BETTER RECONSTRUCTION, WORSE TRANSFER: PROBING REPRESENTATION GEOMETRY IN CROSS-REYNOLDS GENERALISATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural PDE surrogates degrade sharply when the governing physical parameter shifts. How the degradation distributes across the components of the pipeline has not been measured, so improvement effort is allocated by intuition. We address this on a zero-shot cross-Reynolds task in forced 2D Navier-Stokes, using a retrieval predictor as a diagnostic instrument, not because it is competitive, but because it is decomposable: its encoder, database, selection rule and autoregressive loop permit nested diagnostic substitutions. In seed 42, these give a 16.2% encode–decode reference, a 5.7-point reference-to-refreshed-oracle gap and a 14.0-point state-refresh gap; neither contrast distinguishes error injection from dynamical amplification. Holding representation and output path fixed, retrieved increments transfer better than the two tested learned-transition implementations, though by smaller margins than the equal-width PCA–ConvAE contrast; this does not make the increment mechanism irrelevant. Representation interventions then decouple reconstruction from transfer. A random projection of equal width yields 99.42% error, level with its own 99.27% reconstruction reference error, so reducing dimension is not by itself sufficient. A linear encoder of equal width reconstructs the target regime better (10.99% against 16.20%) yet transfers worse (41.93% against 38.30%, paired CI [3.26, 4.00]), a reversal that holds for all three autoencoder training seeds, so reconstruction fidelity does not determine transfer ordering. Widening that linear family from 64 to 512 improves the reconstruction reference error 3.16-fold while leaving transfer error unimproved, so capacity does not determine it either. Tested nonorthogonal invertible linear reparameterisations preserve reconstruction to numerical tolerance but increase mean deployment error by up to 4.75 points, showing that transfer depends on the retrieval coordinate frame even when reconstruction is held fixed. On the physical-output route, a fourth encoder trained with no reconstruction objective transfers 11.92 points worse (paired CI [11.30, 12.54]) while passing its encoder collapse diagnostics, so a non-reconstruction objective is not by itself sufficient either. Eight further candidates have mixed outcomes, including beneficial database expansion, a nonsignificant curvature-proxy association and an untested phase-alignment control. The established result is a decoupling between reconstruction fidelity and deployment ordering; organisation is a suggested explanation still to be identified. All claims are scoped to this benchmark and a tenfold parameter shift. We do not present a method: no tested diagnostic variant beats its canonical baseline.

1 INTRODUCTION

When a predictor moves to a different physical regime, a larger rollout error identifies a failure without locating it. Representation, available source transitions, neighbour selection, and feedback can each change the outcome, so improving the most visible component need not address the largest discrepancy. We use a retrieval predictor as a diagnostic instrument, not because it is competitive, but because it is decomposable.

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.

The instrument exposes which query is supplied and which source increment is borrowed. Holding the remaining pipeline fixed while changing those choices makes it possible to compare diagnostic levels on the same target trajectories. Our question is how those measured gaps and representation interventions constrain explanations of transfer on one benchmark; a strong method ranking alone would not answer it.

The seed-42 ladder moves from a 16.1953% encode–decode reference to 21.8518% with a refreshed true latent base and direction-oracle increments, then to 35.8606% using the same oracle indices without state refresh, and finally to 38.3001% in deployment. The gaps are 5.6565, 14.0088 and 2.4395 pp: conditional comparisons, not independent error sources or a universal selection limit.

Representation interventions then break several attractive explanations. Equal-width PCA reconstructs better but transfers worse than ConvAE by 3.6317 pp [3.2635, 4.0007], and widening PCA from 64 to 512 improves the reference error 3.16-fold without improving deployment. Yet the JEPA intervention trained without a reconstruction objective also transfers worse on its comparison route, by 11.9226 pp [11.3002, 12.5385]. We separately operationalise organisation as the retrieval coordinate frame and intervene on it in Section 5.4; this tests coordinate sensitivity without identifying a unique responsible property.

U-Net's lead is 3.6725 pp at $\alpha = 0$, 3.5314 pp at $\alpha = 1$ and 2.9610 pp at $\alpha = 2$. The smaller-than-3.62 expectation is satisfied, but this small reduction does not establish a substantial change. These metric-specific comparisons do not establish a metric-independent account of physical fidelity.

This is a diagnostic study and does not introduce a method. Its measurements are exploratory follow-up diagnostics, with prior expectations and the recorded stopping decision retained. The scope is the tested representation and predictor choices under the stated viscosity shift.

Our primary contribution is the decoupling of reconstruction fidelity from cross-Reynolds deployment ordering, together with a fidelity-preserving intervention on retrieval coordinates: across seven tested frames, deployment changes by up to 4.75 pp while reconstruction is preserved to numerical tolerance. Three supporting diagnostics are the implementation-resolved nested oracle ladder, eight candidate explanations with mixed evidence rather than eight exclusions, and the paired spectral-norm evaluation audit. The four representation interventions retain the opposite-direction fourth result; the incomplete spectral-fidelity observation in Section 7.1 is not a delivered principal empirical contribution.

2 Related work

Operators and autoregressive predictors. Fourier neural operators use spectral representations (Li et al., 2021); DeepONet and operator learning offer function-space formulations (Lu et al., 2021; Kovachki et al., 2023). Convolutional encoder–decoders, gated recurrent units and message-passing solvers (Ronneberger et al., 2015; Cho et al., 2014; Brandstetter et al., 2022b) are comparative references here: we diagnose a fixed predictor's errors rather than propose an architecture.

Distribution shift and PDE evaluation. PDE benchmarks foreground equation settings and evaluation protocols (Takamoto et al., 2022; Gupta & Brandstetter, 2023; Koehler et al., 2024); weak baselines and reporting biases have also been audited (McGreivy & Hakim, 2024). Scale-consistent learning studies zero-shot cross-Reynolds prediction using rescaled-domain training consistency and explicit scale conditioning (Li et al., 2025). Its autoregressive and space–time Navier–Stokes evaluations use regime-dependent resolutions and temporal histories (Tables 2 and 6), rather than our fixed retrieval instrument and context protocol. These are protocol differences, not a controlled numerical comparison. Context-informed dynamics and learned fluid discretizations address generalization through other mechanisms (Kirchmeyer et al., 2022; Kochkov et al., 2021). We claim no unique zero-shot precedent.

Analog prediction and latent dynamics. Analog forecasting and retrieval predict from observed states or historical patterns (Lorenz, 1969; Lguensat et al., 2017; Han et al., 2025), while latent world models learn state evolution (Hafner et al., 2019; 2025) and joint-embedding objectives predict representations rather than reconstruct pixels (Assran et al., 2023; Bardes et al., 2024; Assran et al.,

Table 1: Reference predictors on the shared $N = 500$ target evaluation. FNO and U-Net aggregate errors are reported in Shi (2026), Table 2. Uncertainty follows each method's reporting convention rather than a homogenized definition.

