete dSprites fit–structure trajectories, averaged over ten paired seeds. Lines show each family's two-dimensional nondominated envelope; the four-objective audit also includes target orientation error.

At the aggregate level, an MPT operating point dominates all 21 Shared-MLP points and 19 of 21 conditional-neural-ODE points in heldout MSE, target error, leakage, and path error. Repeating the calculation within each seed gives mean dominance counts of 20.8/21 and 18.8/21. This envelope is descriptive because it inspects complete test trajectories; the strict comparison below selects a pair using validation metrics only.

Table 14: Test metrics for validation-matched dSprites operating points. Selection uses validation MSE and target error only.

Metric	MPT	Shared MLP	MPT wins	Exact p
Heldout MSE	5.55×10^{-4}	6.95×10^{-4}	9/10	0.0215
Target error (bins)	0.286	0.284	4/10	0.754
Non-target leakage	9.89×10^{-3}	9.53×10^{-3}	4/10	0.754
Path error	1.65×10^{-5}	5.85×10^{-4}	10/10	0.00195

The validation-matched pair has similar target error. MPT has lower heldout MSE in 9/10 seeds and lower path error in 10/10, but the leakage comparison does not detect a difference. No conditional-neural-ODE operating point meets both prespecified matching tolerances, so this matched comparison excludes that family.

The Pareto audit uses all 630 per-seed trajectory points: ten seeds, three families, seven regularization weights, and three milestones. The aggregate four-objective front contains three MPT and two conditional-neural-ODE operating points; Shared MLP contributes none. The two neural-ODE points obtain slightly smaller leakage or path error only at heldout MSE between 6.79×10^{-3} and 8.69×10^{-3} and target orientation error above 5.3 bins. The three MPT points have heldout MSE near 2.6×10^{-4} and orientation error near 0.21 bins.

The original validation-fit-constrained rule allowed a larger difference in predictive fit. It gave MPT versus Shared MLP heldout MSE 2.63×10^{-4} versus 4.60×10^{-3}, leakage 7.44×10^{-4} versus 1.21×10^{-2}, and path error 1.66×10^{-6} versus 2.51×10^{-2}, with 10/10 wins on each metric. That result combines accuracy and preservation differences. The stricter matching test does not establish a leakage advantage at comparable accuracy.

B.7 SHAPES3D TRANSITION STRATA AND ROUTED ENDPOINTS

Aggregate endpoint error can hide a difference between forward and reverse queries. Here forward means that all nonzero coordinates of α are positive, and reverse that all are negative. Single-factor queries have exactly one nonzero coordinate; multi-factor queries have more than one. These labels refer to code changes, including when a cyclic factor wraps. We compare each stratum and then compare direct predictions with compositions of observed maps.

Table 15: Shapes3D direct-transition strata for fit-only MPT and SharedLowRank.

Stratum	MPT MSE	LowRank MSE	Paired wins
Forward	0.001605	0.001996	10/10
Reverse	0.002072	0.007106	10/10
Single forward	0.001416	0.002183	10/10
Single reverse	0.001957	0.008035	10/10
Multi forward	0.001822	0.001782	5/10
Multi reverse	0.002203	0.006043	10/10

For long paths, MPT direct endpoint MSE is 0.002489 and its composed endpoint MSE is 0.002504. SharedLowRank improves from direct MSE 0.002682 to composed MSE 0.002195 by following observed unit edges. Its direct and composed predictions disagree more, but its composed prediction is closer to the target.

B.8 ZERO-PRESERVING CONDITIONAL ODE CONTROL

To test whether zero-change identity suffices without reversal or linear scaling, we define Zero-NODE by

$$\dot{z} = h_\theta(z, t, \alpha) - h_\theta(z, t, 0). \tag{30}$$

The subtraction sets the field to zero at zero control, while allowing nonlinear, time-dependent dynamics at other codes. Zero-NODE receives the same node difference as MPT and uses dense networks on Shapes3D and Norman and periodic convolutions on PDE fields. Autonomous-zero removes t; autonomous-odd uses Eq. (36) to impose sign symmetry as well. Each subtraction requires two network evaluations. Appendices B.9 and B.14 compare these choices.

Zero-NODE's parameter count differs from MPT by less than 0.2%; the autonomous controls differ by 0.31% on Shapes3D and 1.98% on Burgers. Update caps and training RK4 steps match. The Shapes3D autonomy comparison reuses historical MPT fits with a different learning-rate and clipping protocol, as detailed below.

Shapes3D uses seeds 100–109, frozen reconstruction-trained representations, three learning rates $\{10^{-4}, 3 \times 10^{-4}, 10^{-3}\}$, and 12,000 updates per candidate. Within each rate, observed-edge validation MSE selects the checkpoint; disjoint validation-chord MSE selects the rate. Target-factor error was prespecified as the primary endpoint of this follow-up, distinct from the original latent-MSE endpoint. Table 17 reports both. No target labels enter transport training or selection. The selected rate is 10^{-3} in all ten runs.

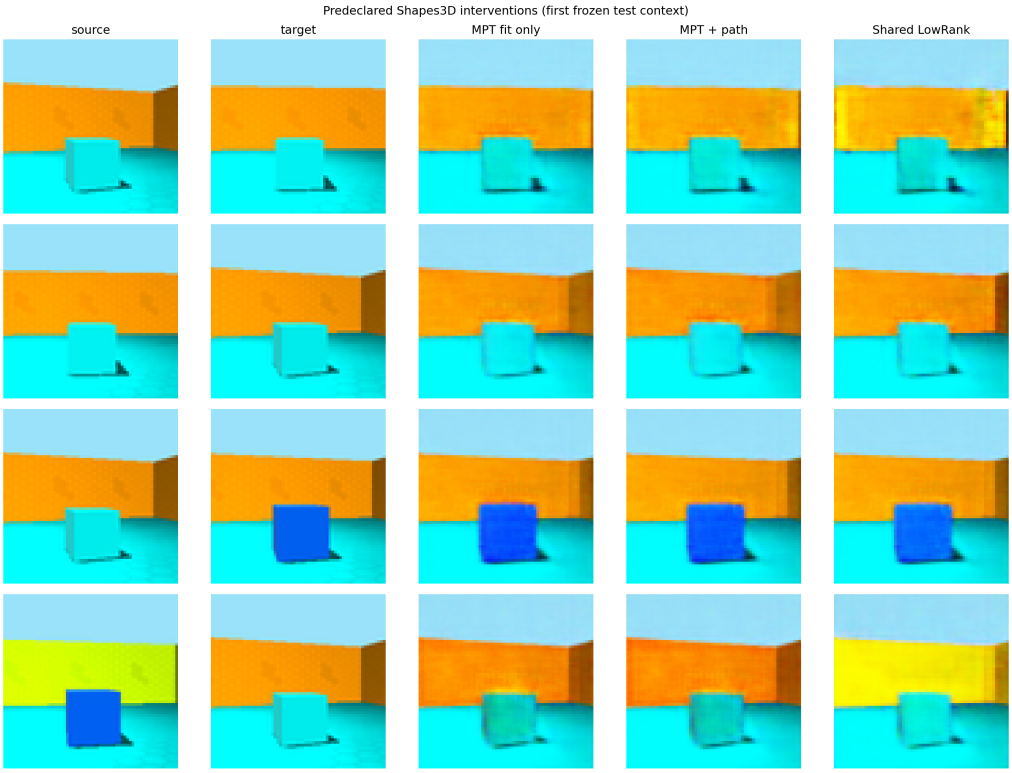

Figure 3: Predeclared Shapes3D interventions. Columns are source, true target, fit-only MPT, path-regularized MPT, and SharedLowRank. Rows use the first frozen test context; no example was selected after inspecting model outputs.

Table 16: Zero-preserving control budget. MAC ratios count static multiply–accumulate operations per sample and finite transport, excluding activation, memory, and solver overhead. Equal stage counts do not imply equal arithmetic cost or wall time.

Task	MPT parameters	Zero-NODE	Parameter ratio	MAC ratio
Shapes3D	49,984	50,036	1.0010	2.019
PDE	12,097	12,080	0.9986	1.998
Norman	100,992	100,918	0.9993	2.042

The control's observed-edge MSE is $(8.579 \pm 2.132) \times 10^{-5}$ and decoded image MSE is 0.004169 ± 0.000707. Its query-level latent RMS displacement is 0.11848 ± 0.00495, using the historical r_q aggregation in Appendix B.3. The original benchmark's joint target-and-preservation success rates are 0.716 for MPT, 0.541 for SharedLowRank, 0.080 for the unconstrained conditional ODE, and 0.069 for Shared MLP. The new control did not collect that joint statistic; leakage and target error in Table 17 are scored separately.

PDE controls reuse seeds 300–309, the frozen data and graph splits, and the same three learning rates as MPT. Every candidate receives 12,000 updates. Checkpoints minimize observed-edge validation MSE; rates minimize validation-chord MSE. Evaluation uses all test contexts and 64 RK4 steps. Zero-NODE's direct Burgers error is higher than MPT's in all ten seeds, while its mean routed error is lower (Table 2). Its 32- and 64-step direct MSE means are 1.868782×10^{-4} and 1.868784×10^{-4}.

Advection's four training edges all carry code 0.25. Zero-NODE fits their validation endpoints to mean MSE 4.228×10^{-5}, but its direct predictions have large errors on both larger positive shifts and reversals. The 32- and 64-step all-heldout direct MSE means are 2.879747 and 2.879955, so solver

Table 17: Shapes3D zero-preserving control. Entries follow the scale stated in each row. The control column is mean ± sample standard deviation across ten seeds. Intervals are descriptive paired t intervals, $\bar{d} \pm t_{0.975,9} s_d/\sqrt{10}$, for d = MPT − Zero-NODE; negative favors MPT. These are marginal, not simultaneous, intervals. Each metric has ten paired wins (exact two-sided sign $p = 0.001953$).

Metric (scale)	MPT mean	Zero-NODE	Difference 95% CI
Target factor (10^{-3})	17.785	36.580 ± 5.214	$[-20.735, -16.855]$
Latent MSE (10^{-3})	1.839	5.874 ± 1.473	$[-4.999, -3.071]$
Leakage (10^{-3})	0.05285	1.7545 ± 1.1939	$[-2.5508, -0.8526]$
Path error (10^{-5})	0.8056	28.8121 ± 4.9282	$[-31.6000, -24.4129]$

Table 18: Controlled advection: direct and routed errors for the zero-preserving control. Dispersion is the sample standard deviation across ten seeds. The all-heldout direct row is the primary endpoint; forward and reverse rows separate the four positive and ten negative queries.

Metric	MPT	Zero-NODE
All-heldout direct MSE	0.003216 ± 0.000224	2.87995 ± 1.97047
Forward direct MSE	0.006015 ± 0.000415	1.61870 ± 0.72240
Reverse direct MSE	0.002097 ± 0.000150	3.38446 ± 2.65935
Forward routed MSE	0.006036 ± 0.000416	0.008015 ± 0.003431

refinement does not close the endpoint gap. All ten advection runs select 10^{-3}; Burgers selects it in nine runs and 3×10^{-4} in one. Performance beyond this learning-rate range remains untested.

B.9 SHAPES3D AUTONOMY AND ODD-CONTROL COMPARISON

The Shapes3D comparison asks whether reversal helps when autonomy is held fixed, and whether linear scaling improves further. It evaluates the two fields in Eq. (36) on the frozen representations and graphs for seeds 100–109. All models receive the same node differences. The primary comparison is MPT versus autonomous-odd on latent MSE across 30 missing edges and 128 fixed contexts per edge. Image metrics use 32 contexts per edge. Table 3 reports seed means and sample standard deviations. Paired percentile intervals resample the ten seed-level differences 10,000 times with bootstrap seed 18018; secondary intervals are marginal, without multiplicity adjustment.

Each new autonomous arm has 49,831 parameters, versus 49,984 for MPT. Three rates, $\{10^{-4}, 3 \times 10^{-4}, 10^{-3}\}$, receive 12,000 updates each, with batch size 128 and eight RK4 steps, giving 60 completed candidates. The two new arms share initial network weights and candidate minibatch sequences within a seed. Observed-edge validation selects checkpoints; validation chords select rates. All twenty selected models use 10^{-3}. Each arm selects the final checkpoint in nine seeds and the 6,000-update checkpoint in seed 107; validation error is still falling at the budget limit in those nine runs.

The earlier MPT fits use the same update cap, batch size, training solver, and representations, but a fixed rate 3×10^{-4} chosen during development. Their clipping threshold is 10, compared with 5 for the new controls, and their initialization and sampling procedures differ. The comparison includes these optimization differences as well as control linearity. We retain the prespecified odd-control comparison and the original MPT fits, without selecting a replacement MPT model from test results.

The odd model has lower latent and decoded image MSE in every seed. The MPT-minus-odd differences are 1.09351×10^{-3} and 1.07895×10^{-3}, with paired 95% intervals $[0.81247, 1.37113] \times 10^{-3}$ and $[0.92733, 1.21285] \times 10^{-3}$. The mean difference has the same sign in every stratum (Table 19). Zero-NODE has low single- and multi-forward MSE, 0.703×10^{-3} and 0.434×10^{-3}, but large reverse errors, 11.902×10^{-3} and 10.332×10^{-3}. Its poor aggregate score is driven mainly by reversals.

Table 19: Shapes3D autonomy comparison by transition stratum. Latent MSE means are scaled by 10^{-3}. Wins count seed-level odd-control improvements over MPT, not independent edges.

Stratum	MPT	Autonomous odd	Odd wins
Single forward	1.4156	0.4967	10/10
Single reverse	1.9570	1.1605	9/10
Multi forward	1.8223	0.3352	10/10
Multi reverse	2.2034	0.9645	9/10

Odd control lowers target-factor error from 0.017785 to 0.009401; the paired difference interval is $[0.005075, 0.011826]$. Latent joint target-and-preservation success increases from 0.7164 to 0.8612 and decoded joint success from 0.7871 to 0.8742. Mean latent RMS displacement is 0.09728 for MPT and 0.10000 for odd control; this measures displacement magnitude, not a ratio to the true movement. Mean non-target leakage is lower for odd control, but its paired interval $[-2.222, 7.480] \times 10^{-5}$ includes zero, as does the decoded leakage interval. The target-prediction gain is supported; the leakage difference remains uncertain.