Predictor	Parameters	RelErr (%)	Uncertainty; seeds
Persistence	0	61.08	not reported
FNO	465,717	46.68	$\pm$0.11; 3
PCA relay, $d = 36$	0	42.250317	not reported
kNN-copy	0	41.012695	not reported
ConvAE-GRU	1,706,081	41.23	SD 0.39; 3
RSSM-style	5,274,337	41.00	SD 0.02; 3
ConvAE-Relay, latent	1,606,561	38.343243	see Appendix F; 3
ConvAE-Relay, physical	1,606,561	38.92515182	seed 42
U-Net	7,768,001	34.72	$\pm$0.60; 3

GRU and RSSM intervals are [40.91, 41.54] and [40.60, 41.39], respectively. FNO and U-Net's $\pm$ values are labeled SD in Shi (2026), Table 2, but unspecified in the retained result records.

2025). Retrieval exposes query and selection interventions; comparisons across latent encoders cannot isolate a learned-versus-retrieved effect.

Neighbour geometry and rollout fidelity. Nearest-neighbour concentration results require distributional conditions (Beyer et al., 1999; Durrant & Kabán, 2009), not ambient dimension alone in correlated fields; long-rollout and spectral diagnostics extend single-training-step evaluation (Lippe et al., 2023; Gao et al., 2026). We therefore report query geometry, stepwise error and norm sensitivity together, preserving each diagnostic's scope (Appendix I).

3 Setting and the diagnostic instrument

3.0 Exploratory follow-up, not confirmation

The diagnostics are *exploratory follow-up*, not confirmatory experiments, with known benchmark outcomes in view. Prior expectations and measured outcomes appear in Appendix G; diagnostic rules and analytic checks appear in Appendix H.

No multiplicity correction is applied. Paired bootstrap intervals are descriptive, and all diagnostics are exploratory rather than confirmatory tests; Table 1 retains the reference predictors' distinct uncertainty conventions.

3.1 Benchmark and evaluation protocol

We evaluate on 500 target-regime trajectories, predicting frames 10 to 19 from frame 9; Appendix A gives the full evaluation protocol, array shapes and frame ranges.

Our primary statistic is the mean relative field error,

$$\mathrm{RelErr} = \frac{100}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} \frac{\|\widehat{\omega}_{n,t} - \omega_{n,t}\|_2}{\|\omega_{n,t}\|_2 + 10^{-10}}, \qquad N = 500, \quad T = 10. \tag{1}$$

Intervals are paired bootstrap intervals over the shared evaluation trajectories (Efron, 1979); Appendix A states the resampling count, seed, pairing unit and seed-averaging convention. The intervals condition on the trained models, not an unrestricted population of training seeds.

Table 1 reports 34.72% for U-Net; the retained stepwise record in Appendix C gives 34.668555%. Both records are reported, but their protocol difference remains unreconciled: checkpoint, prediction-cache, sample-set and normalization provenance have not been isolated.

3.2 Physical and dynamical time

Source and target share snapshot spacing, not turnover time; Appendix A gives both turnover proxies and the unmeasured boundary case.

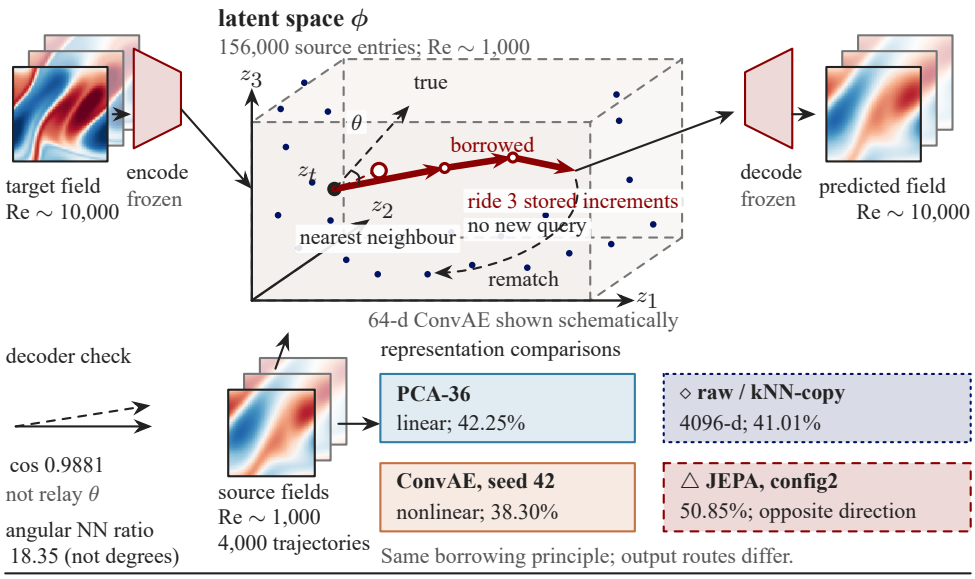


Figure 1: Matching and borrowing in latent space (F1). Real fields illustrate the pipeline; the latent volume and arrows are shown schematically. Increment angle, decoder check and angular nearest-neighbour ratio are distinct measurements. Chips retain their recorded output routes (raw kNN-copy: $N = 500$); JEPA is config2.

3.3 Two output routes

Figure 1 shows the retrieval predictor; Appendix A states the matching rule, ride length and decode cadence.

Figure 1 locates encoding, retrieval, borrowed increments and decoding in the diagnostic instrument.

3.4 Resolution audit

The specified spectral diagnostic classifies source and target as sufficient and the boundary regime as questionable. Appendix A gives the ratios, criterion and scope.

4 Nested oracle comparisons for transfer error

4.1 Diagnostic levels and their operations

The operations table defines the diagnostic levels. Their differences are not independent additive error sources: relative L_2 is a ratio of norms. The reconstruction reference evaluates the encoder-selected latent point, not an optimum over decoder outputs (Appendix F).

The direction oracle uses the true target physical increment without query screening, with identical indices in both levels. The 14.0088 pp contrast measures state refresh, not erroneous-query neighbour selection; the 5.6565 pp gap is conditional on that oracle and output route.

Both oracle levels use every-step selection. Canonical selection with state refresh was not run; the cadence control in Section 4.2 changes rematching without refreshing the base. These contrasts do not distinguish accumulated increment error from dynamical amplification.

The ladder plot is retained in Appendix F, Figure 11; the per-level operation table remains here.

Per-level operations

Level	Query and latent base	Selection / candidate set	Cadence and output
Reference	True scored state; freshly encoded	None; no candidates	Each scored frame: $D(E(x_t))$
Refreshed oracle	No latent query; base $E(x_t)$ reset each step	True physical-increment direction cosine; all 156,000	New oracle index each step; $D(E(x_t) + \Delta z_i)$
Free oracle	No latent query; accumulated z_t, initially $E(x_9)$	Same direction-oracle indices and full candidate set	Same every-step indices; $D(z_t + \Delta z_i)$
Canonical	Generated latent cosine query; accumulated z_t	Latent cosine top-1; all 156,000	$N_{\mathrm{RIDE}} = 3$, boundary rematch; decoded latent update

4.2 Attained selector improvements and cadence

Two oracle objectives attain different improvements over the canonical selector: physical increment-direction cosine gives 2.439408 pp [2.273871, 2.604401], whereas the magnitude-labelled oracle, which minimizes normalized squared physical increment mismatch rather than maximizing magnitude, gives 3.412029 pp [3.256380, 3.570465]. Neither optimizes final decoded trajectory error. These attained improvements do not limit an optimal selector's gain: $L^* \leq L_{\mathrm{dir}}$ implies $L_{\mathrm{base}} - L_{\mathrm{dir}} \leq L_{\mathrm{base}} - L^*$. Changing the objective changes the achieved gain; it does not establish a performance limit.

Both oracles choose an increment every step, unlike the canonical three-step riding cadence. A new same-code control gives $N_{\mathrm{RIDE}} = 1$ error 38.2660206%, versus 38.3000470% at three steps. The signed cadence improvement is 0.0340265 pp [-0.0531442, 0.1174261]; the same-cadence direction-selector improvement is 2.4053815 pp [2.2435583, 2.5694817]. Their means sum to the historical 2.439408 pp comparison. This decomposes that attained contrast, not every selection strategy's potential. The rerun agrees with the diagnostic cache, which differs slightly from the recorded 38.300117% deployment; both records and bootstrap conventions are retained in Appendix F.

The five $N = 500$ stepwise curves remain in Appendix C, Figure 7.

5 What the representation changes

Figure 3 compares reconstruction and deployment.

5.1 Distance from the source representation

The target-to-source ratio of angular nearest-neighbour distances is 18.35345. The corresponding Euclidean ratio is 7.8144152. The angular ratio compares target-to-source distance with within-source distance, so a value of one denotes equality of the two reference distances; neither ratio directly measures a manifold dimension.

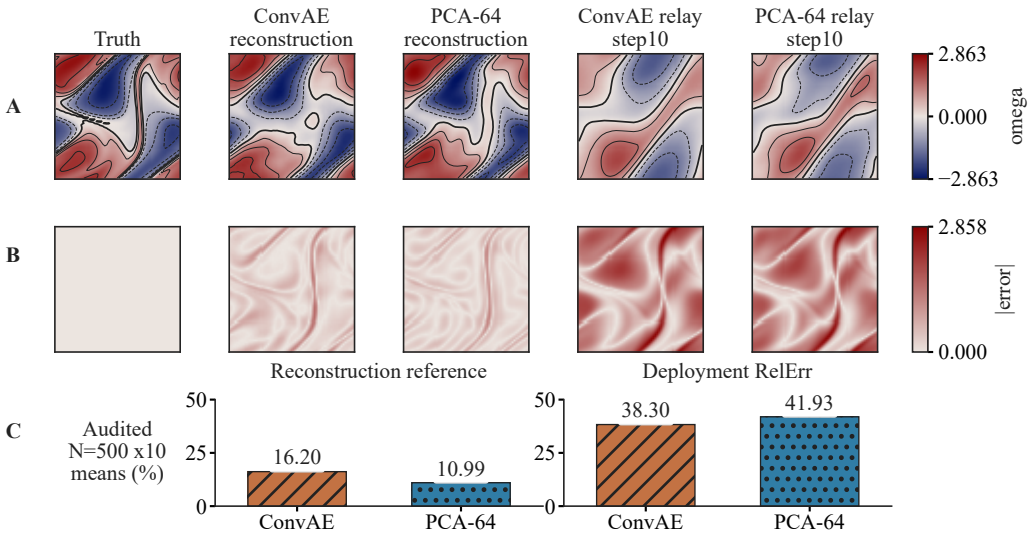
Linear variance complexity rises across regimes (Appendix F); we treat this as a description, not an explanation.

5.2 Four interventions on representation

Figure 2 gives a field-level example; Figure 3 places the reversal among the aggregate intervention results.

The interventions decouple reconstruction fidelity from deployment ordering (Table 2, Figures 2–3). Random projection alone does not recover ConvAE; its reconstruction reference is already near 100%, so this control does not isolate latent geometry.

At equal width and a common latent-output route, PCA-64 transfers worse than ConvAE by 3.6317 pp [3.2635, 4.0007] despite reconstructing better. This reversal rejects reconstruction fidelity as a sufficient ordering criterion among these encoders.



Signed contours: positive solid; negative dashed; zero bold.

Figure 2: Reconstruction and deployment reversal. Top: truth, two frame-19 reconstructions, and two step-10 relay predictions; middle: absolute errors with one shared scale. ConvAE seed 42; trajectory 105, selected before field inspection by nearest-to-median ten-step paired deployment difference (PCA-64 minus ConvAE: +2.187634 pp; smallest-ID tie-break). Bars are audited $N = 500 \times 10$-step means, not snapshot errors. Signed contours are solid positive, dashed negative and bold zero.

Table 2: Representation interventions. "Ref." denotes a reconstruction reference except for JEPA's readout reference. ConvAE and PCA rows use seed-42 ConvAE for comparison; random projection and JEPA report three training seeds. JEPA uses the physical-output comparison route.

Representation	Width	Ref. (%)	Deployed (%)	ϵ_ϕ (%)
ConvAE, seed 42	64	16.195261	38.300117	70.748903
Random projection	64	99.27	99.42	n/a
PCA	64	10.987268	41.931767	74.188400
PCA	512	3.471558	42.012421	74.062400
JEPA, config2	64	53.7193	50.8477	91.47–93.44

Aligned fields, absolute errors and audited means are shown in Figure 2.

The non-monotone PCA width sweep improves reconstruction without repairing transfer within this linear family in the tested regime (Figure 3; Appendix F).

The fourth intervention runs opposite to the inference that any departure from reconstruction training should help: JEPA config2 transfers 11.9226 pp worse [11.3002, 12.5385] on the physical-output comparison route. Its encoder passes the recorded collapse gates (participation ranks 18.29–19.64 versus 10.60); this does not isolate the predictive objective from the rest of the representation.

Encoder non-collapse and rollout queries approaching the source distribution can coexist; this does not explain why config1b lies below its readout reference. Full configuration-specific scores and angular diagnostics are retained in Appendix F.11.

A small dynamics penalty also improves the matched comparison numerically: $\lambda = 0.0006$ gives 38.388809%, compared with matched $\lambda = 0$ at 38.535474% (Table 3). The canonical three-seed baseline remains lower at 38.343243%, and only seed 42 has a paired interval excluding zero; the other two intervals include zero. The matched improvement and the absence of a canonical improvement must therefore be reported together, rather than described as a uniformly null sweep.