Earlier image controls. The original observed-edge MSE means were 5.89×10^{-4} for fit-only MPT, 5.61×10^{-4} for Shared MLP, 6.23×10^{-4} for the conditional neural ODE, and 4.70×10^{-4} for SharedLowRank. Shared MLP and the conditional ODE had heldout latent MSE 1.811×10^{-2} and 1.491×10^{-2}, leakage 1.23×10^{-2} and 8.47×10^{-3}, and image MSE 2.819×10^{-2} and 2.446×10^{-2}. MPT outperforms these controls and Zero-NODE, while autonomous-odd is more accurate than MPT. Figure 3 shows the original predeclared examples, before the autonomous models were trained.

B.10 Task-matched FE-NODE and Neural Context Flow controls

FE-NODE estimates basis coefficients from support observations without a gradient update (Ingebrand et al., 2024, Sections 3.1–3.2). NCF fits a context from task observations and also describes parameter-induced zero-shot adaptation via inverse regression (Nzoyem et al., 2025, Appendix A.2). Our main query supplies metadata but no query-edge observations. We compare training-edge metadata adapters under this restriction and separately measure how test-edge support changes prediction. The two regimes answer different questions and are not pooled into one ranking.

In the no-support regime, each method learns shared dynamics and a task code for each training edge. A zero-intercept ridge map sends the graph difference α to the FE-NODE coefficient; an affine ridge map sends it to the NCF context. The intercept accommodates the common translation of context coordinates. Adapters use training edges only, and disjoint validation chords select learning rates. Nothing is fitted on the requested test edge.

The implementations retain the defining components of the source methods. FE-NODE uses a learned neural-ODE basis, episodic support–query training, and Gram-corrected ridge coefficient estimation. NCF uses learned edge contexts, differentiable first-order Taylor self-modulation during training, alternating shared-weight and context updates, and the Nearest-First option for context pools. The graph adapters are the task-specific additions needed for zero-support prediction. Parameter counts range from 0.966 to 1.000 times the corresponding MPT count.

Finite-action parameterization. In the task-matched FE-NODE implementation, each basis field g_i is integrated independently from the same source state. With the column-vector convention for codes, prediction is

$$\Psi_i(x) = \phi_1^{g_i}(x) - x, \qquad c(\alpha) = A\alpha, \qquad \widehat{T}_\alpha^{\mathrm{FE}}(x) = x + \sum_i c_i(\alpha)\Psi_i(x). \tag{31}$$

The implementation approximates each basis flow by RK4 and then combines its displacement. Coefficients enter after integration, not inside a combined ODE. For the no-support linear adapter, the finite increment is linear in α; MPT instead integrates $V(x)\alpha$, producing the state-dependent interaction terms in Eq. (10). If coefficients were inserted before integration, a linear adapter could be absorbed into the dictionary as VA; that is a different parameterization from Eq. (31). The comparison also differs in episodic training and coefficient estimation, so it does not isolate integration

order alone. Test-edge adaptation replaces $A\alpha$ with support-estimated coefficients while keeping the same finite-action formula.

Table 20: No-test-edge-support closest-method comparison over ten common seeds. Values are mean MSE except Shapes3D leakage. Lower is better; MPT wins every paired direct-MSE comparison (two-sided sign $p = 0.001953$).

Benchmark	Metric	MPT	FE-NODE $+\alpha$	NCF $+\alpha$
Shapes3D	Direct latent MSE	1.839×10^{-3}	4.689×10^{-3}	1.708×10^{-2}
	Non-target leakage	5.285×10^{-5}	3.164×10^{-4}	1.730×10^{-2}
	Path disagreement	8.056×10^{-6}	9.062×10^{-4}	1.220×10^{-2}
	Decoded image MSE	2.099×10^{-3}	5.174×10^{-3}	2.786×10^{-2}
Burgers	Direct MSE	1.162×10^{-4}	1.001×10^{-3}	1.884×10^{-4}
	Routed MSE	1.162×10^{-4}	1.206×10^{-4}	1.255×10^{-4}
	Path disagreement	3.08×10^{-10}	8.520×10^{-4}	4.905×10^{-5}
Advection	All-edge direct MSE	3.216×10^{-3}	12.506	1.957
	Positive routed MSE	6.036×10^{-3}	4.455×10^{-4}	6.786×10^{-3}
	Reverse direct MSE	2.097×10^{-3}	12.724	1.999

The largest errors occur in direct prediction on missing edges. FE-NODE nearly matches MPT on routed Burgers MSE and gives the lowest positive advection route error, showing that its learned basis can predict local changes. Direct queries instead require the adapter to extrapolate from training-task coordinates to a missing chord or reversal. NCF is the closer direct comparator on Burgers; MPT lowers its error by 38.3%, with the same direction in all ten seeds. The advection comparison tests sign extrapolation: reverse codes lie well outside both adapters' observed task coordinates.

The second regime uses target examples from each reported test edge. FE-NODE re-estimates its coefficient from those examples, while NCF optimizes a new context with shared weights frozen. Table 21 keeps these rows separate because their information exceeds that available to no-support MPT.

Table 21: Closest-method few-shot adaptation. MPT receives no test-edge examples; FE-NODE and NCF receive the listed number per test edge.

		MPT	Adapted model: support size		
Benchmark	Model	No support	8	32	128
Shapes3D	FE-NODE	1.839×10^{-3}	3.780×10^{-3}	3.725×10^{-3}	3.701×10^{-3}
	NCF		8.325×10^{-3}	8.277×10^{-3}	8.233×10^{-3}
Burgers	FE-NODE	1.162×10^{-4}	4.065×10^{-4}	3.973×10^{-4}	3.940×10^{-4}
	NCF		1.677×10^{-4}	1.657×10^{-4}	1.661×10^{-4}
Advection	FE-NODE	3.216×10^{-3}	1.109	1.098	1.092
	NCF		1.857	1.845	1.843

For each seed, method, and benchmark, the main comparison evaluates three learning rates and checkpoints at 3,000, 6,000, and 12,000 updates. Observed-edge validation error selects a checkpoint within each rate; validation-chord error then selects the rate. The selected rates often lie on a grid boundary: the largest FE-NODE and NCF rate in 9/10 Shapes3D seeds, the largest NCF rate in 8/10 Burgers seeds, and the smallest FE-NODE and NCF rate in 9/10 and 10/10 advection seeds. Most selected checkpoints are at 12,000 updates, and several observed-edge losses are still decreasing.

To test sensitivity to this training budget, we continue selected runs at nearby learning rates (Table 22). Seeds were fixed in advance. Model weights and NCF contexts were warm-started, optimizer states were reset, and test edges were excluded from training and selection. Each original checkpoint competed with two continuations of at most 12,000 additional updates. Shapes3D and

Burgers used learning-rate multipliers 1 and 3; advection used 0.3 and 1 because its original choices were at the lower grid boundary.

Table 22: Training-budget and learning-rate sensitivity on three fixed seeds per benchmark. Entries are direct MSE without test-edge support. Extended-budget selection chooses among the initial fit and two warm-start continuations using validation chords only. The MPT row reuses the initial fits on the same seeds.

Benchmark	Control	MPT	Initial fit	Extended-budget selection
Shapes3D	FE-NODE $+\alpha$	1.684×10^{-3}	4.583×10^{-3}	4.633×10^{-3}
	NCF $+\alpha$	1.684×10^{-3}	1.419×10^{-2}	1.556×10^{-2}
Burgers	FE-NODE $+\alpha$	1.232×10^{-4}	1.157×10^{-3}	1.253×10^{-3}
	NCF $+\alpha$	1.232×10^{-4}	1.750×10^{-4}	1.929×10^{-4}
Advection	FE-NODE $+\alpha$	3.114×10^{-3}	12.565	12.565
	NCF $+\alpha$	3.114×10^{-3}	1.9602	1.9598

A continuation is selected in 14/18 method–seed units, while direct test MSE improves in 8/18. Relative to the initial fits, selected mean direct MSE changes by $+1.07\%$ and $+10.76\%$ for Shapes3D FE-NODE and NCF, by $+7.20\%$ and $+9.62\%$ on Burgers, and by 0.00% and -0.019% on advection. MPT remains lower in all 18 paired direct-MSE comparisons. This three-seed warm-start study addresses nearby training budgets and rates; the ten-seed fixed-protocol table remains the confirmatory comparison.

B.11 PDE FIELD PROTOCOL AND DIRECTIONAL DIAGNOSTICS

The Burgers protocol uses official PDEBench terminal fields (Takamoto et al., 2022) at viscosities 0.001, 0.002, 0.004, 0.008, and 0.01, indexed by nodes $0, \ldots, 4$ in that order. All 10,000 contexts pass initial-condition pairing checks before splitting; evaluation uses all 2,000 test contexts. Adjacent positive viscosity edges train the models, chords $0 \to 2$ and $2 \to 4$ select learning rates, and the four remaining longer forward edges test scale extension. This protocol contains no reverse Burgers queries. Reversal is tested with generated periodic advection: 2,048 paired contexts at speeds 0, 0.25, 0.5, 0.75, and 1, corresponding to displacements 0, 0.05, 0.10, 0.15, and 0.20 at duration 0.2. It uses the same node ordering, positive adjacent training graph, and validation chords, with all remaining positive and reverse edges held out. For either scalar parameter p_e, codes are $\beta_e = (p_e - p_{\min})/(p_{\max} - p_{\min})$ and normalized parameter distance is $|\beta_{e'} - \beta_e|$.

For each of seeds 300–309, every trainable family is run at three learning rates and checkpoints at 3,000, 6,000, and 12,000 updates. Selection uses validation-chord MSE with 16 RK4 steps; test evaluation uses 64 steps. MPT has 12,097 parameters. Relative parameter counts are 1.000 for SharedLowRank, 1.020 for the shared conditional CNN, 0.999 for the conditional neural ODE, and 1.009 for FNO. The shared conditional row was named "Shared MLP" in earlier experiment summaries; its implementation uses periodic convolutions. The objective is paired standardized-field MSE only.

On Burgers, MPT direct MSE is $(1.162 \pm 0.082) \times 10^{-4}$ across seeds, lower than every trainable baseline in all ten seeds. The routed ordering differs: MPT beats LowRank and Shared CNN in 9/10 seeds, the conditional neural ODE in 8/10, and FNO in only 2/10. Direct MSE rises with normalized parameter distance, from 8.58×10^{-5} at distance 0.667 to 1.47×10^{-4} at distance 1. MPT's direct and routed predictions remain nearly identical across this range.

The aggregate advection score combines seven forward and reverse strata, whose rankings differ. MPT forward direct, forward routed, reverse direct, and reverse routed MSE are 6.015×10^{-3}, 6.036×10^{-3}, 2.097×10^{-3}, and 2.099×10^{-3}. It beats every baseline in all ten seeds on forward direct and both reverse metrics. On forward routed MSE, it loses every seed to Shared CNN and LowRank and nine seeds to FNO, while beating the conditional neural ODE in every seed. Composing learned local maps works well on observed positive routes; the signed autonomous field is most useful for reversals.

For a paired source and target (x, y) with forward code α, internal cycle MSE scores $\widehat{T}_{-\alpha}(\widehat{T}_{\alpha}(x))$ against x; grounded reverse MSE scores $\widehat{T}_{-\alpha}(y)$ against x. Both average standardized coordinate errors over contexts and the four heldout forward queries. Their means are 2.44×10^{-8} and 3.88×10^{-3}, respectively.

High-frequency amplification is the mean of $\left\|\widehat{T}_{\alpha}(x+\delta) - \widehat{T}_{\alpha}(x)\right\|_2 / \|\delta\|_2$ over contexts and queries. The perturbation δ is a mode-31 sine wave on the 128-point periodic grid, scaled to RMS 10^{-3} in standardized coordinates; training signals use modes 1–16. Reverse amplification is 0.329. Forward amplification averages 11.23 and grows with shift length: 4.43, 9.60, and 21.30 at distances 0.5, 0.75, and 1. Exact translation has amplification one. This off-spectrum instability contrasts with the accurate benchmark endpoints and rules out recovery of the full translation operator.

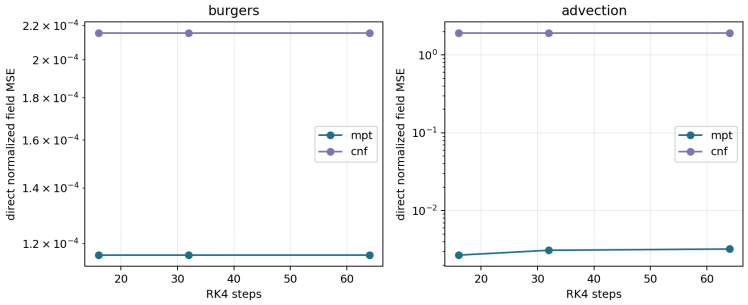

Figure 4: RK4 refinement for the PDE field neural ODE models. MPT's mean 32-to-64 relative change is 5.49×10^{-5} on Burgers and 3.63% on advection. The latter is below the prespecified 5% criterion but indicates greater solver sensitivity on advection.

B.12 Norman Perturb-seq across condition splits

This study tests whether response rankings persist across heldout combinations within one screen. The later objective comparison holds these splits fixed and removes the random-pair variance penalty.

Five-split study. The follow-up fixes split seeds $1401, 1411, 1421, 1431, 1441$ and training seeds $510, 511, 512$ before running. For each split, expression-based gene selection and TruncatedSVD fitting use only control, observed singles, and training combinations. Validation and test combinations do not enter this fit. MPT, SharedLowRank, Shared MLP, and Zero-NODE train on the same resulting coordinates, using learning rates $\{3\times10^{-4}, 10^{-3}\}$, batch size 128, and 6,000 updates. Validation condition-mean latent MSE selects among checkpoints at 1,500, 3,000, and 6,000 updates and then among rates. No DE20 test score is used for selection.