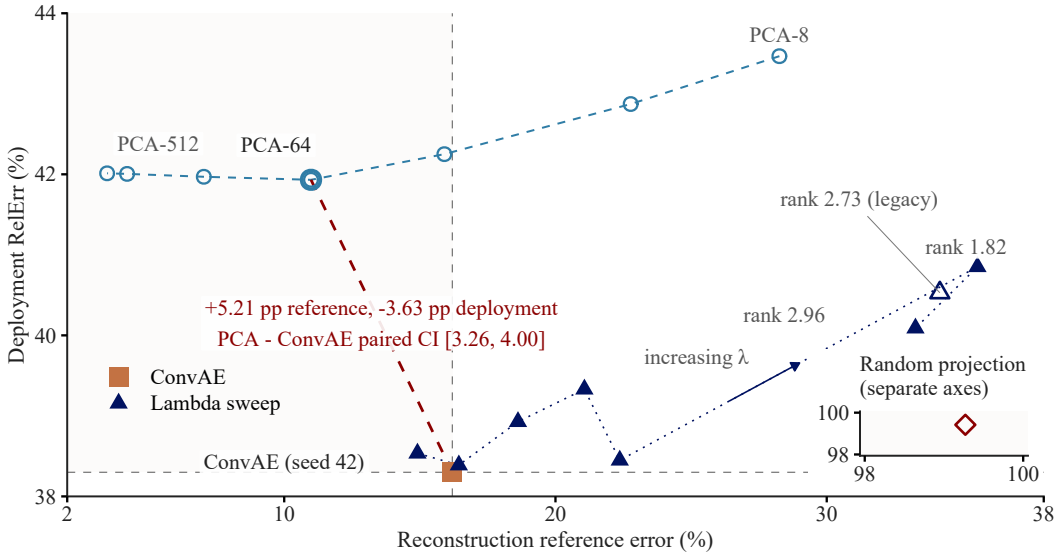


Figure 3: Reconstruction reference error does not order deployment error: PCA-64 reconstructs 5.21 pp better than ConvAE yet transfers 3.63 pp worse (paired PCA-minus-ConvAE interval [3.2635, 4.0007]). Shaded-region points reconstruct better yet transfer worse than ConvAE. PCA's width sweep and the connected λ path are non-monotone; ConvAE is seed 42 and the sweep uses three-seed means. Rank 2.73 is a separate legacy $\lambda = 0.1$ record. Random projection (99.27% reference, 99.42% deployed) needs separate inset axes; relative L_2 is not bounded at 100%. JEPA's physical-output, readout-reference point is separate in Appendix F, Figure 12.

Table 3: The restricted improvement and the canonical comparison share one table. All deployment means use three seeds. The canonical row is not the matched sweep control.

Training condition	Reference (%)	Deployed (%)
Canonical ConvAE	14.710714	38.343243
Matched $\lambda = 0$	14.911939	38.535474
Matched $\lambda = 0.0006$	16.437883	38.388809

5.3 The intervention ordering is not monotone

The reconstruction-trained ConvAE transfers better than equal-width PCA by 3.6317 pp, yet replacing reconstruction training with the tested JEPA objective worsens the same-route comparison by 11.9226 pp. At the high dynamics penalty $\lambda = 0.06$, participation rank falls to 1.82 and deployment error rises to 40.85141%. Within these interventions, neither improved reconstruction, greater linear width, nor removal of reconstruction supplies a monotone route to improved deployment. The useful implication is a requirement to test representation changes against the deployed task, not a prescription for a new training objective.

5.4 Retrieval geometry under fidelity-preserving reparameterisation

For invertible A, $E_A = AE$ and $D_A = D \circ A^{-1}$ leave $D_A(E_A(x)) = D(E(x))$ unchanged analytically; source increments become $A\Delta z$. Encoder, decoder, dimension, recoverable information and rollout rule remain fixed. Two orthogonal transforms give $5{,}000/5{,}000$ index agreement, as cosine invariance predicts, and deployment changes $+7.15 \times 10^{-7}$ and -2.34×10^{-6} pp. Forced-A0-index controls agree within 9.452343×10^{-6} pp (Appendix F.10).

Across seven conditions the reconstruction reference stays within 7.851874×10^{-6} pp of 16.195261%. Among the tested nonorthogonal conditions, mean deployment error rises as applied-index agreement falls: 62.14% gives $+0.0097$ pp, 38.04% gives $+0.0864$ pp, 20.80% gives $+0.6267$ pp, and source whitening at 7.50% gives $+4.7538$ pp. Paired intervals include zero at condition

Table 4: Eight candidates and evidence forms. Full measurements and diagnostic scope: Appendix H.

Candidate	Evidence form	Measurement and scope
Mismatch growth	Observed intervention effect	True-query angle contrast 11.155441 [10.723831, 11.589677]; no mechanism exclusion.
Temporal context	No benefit in tested variants	All three variants worse; paired intervals in Appendix H.2.
Database coverage	Positive effect, diminishing gains	Full database improves on tested subsets; all three intervals exclude zero.
Similarity metric	Improvement depends on objective	Direction/mismatch oracles improve by 2.439408/3.412029 pp; larger gains not excluded.
Curvature	No significant proxy association	$\rho = -0.0782$, $p = 0.0807$; not a rejection of intrinsic curvature.
Phase coordinates	Test incomplete	No phase-aligned control array; no substitute constructed.
Dynamics objective	Adverse high-penalty effect	Participation rank 10.09 to 1.82; small matched improvement remains an exception.
Step-1 anomaly	Association reverses after adjustment	Controlled minus raw Spearman is -0.740103.

numbers 2 and 5 but exclude zero at 20 and for whitening (Table 28). Orthogonal changes remain within numerical tolerance.

This operationalises retrieval coordinates through fidelity-preserving linear intervention, over the seven tested frames; it establishes neither a mathematical local optimum nor arbitrary-geometry behavior. Whitening's adverse effect does not establish that anisotropy ought universally to be retained or identify which variance directions carry transferable information. Whitening worsened transfer, contrary to our expectation (Appendix G).

6 Eight candidate explanations

Evidence is described by its form rather than letter grades: observed changes with their sign, no benefit from tested variants, no statistically clear proxy association, and incomplete tests. These forms do not imply eight exclusions or equal causal identification.

Stage 3 admitted only nonzero penalties with mean reconstruction reference error at most 15.411939%. The lowest nonzero penalty gave 16.437883%, so none qualified and Stage 3 was **not run**. An unrun stage is not a null-effect result.

7 What the evaluation metric can support

7.1 Two orderings need not coincide

The qualitative observation is a disagreement between the RelErr ordering and the high-wavenumber spectral-fidelity ordering in the measured comparison. Complete per-method spectral errors and ranking are not reported; RSSM and kNN-copy spectral comparisons are unavailable. We therefore do not extend that observation to the full method table or use it as a quantified ranking claim.

7.2 Spectral weighting and the prediction

The recorded prediction states that U-Net's lead would fall below 3.62 pp under the $\alpha = 1$ energy-style norm; the measured margin is 3.5314 pp, satisfying that literal inequality. Across $\alpha = 0, 1, 2$, U-Net-minus-relay differences are -3.6725, -3.5314 and -2.9610 pp (Table 5); the first-to-second reduction is the recorded 0.141 pp. This small reduction does not establish a substantial change, and the satisfied literal inequality is not a failed prediction. All three paired intervals exclude zero. Their overlap is not a paired test of the change in margins, whose interval is unavailable.

Table 5: Paired norm comparison on the spectral evaluation. These $\alpha = 0$ values are a separate evaluation record from Table 1; they are not silently substituted for its method means.