Training objective and selection. Let P_0 and P_q denote the empirical control and perturbed-cell distributions in the $d = 48$ latent coordinates, with intervention code $\alpha_q = \beta_q - \beta_0$. Each update chooses a training condition q uniformly, then samples $X \sim P_0$ and $Y_q \sim P_q$ independently with replacement. All trained transport arms in the fixed-split and five-split studies minimize

$$\mathcal{L}_{\text{cell}} = \mathbb{E}_q \mathbb{E}_{P_0 \otimes P_q}\left[\frac{1}{d}\left\|\widehat{T}_{\alpha_q}(X) - Y_q\right\|_2^2\right]. \tag{32}$$

This is pointwise MSE on random cell pairs, with no MMD, cycle, path, or identity loss. AdamW uses zero weight decay and gradient clipping at norm 5. The MMD objectives in Appendix B.17 belong to separate synthetic experiments.

For fixed q, put $U_q = \widehat{T}_{\alpha_q}(X)$ and $\mu_q = \mathbb{E}Y_q$. The independent sampling above yields the exact loss decomposition

$$\mathbb{E}\,\|U_q - Y_q\|_2^2 = \|\mathbb{E}U_q - \mu_q\|_2^2 + \operatorname{tr}\operatorname{Cov}(U_q) + \operatorname{tr}\operatorname{Cov}(Y_q). \tag{33}$$

Centering both variables and using independence makes the cross term vanish. The target covariance is constant in the model. The other terms reward both an accurate mean and low predicted covariance, even though evaluation scores the mean alone. Checkpoint and rate selection use mean latent error across validation conditions, with 256 fixed control cells and the full sampled condition mean. Evaluation uses 300 control cells. Appendix B.13 replaces random-pair MSE with full-pool mean matching to test whether this variance penalty explains the response ranking.

Evaluation transports 300 control cells and compares their predicted mean with the heldout condition mean. The frozen SVD decoder maps latent predictions to gene space. DE20 comprises the twenty largest absolute raw target-minus-control mean changes in the selected 2,048-gene panel; non-DE error uses the 1,024 smallest changes. These are response-magnitude rankings, not a new statistical differential-expression test. Every model is scored on the same genes for a given condition.

Table 23: DE20 by heldout-condition split. Each transport entry averages three training seeds; additive is deterministic. The last row averages the five split means.

Split	MPT	Zero-NODE	Shared MLP	SharedLowRank	Additive
1401	0.083656	0.066017	0.055612	0.071789	0.169407
1411	0.119464	0.065782	0.101046	0.092715	0.117312
1421	0.075368	0.064461	0.062743	0.062241	0.078757
1431	0.095949	0.061025	0.059810	0.072728	0.089194
1441	0.095007	0.086056	0.085104	0.082848	0.110836
Mean	0.093889	0.068668	0.072863	0.076464	0.113101

For a comparator b, we first average paired differences over the three training seeds within each split, then average those five effects. The prespecified interval resamples the five split effects 10,000 times (bootstrap seed 1601). Table 24 gives percentile intervals. The five overlapping test sets cover 35 distinct combinations from one screen, so these intervals describe sensitivity within the chosen split scheme rather than population-level biological uncertainty. With only five split effects, a unanimous direction gives exact two-sided sign-test $p = 0.0625$; the bootstrap intervals and sign tests answer different inferential questions at this sample size.

Table 24: Norman DE20 difference, MPT minus comparator, after averaging training seeds within each split. Positive favors the comparator. Intervals are marginal split-bootstrap intervals; no multiplicity adjustment is applied.

Comparator	Mean difference	Bootstrap 95% CI	MPT split wins
Zero-NODE	$+0.025220$	$[0.011471, 0.041376]$	0/5
Shared MLP	$+0.021026$	$[0.012695, 0.029357]$	0/5
SharedLowRank	$+0.017424$	$[0.012236, 0.023125]$	0/5
Additive singles	-0.019212	$[-0.053266, 0.002885]$	3/5

Zero-NODE also lowers gene MSE and non-DE error in all five split means. The MPT-minus-Zero-NODE non-DE difference is 7.230×10^{-5}, with split-bootstrap interval $[5.584, 9.670] \times 10^{-5}$. Path disagreement favors MPT instead: 0.01128 ± 0.00463 against 0.15969 ± 0.04576, with standard deviations across split means. Their movement ratios are 1.028 ± 0.066 and 1.011 ± 0.050. The nonlinear field is both more accurate and less internally compositional on this task.

One official GEARS run (seed 510) accompanies each split. Its native DE20 is 0.10898 ± 0.03196, gene MSE 0.009687 ± 0.003091, and non-DE error 0.005102 ± 0.002580. Passing its prediction through the corresponding frozen SVD decoder gives DE20 0.11590 ± 0.02608 and non-DE error 0.000707 ± 0.000231, keeping raw targets and gene rankings fixed. Projection again accounts for much of the native non-DE gap. These five runs are descriptive external references, not part of the paired training-seed comparison. The optional common-subspace score projects targets too and is a different endpoint from the raw-target tables.

Earlier fixed-split study. The original Norman protocol uses the official GEARS-preprocessed h5ad from the combinatorial CRISPRa screen (Norman et al., 2019). Eligible mechanisms are chosen by intervention availability; expression-based gene selection and the SVD fit use training condi-

tions only. The selected graph contains 24 single-gene mechanisms, 32 training pairs, 11 validation pairs, and 11 heldout pairs. Each condition contributes at most 768 cells; control contributes at most 2,048. Training draws control and perturbed cells independently, and evaluation compares condition means. No claim relies on a biological one-to-one cell correspondence.

Table 25: Earlier fixed-split transport comparison, retained separately from the five-split study. Entries are mean ± standard deviation over seeds 400–409; additive is deterministic. These standard deviations measure training variation within one split.

Model	DE20 MSE	Gene MSE	Non-DE error
MPT	$0.09474 \pm .01835$	$0.004290 \pm .000727$	$0.000464 \pm .000045$
Shared MLP	$0.06075 \pm .01090$	$0.003652 \pm .000383$	$0.000503 \pm .000050$
Conditional neural ODE	$0.09388 \pm .04515$	$0.005157 \pm .001903$	$0.000563 \pm .000115$
SharedLowRank	$0.07905 \pm .01092$	$0.005013 \pm .000483$	$0.000625 \pm .000054$
Additive singles	0.16941	0.008132	0.000652

Table 26: Earlier fixed-split, ten-seed MPT–GEARS comparison with post-hoc paired statistical reanalysis. Native GEARS remains in gene space; matched GEARS passes through the frozen 48-dimensional SVD decoder. Both p columns use matched seeds; tests are two-sided and unadjusted for multiplicity.

Evaluation	Metric	MPT	GEARS	MPT wins	Sign p	Paired flip p
Native	Gene mean MSE	0.004290	0.015986	10/10	0.00195	0.00195
Native	DE20 MSE	0.094736	0.078513	3/10	0.34375	0.05469
Native	Delta Pearson	0.873365	0.754186	10/10	0.00195	0.00195
Native	Opposite direction	0.002727	0.021818	10/10	0.00195	0.00195
Native	Non-DE error	0.000464	0.009987	10/10	0.00195	0.00195
Decoder matched	Gene mean MSE	0.004290	0.006757	10/10	0.00195	0.00195
Decoder matched	DE20 MSE	0.094736	0.112124	8/10	0.10938	0.04492
Decoder matched	Delta Pearson	0.873365	0.795668	10/10	0.00195	0.00195
Decoder matched	Opposite direction	0.002727	0.028636	10/10	0.00195	0.00195
Decoder matched	Non-DE error	0.000464	0.000973	10/10	0.00195	0.00195
Common subspace	Gene mean MSE	0.002739	0.005206	10/10	0.00195	0.00195

Paired statistical reanalysis. For each endpoint, d_s is the MPT-minus-GEARS difference at training seed s. The sign test counts positive and negative differences after excluding ties. The paired flip test enumerates all 2^{10} sign vectors and compares $|10^{-1}\sum_s \epsilon_s d_s|$ with $|10^{-1}\sum_s d_s|$. Its exactness assumes joint sign symmetry under the null; its smallest attainable two-sided p is $2/2^{10} = 0.001953125$. These are exploratory comparisons of training-seed variation on one fixed split, with no multiplicity correction.