α	U-Net (%)	Relay (%)	Difference (pp)	Paired 95% CI
0	34.6686	38.3411	-3.6725	$[-3.8779, -3.4723]$
1	29.6312	33.1626	-3.5314	$[-3.8017, -3.2596]$
2	30.9576	33.9186	-2.9610	$[-3.3176, -2.5970]$

8 Learned versus retrieved transitions

Within the stated ConvAE, decoded-output comparison, latent relay gives 38.34324%. GRU gives 41.23%, and the RSSM-style predictor gives 41.00%. Their reported uncertainty conventions and parameter counts appear in Table 1. This comparison places the retrieved transition ahead in that controlled column, but does not identify retrieval as superior across encoders, output routes, or model classes. Paired trajectory distributions are shown in Appendix B, Figure 6. The physical-output ConvAE baseline is 38.92515%. JEPA config2 gives 50.8477%, with paired difference 11.9226 pp [11.3002, 12.5385]. Because the systems use different representations, this same-output-route comparison tests complete representation–prediction choices and cannot isolate a learned-versus-retrieved transition effect.

9 Limitations

The empirical scope is one benchmark, one grid, one tenfold source-to-target viscosity shift, and a boundary regime whose resolution adequacy is marginal under the specified audit. The reconstruction reference is not a decoder-output lower bound, and oracle differences are not independent causal components. Capacity descriptors are not attractor dimensions; a nonsignificant curvature-proxy association does not reject intrinsic curvature. The exploratory design and unrun Stage 3 limit generalization.

The title describes a reconstruction–transfer reversal and a probe of representation geometry. Reconstruction fidelity does not order deployment error, and retrieval-coordinate interventions have measurable effects; we do not identify which property of the frame is responsible.

Protocol limitations remain. No source-domain held-out ladder protocol was established, so we cannot isolate the cross-domain increment in these contrasts. Oracle levels currently use only seed 42; the selection-by-refresh factorial comparison is incomplete. U-Net's 34.72% main-table result differs from the 34.668555% mean of the retained stepwise record; their protocol difference remains unreconciled. Learned and retrieved transitions share the ConvAE and decoded-output route, but their transition modules are not architecture-matched. JEPA's reported paired contrast uses the seed-42 physical baseline, not a three-versus-three control; the separate three-seed physical mean is 39.033728%, without a newly certified paired interval here.

We do not present a method: no variant beats the baseline. "Variant" means tested diagnostic-predictor modifications on their canonical route: the best penalty mean is 38.388809% versus 38.343243%, and JEPA config2 is 50.8477% versus physical-route 38.92515182%. This excludes external predictors: U-Net's 34.72% is lower than relay's. The multiscale explanation is untested.

10 Conclusion

On this benchmark, nested oracle comparisons expose a substantial state-refresh effect, while attained selector improvements vary with their objective and cadence. The equal-width reconstruction reversal, PCA width sweep and opposite-direction JEPA result decouple fidelity from deployment ordering; fidelity-preserving linear interventions additionally probe retrieval-coordinate geometry without identifying a unique geometric mechanism or proposing a new method.

AI USE STATEMENT

In this work, we used generative AI tools for tasks in the required disclosure list: developing the conceptual framework of the diagnostic decomposition; formulating the mathematical claims used in the fidelity-preserving reparameterisation, including the identity $D_A(E_A(x)) = D(E(x))$ and the invariance of normalised cosine similarity under orthogonal transformations; proposing and refining hypotheses; designing the experiments and providing feedback on methodology; implementing the methods, evaluation pipelines, and analysis scripts; cleaning and reformatting datasets for the experiments; supporting the qualitative classification of candidate explanations; and interpreting results.

We have not used generative AI tools to generate synthetic datasets; all fields come from a previously released benchmark. We have not used generative AI tools for translation. The writing of formal proofs is not applicable to this work, as it contains none.

Additionally, we used generative AI tools for tasks in the recommended disclosure list: creating and modifying the figures; creating and editing software code; drafting parts of this paper; summarising and analysing existing literature and identifying relevant work; identifying gaps and brainstorming; searching for information; editing the manuscript for readability; formatting references; and proposing the title and keywords.

We have reviewed all AI-assisted work. We checked reported numerical results against retained arrays and source records. Numerical reproduction checks with known analytic targets were used where available: the orthogonal reparameterisations were required to reproduce identical retrieval indices, and the affine decoder check was required to reproduce its analytic values.

The authors take responsibility for the final content of this work, including all text, claims, figures, and artifacts produced with the aid of generative AI. All scope limitations, adjudications of what the evidence supports, and the decision not to claim a method are the authors' own.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Assran_Self-Supervised_Learning_From_Images_With_a_Joint-Embedding_Predictive_Architecture_CVPR_2023_paper.html.

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. arXiv preprint arXiv:2506.09985, 2025. URL https://arxiv.org/abs/2506.09985.

Randall Balestriero and Yann LeCun. How learning by reconstruction produces uninformative features for perception. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 2566–2585, 2024. URL https://proceedings.mlr.press/v235/balestriero24b.html.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. arXiv preprint arXiv:2511.08544, 2025. URL https://arxiv.org/abs/2511.08544.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=QaCCuDfBk2.

Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is "nearest neighbor" meaningful? In *Database Theory – ICDT'99*, volume 1540 of *Lecture Notes in Computer Science*, pp. 217–235. Springer, 1999. doi: 10.1007/3-540-49257-7_15. URL https://link.springer.com/chapter/10.1007/3-540-49257-7_15.

Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: Applications to image and text data. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '01)*, pp. 245–250. ACM, 2001. doi: 10.1145/502512.502546. URL https://doi.org/10.1145/502512.502546.

Guido Boffetta and Robert E. Ecke. Two-dimensional turbulence. *Annual Review of Fluid Mechanics*, 44:427–451, 2012. doi: 10.1146/annurev-fluid-120710-101240. URL https://doi.org/10.1146/annurev-fluid-120710-101240.

Johannes Brandstetter, Max Welling, and Daniel E. Worrall. Lie point symmetry data augmentation for neural PDE solvers. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 2241–2256, 2022a. URL https://proceedings.mlr.press/v162/brandstetter22a.html.

Johannes Brandstetter, Daniel E. Worrall, and Max Welling. Message passing neural PDE solvers. In *International Conference on Learning Representations (ICLR)*, 2022b. URL https://arxiv.org/abs/2202.03376.

Ashesh Chattopadhyay, Ebrahim Nabizadeh, and Pedram Hassanzadeh. Analog forecasting of extreme-causing weather patterns using deep learning. *Journal of Advances in Modeling Earth Systems*, 12(2):e2019MS001958, 2020. doi: 10.1029/2019MS001958. URL https://doi.org/10.1029/2019MS001958.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder

for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014. doi: 10.3115/v1/D14-1179. URL https://aclanthology.org/D14-1179/.

Mengdi Chu, Yang Liu, Ayan Biswas, and Han-Wei Shen. Do physics foundation models learn generalizable physics? A bias-aware benchmark across physical regimes and distribution shifts. arXiv preprint arXiv:2605.29283, 2026. URL https://arxiv.org/abs/2605.29283.

Robert J. Durrant and Ata Kabán. When is 'nearest neighbour' meaningful: A converse theorem and implications. *Journal of Complexity*, 25(4):385–397, 2009. doi: 10.1016/j.jco.2009.02.011. URL https://doi.org/10.1016/j.jco.2009.02.011.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Elena Facco, Maria d'Errico, Alex Rodriguez, and Alessandro Laio. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Scientific Reports*, 7(1): 12140, 2017. doi: 10.1038/s41598-017-11873-y. URL https://doi.org/10.1038/s41598-017-11873-y.

J. Doyne Farmer and John J. Sidorowich. Predicting chaotic time series. *Physical Review Letters*, 59 (8):845–848, 1987. doi: 10.1103/PhysRevLett.59.845. URL https://doi.org/10.1103/PhysRevLett.59.845.

Ewen Frogé, Carlos Granero-Belinchón, Stéphane G. Roux, Nicolas B. Garnier, and Thierry Chonavel. Analog-based forecasting of turbulent velocity: Relationship between unpredictability and intermittency. *Europhysics Letters*, 149(1):13001, 2025. doi: 10.1209/0295-5075/ada4d3. URL https://doi.org/10.1209/0295-5075/ada4d3.

Jonathan Gallagher and Roberto Guglielmi. Toward goal-agnostic joint-embedding predictive control of partial differential equations. arXiv preprint arXiv:2607.21644, 2026. URL https://arxiv.org/abs/2607.21644.

Peiran Gao, Eric Trautmann, Byron Yu, Gopal Santhanam, Stephen Ryu, Krishna Shenoy, and Surya Ganguli. A theory of multineuronal dimensionality, dynamics and measurement. bioRxiv 214262, 2017. URL https://www.biorxiv.org/content/10.1101/214262v2.

Zhiwei Gao, Liu Yang, and George Em Karniadakis. Spectral audit of in-context operator networks. arXiv preprint arXiv:2606.02427, 2026. URL https://arxiv.org/abs/2606.02427.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://arxiv.org/abs/2111.00396.

Jayesh K. Gupta and Johannes Brandstetter. Towards multi-spatiotemporal-scale generalized PDE modeling. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=dPSTDbGtBY.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2555–2565, 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025. doi: 10.1038/s41586-025-08744-2. URL https://doi.org/10.1038/s41586-025-08744-2.

Sungwon Han, Seungeon Lee, Meeyoung Cha, Sercan Ö. Arik, and Jinsung Yoon. Retrieval augmented time series forecasting. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21774–21797, 2025. URL https://proceedings.mlr.press/v267/han25d.html.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/cf73d57b6dcda32b293df7c2d5341f49-Abstract-Conference.html.

Peiyan Hu, Haodong Feng, Hongyuan Liu, Tongtong Yan, Wenhao Deng, Tianrun Gao, Rong Zheng, Haoren Zheng, Chenglei Yu, Chuanrui Wang, Kaiwen Li, Zhi-Ming Ma, Dezhi Zhou, Xingcai Lu, Dixia Fan, and Tailin Wu. RealPDEBench: A benchmark for complex physical systems with real-world data. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2601.01829.

William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference on Modern Analysis and Probability*, volume 26 of *Contemporary Mathematics*, pp. 189–206, Providence, RI, 1984. American Mathematical Society. doi: 10.1090/conm/026/737400. URL https://doi.org/10.1090/conm/026/737400.

Ian T. Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016. doi: 10.1098/rsta.2015.0202. URL https://doi.org/10.1098/rsta.2015.0202.

Siavash Khodakarami, Vivek Oommen, Nazanin Ahmadi Daryakenari, Maxim Beekenkamp, and George Em Karniadakis. Spectral bias in physics-informed and operator learning: Analysis and mitigation guidelines. arXiv preprint arXiv:2602.19265, 2026. URL https://arxiv.org/abs/2602.19265.

Matthieu Kirchmeyer, Yuan Yin, Jérémie Donà, Nicolas Baskiotis, Alain Rakotomamonjy, and Patrick Gallinari. Generalizing to new physical systems via context-informed dynamics model. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 11283–11301, 2022. URL https://proceedings.mlr.press/v162/kirchmeyer22a.html.

Dmitrii Kochkov, Jamie A. Smith, Ayya Alieva, Qing Wang, Michael P. Brenner, and Stephan Hoyer. Machine learning–accelerated computational fluid dynamics. *Proceedings of the National Academy of Sciences*, 118(21):e2101784118, 2021. doi: 10.1073/pnas.2101784118. URL https://doi.org/10.1073/pnas.2101784118.

Felix Koehler, Simon Niedermayr, Rüdiger Westermann, and Nils Thuerey. APEBench: A benchmark for autoregressive neural emulators of PDEs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 120252–120310, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d9875ebcf74bccdc5076acab0dbee62c-Abstract-Datasets_and_Benchmarks_Track.html.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023. URL https://jmlr.org/papers/v24/21-1524.html.

Robert H. Kraichnan. Inertial ranges in two-dimensional turbulence. *The Physics of Fluids*, 10 (7):1417–1423, 1967. doi: 10.1063/1.1762301. URL https://doi.org/10.1063/1.1762301.

Brian Kulis. Metric learning: A survey. *Foundations and Trends in Machine Learning*, 5(4):287–364, 2012. doi: 10.1561/2200000019. URL https://doi.org/10.1561/2200000019.

Elizaveta Levina and Peter J. Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in Neural Information Processing Systems*, volume 17, 2004. URL https://papers.nips.cc/paper/2577-maximum-likelihood-estimation-of-intrinsic-dimension.

Redouane Lguensat, Pierre Tandeo, Pierre Ailliot, Manuel Pulido, and Ronan Fablet. The analog data assimilation. *Monthly Weather Review*, 145(10):4093–4107, 2017. doi: 10.1175/MWR-D-16-0441.1. URL https://journals.ametsoc.org/view/journals/mwre/145/10/mwr-d-16-0441.1.xml.

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 (ICLR)*, 2021. URL https://arxiv.org/abs/2010.08895.

Zongyi Li, Miguel Liu-Schiaffini, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Learning chaotic dynamics in dissipative systems. In *Advances in Neural Information Processing Systems*, volume 35, pp. 16768–16781, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6ad68277e27b42c60ac228c9859fc1a2-Abstract-Conference.html.

Zongyi Li, Samuel Lanthaler, Catherine Deng, Michael Chen, Yixuan Wang, Kamyar Azizzadenesheli, and Anima Anandkumar. Scale-consistent learning for partial differential equations. arXiv preprint arXiv:2507.18813, 2025. URL https://arxiv.org/abs/2507.18813.

Phillip Lippe, Bastiaan S. Veeling, Paris Perdikaris, Richard E. Turner, and Johannes Brandstetter. PDE-Refiner: Achieving accurate long rollouts with neural PDE solvers. In *Advances in Neural Information Processing Systems*, volume 36, pp. 67398–67433, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d529b943af3dba734f8a7d49efcb6d09-Abstract-Conference.html.

Andy Zeyi Liu, Haoran Sun, Lucas Baker, Randall Balestriero, and John Sous. Semigroup-JEPA: Latent dynamics consistency for zero-shot physics generalization. arXiv preprint arXiv:2609.10464, 2026. URL https://arxiv.org/abs/2609.10464.