The original collector instead pooled the twenty outcomes and enumerated $\binom{20}{10}$ unpaired label assignments. That calculation assumes pooled exchangeability and produced the 1.08×10^{-5} values in the earlier table. The paired reanalysis keeps seed matching throughout and leaves means, predictions, and selection unchanged. Native DE20 gives $p = 0.05469$, versus the original unpaired 0.01709; decoder-matched DE20 gives 0.04492, versus 0.03205. We report their mean directions without a general DE20 superiority claim. Both calculations and paired bootstrap intervals are retained in the statistical audit.

Top-DE and non-DE genes are selected from the raw heldout target delta and kept fixed in all three spaces, so projection does not change the gene ranking. Projecting GEARS reduces its gene MSE by 57.7% and its non-DE error by 90.3%, removing 94.7% of the raw preservation gap. A residual difference remains in every seed. Common-subspace gene MSE uses $\|\Delta z\|_2^2/d_{\text{gene}}$ for MPT, where Δz is the predicted-minus-target latent mean and $d_{\text{gene}} = 2{,}048$. GEARS predictions and targets are explicitly projected through the same frozen SVD basis for this score.

At the heldout-condition level, decoder-matched MPT wins gene MSE on 9/11 combinations, DE20 on 8/11, delta Pearson on 11/11, and non-DE error on 10/11. Its gene-MSE losses are `CBL+PTPN12` and `CEBPE+CNN1`. The former carries the largest top-response gap: MPT DE20 MSE is 0.2403 versus 0.0803 for matched GEARS. The latter is the sole condition on which matched GEARS also

has lower non-DE error. These cases explain the competing response and preservation rankings in the original fixed-split comparison. The newer Zero-NODE result improves both quantities, so this tradeoff is not a general property of the benchmark.

All ten GEARS runs use cell-gears 0.1.2 and the same condition split, SVD basis, and preprocessed graph. Independent rescoring reproduces the recorded condition metrics to maximum absolute error 1.2×10^{-6}.

B.13 NORMAN OBJECTIVE COMPARISON: FULL-POOL MEAN MATCHING

Could Zero-NODE's lower response error result from the variance penalty in random-pair training? We replace that objective by mean matching, retaining the five condition splits, three training seeds, and training-only representations of Appendix B.12. For the full pool of $N_0 = 2{,}048$ control states and empirical latent mean μ_q of training condition q, the objective is

$$\mathcal{L}_{\text{mean}} = \mathbb{E}_q \frac{1}{d} \left\| \frac{1}{N_0} \sum_{j=1}^{N_0} \widehat{T}_{\alpha_q}(x_j) - \mu_q \right\|_2^2 . \tag{34}$$

Each update samples a training condition uniformly. A two-pass calculation over chunks of 128 controls obtains the full-pool mean and its exact gradient; it does not replace Eq. (34) by a stochastic minibatch-mean loss. The explicit predicted-covariance term in Eq. (33) is absent. The objective constrains the mean, not the full cell distribution.

Each model–split–seed combination receives rates $\{3 \times 10^{-4}, 10^{-3}\}$ and 6,000 updates, giving 60 candidates and 30 selected fits. The representations, selection endpoint, and evaluation genes remain fixed. Historical random-pair checkpoints are rescored rather than retrained; the optional fresh-random arm was not run. Full-pool updates process more control cells than the historical random-pair batch of 128, so these objectives are not compared at equal FLOPs or wall time.

Let $E_{m,o,s}$ be DE20 for model m and objective o, averaged over the three training seeds within split s. The prespecified primary effect is

$$\Delta = \frac{1}{5} \sum_{s} \left[(E_{\text{MPT,mean},s} - E_{\text{Zero,mean},s}) - (E_{\text{MPT,pair},s} - E_{\text{Zero,pair},s}) \right]. \tag{35}$$

A negative value means that the objective change favors MPT relative to Zero-NODE. Paired percentile intervals resample the five split effects 10,000 times with seed 18018. They describe this overlapping split scheme, not independent biological replication. This bootstrap stream differs from the earlier random-pair analysis in Table 24.

Table 27: Norman objective comparison on DE20. Differences average three seeds within each split and then five splits. The interaction is primary; other rows are secondary paired contrasts. Intervals are marginal split-bootstrap 95% intervals.

Contrast	Mean difference	95% interval
Objective-by-model interaction Δ	-0.01082	$[-0.03080, 0.00686]$
MPT minus Zero-NODE, mean matching	$+0.01440$	$[0.00526, 0.02527]$
Mean matching minus random pair, MPT	-0.01830	$[-0.03938, 0.00929]$
Mean matching minus random pair, Zero-NODE	-0.00748	$[-0.01770, 0.00510]$

The mean DE20 improvement is 19.5% for MPT and 10.9% for Zero-NODE, but both within-model intervals include zero, as does the primary interaction. Under mean matching, Zero-NODE retains lower DE20 in all five splits. Its mean gene MSE is 0.003035 against 0.003871 for MPT, and non-DE error is 0.000311 against 0.000386; both paired difference intervals favor Zero-NODE. Mean delta Pearson correlation is 0.8989 versus 0.8919. Movement ratios are 1.0665 and 1.1420, respectively, so neither result arises from an identity predictor.

Removing the variance penalty restores much of Zero-NODE's output variation without changing the response ranking. Predicted variance remains below the target value; matching the mean has not established recovery of the cell distribution. No new GEARS model was trained for this objective comparison.

Table 28: Mean latent coordinate variance, averaged across test conditions, training seeds, and splits. The quantity is $d^{-1}\,\mathrm{tr}\,\mathrm{Cov}$, not total covariance trace. Source and target are empirical references.

Model	Random pair	Mean matching
MPT	0.3443	0.6893
Zero-NODE	0.00977	0.6771
Source controls	0.6794	
Target conditions	1.0216	

B.14 Burgers attribution: autonomy and control symmetry

Relative to MPT, Zero-NODE changes both time dependence and control linearity. To distinguish these choices on Burgers, we use the autonomous fields

$$f_{\mathrm{az}}(x,\alpha) = h_\theta(x,\alpha) - h_\theta(x,0), \qquad f_{\mathrm{ao}}(x,\alpha) = \tfrac{1}{2}\big[h_\theta(x,\alpha) - h_\theta(x,-\alpha)\big] \tag{36}$$

with no integration-time input. Both give exact zero-control identity, and f_{ao} is odd in α. Oddness gives inverse flows wherever they exist, but does not impose the scaling relation $f(x,c\alpha) = cf(x,\alpha)$. All four models receive the same node differences; this comparison concerns the action assigned to a code, not access to the code itself.