Edward N. Lorenz. Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric Sciences*, 26(4):636–646, 1969. doi: 10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2. URL https://journals.ametsoc.org/view/journals/atsc/26/4/1520-0469_1969_26_636_aparbn_2_0_co_2.xml.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021. doi: 10.1038/s42256-021-00302-5. URL https://doi.org/10.1038/s42256-021-00302-5.

Nick McGreivy and Ammar Hakim. Weak baselines and reporting biases lead to overoptimism in machine learning for fluid-related partial differential equations. *Nature Machine Intelligence*, 6(10):1256–1269, 2024. doi: 10.1038/s42256-024-00897-5. URL https://doi.org/10.1038/s42256-024-00897-5.

Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina J. Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Stuart B. Dalziel, Drummond B. Fielding, Daniel Fortunato, Jared A. Goldberg, Keiya Hirashima, Yan-Fei Jiang, Rich R. Kerswell, Suryanarayana Maddu, Jonah Miller, Payel Mukhopadhyay, Stefan S. Nixon, Jeff Shen, Romain Watteaux, Bruno Régaldo-Saint Blancard, François Rozet, Liam H. Parker, Miles Cranmer, and Shirley Ho. The well: a large-scale collection of diverse physics simulations for machine learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 44989–45037, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4f9a5acd91ac76569f2fe291b1f4772b-Abstract-Datasets_and_Benchmarks_Track.html.

Masataka Okabe and Kei Ito. Color universal design (CUD): How to make figures and presentations that are friendly to colorblind people. J*FLY web page, 2008. URL https://jfly.uni-koeln.de/color/.

Vivek Oommen, Aniruddha Bora, Zhen Zhang, and George Em Karniadakis. Integrating neural operators with diffusion models improves spectral representation in turbulence modelling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 481(2309): 20240819, 2025. doi: 10.1098/rspa.2024.0819. URL https://doi.org/10.1098/rspa.2024.0819.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(86):2487–2531, 2010. URL https://jmlr.org/papers/v11/radovanovic10a.html.

Stephan Rasp, Stephan Hoyer, Alexander Merose, Ian Langmore, Peter Battaglia, Tyler Russell, Alvaro Sanchez-Gonzalez, Vivian Yang, Rob Carver, Shreya Agrawal, Matthew Chantry, Zied Ben Bouallegue, Peter Dueben, Carla Bromberg, Jared Sisk, Luke Barrington, Aaron Bell, and Fei Sha. WeatherBench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, 16(6):e2023MS004019, 2024. doi: 10.1029/2023MS004019. URL https://doi.org/10.1029/2023MS004019.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351 of *Lecture Notes in Computer Science*, pp. 234–241. Springer, 2015. doi: 10.1007/978-3-319-24574-4_28. URL https://doi.org/10.1007/978-3-319-24574-4_28.

Jianing Shi. Striding across Reynolds numbers: Representation geometry in neural PDE generalisation, 2026. URL https://arxiv.org/abs/2605.30112.

Shashank Subramanian, Peter Harrington, Kurt Keutzer, Wahid Bhimji, Dmitriy Morozov, Michael W. Mahoney, and Amir Gholami. Towards foundation models for scientific machine learning: Characterizing scaling and transfer behavior. In *Advances in Neural Information Processing Systems*, volume 36, pp. 71242–71262, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e15790966a4a9d85d688635c88ee6d8a-Abstract-Conference.html.

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/file/0a9747136d411fb83f0cf81820d44afb-Paper-Datasets_and_Benchmarks.pdf.

A DATASET AND EVALUATION PROTOCOL

The evaluation uses a shared target-regime trajectory set and a fixed source-regime retrieval database. Table 6 distinguishes measured protocol details from unresolved benchmark-generation specifications.

Table 6: Recorded evaluation and retrieval protocol.

Item	Specification
Source regime	$\nu = 10^{-3}$. Source trajectory indices 0–3999; database frames 10–49; 156,000 recorded entries.
Target regime	$\nu = 10^{-4}$. Evaluation trajectory indices 0–499; context frame 9; scored frames 10–19; horizon $T = 10$.
Boundary regime	$\nu = 10^{-5}$. Resolution diagnostic classified as questionable, as detailed in Appendix D.
Retrieval	Unit-normalized 64-dimensional vectors, dot-product similarity, top-1 selection; recorded setting $N_{\mathrm{RIDE}} = 3$.
Training seeds	42, 123, and 7.
Paired uncertainty	10,000 bootstrap replicates, random seed 12345, paired over 500 shared trajectories, 95% intervals. For a three-seed model, average its errors across seeds within each trajectory before pairing against the baseline.
Primary error	$\mathrm{RelErr} = \dfrac{100}{NT} \displaystyle\sum_{n,t} \dfrac{\|\widehat{\omega}_{nt} - \omega_{nt}\|_2}{\|\omega_{nt}\|_2 + 10^{-10}}$.
Grid and solver	All three field arrays have spatial shape 64×64. The complete generation specification for spatial domain, solver, integration step, dealiasing, forcing amplitude and initial-condition distribution is not specified here.
Retrieval construction	Source entries pair start frames 10–48 with their successors within each trajectory; the last frame provides no additional transition. Full generation and boundary-handling specifications remain incomplete.

RelErr measures a relative norm discrepancy. A value of 100% means that the error norm equals the truth norm, before the small denominator stabilizer is considered; it does not identify a particular phase or amplitude failure. The reconstruction reference error evaluates the selected latent point $E(x)$ through $D(E(x))$ and does not optimize over latent points. The readout reference error uses its specified readout path and must be distinguished from the reconstruction reference error.

The retrieval instrument matches unit-normalized 64-dimensional queries by dot product, with top-1 retrieval and $N_{\mathrm{RIDE}} = 3$. The latent route stores latent increments and calls the decoder once per prediction step; the physical route stores field increments and uses no decoder during inference. Changing the route changes the appropriate baseline. Reconstruction and readout reference errors measure the output of their specified encoding or readout paths, while ϵ_ϕ denotes the relative norm discrepancy between borrowed and true increments. A value of 100% for that discrepancy is not a statement about increment correlation. Appendix A specifies the estimator and saved-index mapping.

A.1 EXPLORATORY DESIGN AND PHYSICAL-TIME CONVENTION

The diagnostics are exploratory follow-up measurements, with known benchmark outcomes in view, rather than confirmatory experiments. Prior expectations and measurements appear in Appendix G; diagnostic rules and analytic checks appear in Appendix H.

The main comparisons score frames 10–19 after conditioning on frame 9. Source, target and boundary arrays have shapes $(5000, 64, 64, 50)$, $(10000, 64, 64, 50)$ and $(1200, 64, 64, 20)$, ordered as trajectory, spatial axes, frame. Source and target snapshot spacing is $\Delta t = 1$, giving a horizon of 10 physical units. Equal frame counts do not establish equal turnover horizons.