The follow-up fixes seeds 300–302 and reuses the external Burgers protocol in Appendix B: five viscosities, four adjacent training edges, two validation chords, and four heldout forward edges. Each test edge uses all 2,000 test contexts. The two new arms each have 12,337 parameters, compared with 12,097 for MPT and 12,080 for Zero-NODE. Every new arm–seed pair receives three learning rates, $\{10^{-4}, 3\times10^{-4}, 10^{-3}\}$, for 12,000 updates each, giving 18 candidates. Training uses endpoint MSE only, batch size 64, and four RK4 steps. Observed-edge validation at the training solver resolution selects checkpoints at 3,000, 6,000, and 12,000 updates; validation-chord MSE at sixteen steps selects the rate. Test evaluation uses 64 steps, with 16/32/64-step refinement. The two autonomous arms start from the same network weights for each seed; rates reset those weights and reuse the minibatch sequence.

Table 29 pairs the new arms with saved MPT and Zero-NODE runs on the same three seeds. These historical controls are not new fits, and the table is separate from the ten-seed comparison. Input-file checksums, test edges, seeds, and the context-split rule match the saved protocol. The standard deviations summarize these three seed-specific fits and data splits.

Table 29: Burgers autonomy follow-up, means $\pm$ sample standard deviations over seeds 300–302. MSE columns are scaled by 10^{-4}. Direct MSE is primary; routed MSE and movement are secondary. The first two rows reuse historical fits.

Model	Direct MSE	Routed MSE	Movement ratio
MPT (historical)	$\mathbf{1.0899 \pm 0.0326}$	1.0899 ± 0.0326	0.9681 ± 0.0114
Zero-NODE (historical)	1.3993 ± 0.2859	$\mathbf{1.0625 \pm 0.0629}$	0.9505 ± 0.0290
Autonomous zero	1.6552 ± 0.3142	1.0725 ± 0.0219	0.9623 ± 0.0892
Autonomous odd	1.5134 ± 0.2207	1.1149 ± 0.1144	1.0292 ± 0.0426

MPT lowers mean direct MSE by 34.2% against autonomous-zero and 28.0% against autonomous-odd. Each comparison favors MPT in all three seeds and on each of the four heldout edges within a seed (Table 30). These edge scores share a fitted model and test contexts, so they are not twelve independent replications. Three unanimous seed directions alone do not establish statistical significance: the exact one-sided sign probability is 0.125.

The controls learn useful local maps. Their mean selected observed-edge validation errors are 2.057×10^{-5} and 2.122×10^{-5}, versus 2.214×10^{-5} for MPT. Following the training route reduces autonomous-zero's error by 35.2% and autonomous-odd's by 26.3% relative to their own direct predictions. Autonomous-zero then has slightly lower routed error than MPT in all three seeds, by 1.6% on average. The direct gap is largest on the widest shift, $0 \to 4$: mean MSE is 1.376×10^{-4} for MPT and 2.306×10^{-4} for autonomous-odd. Accurate adjacent maps and explicit sign symmetry do not determine the best direct extension to this larger control.

Table 30: Direct Burgers MSE by seed, scaled by 10^{-4}. Model and rate selection use validation data, not the entries in this table.

Seed	MPT	Zero-NODE	Autonomous zero	Autonomous odd
300	1.1114	1.1163	2.0120	1.7188
301	1.1059	1.3934	1.5337	1.2800
302	1.0523	1.6881	1.4199	1.5414

Mean direct–route disagreement is 3.57×10^{-10} for MPT, 4.83×10^{-5} for autonomous-zero, and 4.52×10^{-5} for autonomous-odd. Burgers has one supplied control coordinate, so MPT's agreement follows from the scalar flow identities; it does not test commutation of multiple learned mechanisms. Endpoint error distinguishes this agreement from a trivial solution: identity incurs MSE 1.751×10^{-3}, about sixteen times larger, while MPT predicts nonzero movement. The task tests forward terminal-field shifts, not unknown-code identification or reverse PDE prediction.

All eighteen candidates complete, and all six selected models use rate 10^{-3} at 12,000 updates. Their observed-edge validation errors fall by a further 5.1–45.5% between 6,000 and 12,000 updates, so longer training may change the comparison. Solver refinement does not change the direct-error ordering: relative 32-to-64-step MSE changes average 6.56×10^{-5} for MPT and less than 1.8×10^{-6} for either new model. Numerical convergence at evaluation and convergence of training are separate checks.

B.15 Exploratory autonomy and odd-control pilot

The same controls were first tested in a short controlled-advection pilot: 128 periodic signals on 32 grid points, three Fourier modes, fixed horizon 0.2, and seeds 630–632. Four adjacent training shifts share code 0.25; one forward chord is reserved for validation. Five other forward chords and ten reversals are evaluated. Each model receives 600 updates at the single fixed rate 10^{-3}; checkpoints use observed-edge validation error. Training uses four RK4 steps and evaluation uses sixteen. Parameters number 897 for MPT, 886 for Zero-NODE, and 945 for each autonomous control.

Table 31: Exploratory controlled-advection pilot, three-seed mean $\pm$ sample standard deviation. Direct field MSE is primary. The routed column is a separate inference interface, not an additional training loss.

Model	Direct MSE	Routed MSE	Movement ratio
MPT	0.01241 ± 0.00236	0.01241 ± 0.00236	1.043 ± 0.018
Zero-NODE	0.28974 ± 0.02798	0.08932 ± 0.02748	0.900 ± 0.121
Autonomous zero	0.73975 ± 0.97775	0.22081 ± 0.28571	1.174 ± 0.176
Autonomous odd	0.06311 ± 0.03951	0.00767 ± 0.00239	0.974 ± 0.023

Removing time dependence alone does not close the direct-error gap. Odd control reduces it, but MPT remains better on direct endpoints in all three seeds. The odd model has lower routed error in two of three seeds and on average. Its mean observed-edge validation MSE is also lower, 0.00121 versus 0.00197. Reverse routes traverse the graph against its observed directions; they are not compositions of observed reverse training actions. A CUDA rerun of the same twelve fits reproduces direct MSE to maximum relative difference 6.84×10^{-6} from the local CPU run. It checks platform reproducibility, not additional seed evidence. The short pilot and the external Burgers follow-up are reported separately.

B.16 Direct and routed inference cost

The cost study asks whether direct prediction is preferable to an observed route at a given accuracy and latency. It uses saved Shapes3D fits on seeds 100/104/109 and Burgers fits on 300/304/309, distinct from the ten-seed tables and the Burgers autonomy subset 300–302. The measured methods are MPT, Zero-NODE, FE-NODE, and SharedLowRank. Each route is a shortest directed path using observed training edges only. Autonomous-zero and autonomous-odd are not timed.

Only 14 of the 30 Shapes3D queries admit such a route: eight single-forward and six multi-forward transitions. Their path lengths are two for nine queries, three for three, four for one, and five for one. All four Burgers queries are reachable, with lengths three, four, two, and three. Direct and routed cost comparisons use exactly these common query sets. Shapes3D accuracy here therefore measures a forward subset, not the all-edge endpoint of Table 3 or the separate long-path study.

The prespecified solver grid is 8/16/32/64 RK4 steps, with timing batches of 1 and 128. Pure-transport endpoint MSE is evaluated on all 128 fixed Shapes3D contexts or 2,000 Burgers contexts per edge, independently of timing batch size. Each latency is the median of five synchronized repetitions after two warmups, then averaged over queries within a seed. Table 32 gives means and sample standard deviations across three seeds. Repetitions and edges are not counted as independent fits. Measurements use an RTX 5090 with PyTorch 2.12 development CUDA 12.8 and Python 3.14.4. They exclude disk I/O, host-to-device transfer, and model loading. Static MAC counts in the artifacts cover linear and one-dimensional convolutional layers, not full runtime FLOPs; memory measurements retain the loaded comparison models.

Table 32: Cost on common directed-route queries, batch 128, eight RK4 steps for ODEs. Entries are three-seed mean $\pm$ sample standard deviation; latency is milliseconds per query batch, not per sample. SharedLowRank has no solver. Eight steps is a fixed grid entry, not selected on test MSE.

Model	Interface	Endpoint MSE	Latency (ms)
Shapes3D: 14 forward queries; MSE in 10^{-3}			
MPT	Direct	1.5641 ± 0.2300	10.775 ± 0.254
MPT	Routed	1.5669 ± 0.2283	27.512 ± 0.435
Zero-NODE	Direct	0.6287 ± 0.1759	7.122 ± 0.112
Zero-NODE	Routed	0.1931 ± 0.0622	18.523 ± 0.577
FE-NODE	Direct	1.3964 ± 0.0477	12.193 ± 0.536
FE-NODE	Routed	0.4312 ± 0.0729	30.984 ± 0.849
SharedLowRank	Direct	1.9991 ± 0.1074	0.163 ± 0.050
SharedLowRank	Routed	1.5295 ± 0.1180	0.333 ± 0.038
Burgers: all four queries; MSE in 10^{-4}			
MPT	Direct	1.2323 ± 0.1067	8.005 ± 0.075
MPT	Routed	1.2324 ± 0.1067	23.918 ± 0.156
Zero-NODE	Direct	1.8932 ± 0.8129	15.650 ± 0.255
Zero-NODE	Routed	1.0623 ± 0.1001	46.930 ± 0.836
FE-NODE	Direct	11.5709 ± 5.2245	31.278 ± 0.772
FE-NODE	Routed	1.2259 ± 0.0121	93.910 ± 2.154
SharedLowRank	Direct	16.1565 ± 2.0182	0.270 ± 0.001
SharedLowRank	Routed	1.3300 ± 0.0674	0.759 ± 0.010

On the forward Shapes3D subset, Zero-NODE's direct predictor is both faster and more accurate than MPT. SharedLowRank's route is also faster with slightly lower mean MSE. Including live image encoding at batch 128 gives direct latency 11.707 ms for MPT and 8.054 ms for Zero-NODE, retaining their ordering. These results agree with the directional analysis: the older all-edge advantage over Zero-NODE came mainly from reversals, which are absent from this cost subset.

On Burgers, MPT direct inference is about three times faster than its own route at nearly equal MSE. Its movement ratio is 0.9348, excluding a near-identity explanation. Zero-NODE's route is more accurate, while SharedLowRank's route is about 10.6 times faster than MPT direct inference at 7.9% higher mean MSE. The preferred predictor depends on the acceptable error and latency.

Across this grid, the largest relative change from eight-step endpoint MSE is below 0.08% in the three-seed mean for each task, model, and interface. Finer integration increases latency (Table 33) without materially changing the accuracy comparison. The experiment artifacts include batch-one measurements and per-query results.

Cached and live image encoding. Transport accuracy uses frozen cached latents; live-input timing also runs the encoder. Replaying its original contiguous batches of 512 reproduces the cache exactly on all fifteen source environments for each cost seed. Batches of 1 and 128 give maximum latent-coordinate differences 2.57×10^{-4}, 2.89×10^{-4}, and 3.05×10^{-4} across seeds. We score the cached and live-input pipelines separately to measure the effect of these floating-point differences.

Table 33: Pure-transport latency across the RK4 grid at batch 128, in milliseconds. Entries average the same three seed-level query means as Table 32. SharedLowRank is unchanged by this grid.

Model	Interface	8 steps	16 steps	32 steps	64 steps
Shapes3D, common 14-query subset					
MPT	Direct	10.77	21.82	44.71	85.02
MPT	Routed	27.51	57.08	112.65	218.71
Zero-NODE	Direct	7.12	14.56	28.25	58.08
Zero-NODE	Routed	18.52	36.12	73.24	144.75
FE-NODE	Direct	12.19	23.77	47.11	94.25
FE-NODE	Routed	30.98	60.53	120.46	241.65
Burgers, all four queries					
MPT	Direct	8.01	15.93	31.75	63.66
MPT	Routed	23.92	47.73	95.26	190.87
Zero-NODE	Direct	15.65	31.23	62.23	124.50
Zero-NODE	Routed	46.93	93.76	186.61	374.79
FE-NODE	Direct	31.28	62.02	122.83	245.60
FE-NODE	Routed	93.91	185.58	369.08	738.93

For each batch, we compare source encodings and transported predictions on matched contexts with the same cached targets. Live-pipeline accuracy uses the first one or 128 contexts, according to batch size, and is reported separately from full-context transport accuracy. For MPT, the mean ratio of prediction RMSE to source-encoding RMSE is 1.013 in both batches. The largest relative endpoint-MSE change is 0.655% at batch one and 0.0497% at batch 128. At 64-step direct inference over all 30 queries, batch-128 seed-mean changes are $+0.00284\%$, -0.00201%, and -0.00644%. Neither batch size changes the seed-mean ordering of the four measured methods. These checks measure the effect of numerical drift for the earlier models; they do not include the new autonomous controls.

B.17 Unpaired boundary

The synthetic unpaired studies ask whether distribution matching and consistency determine a correspondence. They use maximum mean discrepancy (MMD) with a radial basis function (RBF) kernel (Gretton et al., 2012), optionally adding cycle, path, and identity losses, rather than Norman's random-pair MSE. The full objective improves on MMD alone in the original latent commuting benchmark. Under partial observation, however, different correspondences satisfy both distributional and consistency constraints. Those objectives leave the pairing ambiguity unresolved.

C Reproducibility

The experiment records include splits, configurations, dataset checksums, source fingerprints, checkpoints, and evaluation summaries; GEARS records also include gene-space predictions. Model selection uses validation data. Representative Shapes3D and PDE reruns reproduce their saved metrics, and rescoring the earlier Norman predictions agrees to maximum absolute error 1.2×10^{-6}. In the five-split follow-up, reevaluating all sixty trained transport checkpoints and fifteen additive evaluations on CPU reproduces DE20 to maximum absolute difference 4.02×10^{-8}. PDE follow-up summaries are independently reaggregated from per-edge records.

The autonomy, objective, and cost studies include 31 completed evaluations and 120 new training candidates. Their seed-level and split-level summaries, paired intervals, and recorded validation selections are independently recomputed from saved metrics. For these studies, verification uses the recorded metrics rather than a fresh evaluation of every checkpoint; the available result archive does not contain all model weights or raw data. The code supplies entry points for the identification, image, PDE, and single-cell protocols.