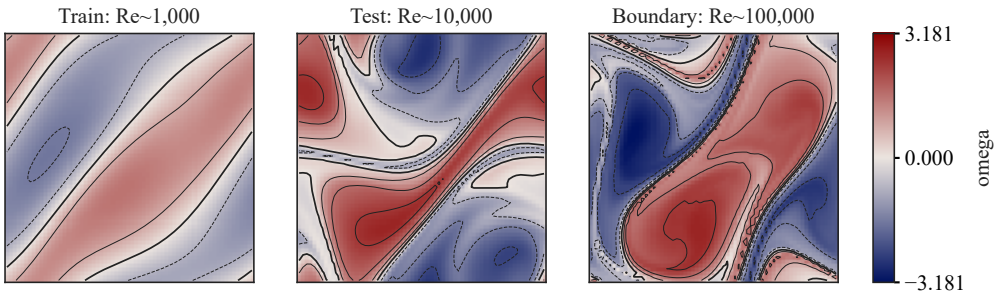


Figure 4: Real vorticity fields in three regimes, with one symmetric color scale. Trajectory 0 and frame 19 were fixed before field inspection in all panels; these are illustrations, not performance-selected exemplars. Forcing is the diagonal single mode $(1, 1)$. Boundary resolution is questionable: $Z_{\mathrm{Nyq}}/Z_{\mathrm{peak}} = 3.945078 \times 10^{-6}$ and tail slope -3.2163, versus target -12.9202; its panel is descriptive, not evidence of a resolved working regime. Signed contours are solid positive, dashed negative and bold zero.

Figure 4 gives the fixed illustrative fields across the three regimes.

The measured energy-turnover proxy is $\tau_E = L_E/U_{\mathrm{rms}}$, with $L_E = \sum_k E_k/k_{\mathrm{phys}} \,/\, \sum_k E_k$ excluding the zero mode and $U_{\mathrm{rms}} = \sqrt{\langle u_x^2 + u_y^2 \rangle}$. Velocity is reconstructed from vorticity through $\widehat{\psi} = \widehat{\omega}/k_{\mathrm{phys}}^2$ on the unit torus. Mean source and target turnover times are 1.2024779220607038 and 0.8164688141725885; dividing the horizon by these means gives 8.316161000996056 and 12.247865229407475 turnovers. The measured horizon ratio is 1.473, while the alternative vorticity proxy gives 1.547; these are different definitions. Sampling uses trajectories 0–499, frames 10–19, and 10,000 trajectory resamples with seed 20260831, a separate uncertainty protocol. Boundary turnover was not measured.

Increment-discrepancy implementation. The recorded single-seed discrepancy is 70.748903%, and the L2 reconstruction audit records a discrepancy of -2.69×10^{-8} pp from its saved reference. The estimator uses

$$\epsilon_\phi = \frac{1}{N} \sum_n \frac{1}{T} \sum_t 100 \frac{\|\Delta\omega_{\mathrm{src}}(r_{nt}) - \Delta\omega_{\mathrm{tgt}}(n,t)\|_2}{\|\Delta\omega_{\mathrm{tgt}}(n,t)\|_2}. \tag{2}$$

Fields are flattened and cast to float64 before subtraction, with no denominator stabilizer. Saved row r maps to source trajectory $\lfloor r/39 \rfloor$ and start frame $10 + (r \bmod 39)$; target transitions run from $9 \to 10$ to $18 \to 19$. Matching uses true-state unit-normalized latent-cosine top-1 retrieval, not the direction-optimizing oracle.

A.2 Protocol and implementation distinctions

The physical-output algorithm has error 38.92515182%; the latent-output implementation has error 38.34%. Four protocol and implementation distinctions are itemized in Appendix A. The latent-increment and physical-increment implementations produce different results. The latent route stores latent increments and invokes the decoder at each prediction step; the physical route stores field increments and does not invoke the decoder during inference. Table 7 specifies these distinct measurement conventions.

Table 7: Protocol and implementation distinctions relative to Shi (2026); distance and seed entries specify the present measurement conventions.

Quantity	Specification
Forcing	The forcing is the diagonal single mode $(1, 1)$; Shi (2026), Section 3.1, describes Kolmogorov forcing at $k_f = 4$.
Output route	The physical-increment inference specified in Shi (2026), Section 3.3 and Algorithm 1, has error 38.92515182% in the present measurement; the latent-increment implementation produced the 38.34% result.
Distance	The ratio 18.28 denotes an angular or cosine-distance comparison. The Euclidean nearest-neighbour ratio is 7.8144152.
Seed average	The value $\epsilon_\phi = 70.75\%$ is single-seed. The three-seed mean is 70.428636%.

A.3 PCA FITTING PROVENANCE

PCA retains variance directions defined by the fitting sample (Jolliffe & Cadima, 2016); changing the fitting regime therefore changes the intervention. Source-fitted PCA-36 has $\epsilon_\phi = 74.621200\%$, with the retained full-precision record 74.62120056152344%. Its PCA fit uses source-regime trajectories 0–3999, frames 10–49, and $160{,}000$ fields. The competing record, 74.918317%, uses a target-regime fit on trajectories 0–199, frames 10–29, and $4{,}000$ fields. The two protocols share only 66.1% of matching indices. We retain the source-fitted value in the intervention comparison because it matches the fitting domain of the other PCA rows; these records are different protocols, rather than alternative rounding of one measurement.

A.4 RESOLUTION AUDIT

The source and target regimes fall in the sufficient category of the recorded spectral diagnostic. Their ratios $Z_{\mathrm{Nyq}}/Z_{\mathrm{peak}}$ are 8.731661×10^{-18} and 2.201566×10^{-11}. The boundary ratio is 3.945078×10^{-6} and falls in the questionable category. This audit is a specified resolution diagnostic, rather than a grid-convergence experiment.

Table 8: Resolution audit. The declared categories are sufficient below 10^{-6}, questionable between 10^{-6} and 10^{-4}, and under-resolved above 10^{-4}.

Regime	$Z_{\mathrm{Nyq}}/Z_{\mathrm{peak}}$	Tail slope	Category
$\nu = 10^{-3}$	8.731661×10^{-18}	-3.6700	Sufficient
$\nu = 10^{-4}$	2.201566×10^{-11}	-12.9202	Sufficient
$\nu = 10^{-5}$	3.945078×10^{-6}	-3.2163	Questionable

A.5 TWO SPECTRAL-RESPONSE MEASUREMENTS

Figure 5 gives the spectral-response slope and half-gain summaries.

The single-mode probe and real-field spectral-ratio measurements give slopes of opposite sign for each reported representation. For ConvAE, the slopes are -3.0622309 and $+9.7915934$ over $k = 2$–43. For PCA-64, they are -9.3653414 and $+3.2427866$ on the same fitting interval. Table 9 records each measurement separately; these measurements do not authorize a single filtering interpretation of the encoder.

A2: registered spectral-response summaries

Fit interval k=2-43; opposite signs, not a low-pass claim

Representation	Estimator	Log slope	Half-gain k
ConvAE seed 42	Single-mode probe	-3.0622309	7.0582684
ConvAE seed 42	Real-field ratio	+9.7915934	NaN (not detected)
PCA-64	Single-mode probe	-9.3653414	0.0000000
PCA-64	Real-field ratio	+3.2427866	5.8449146

Figure 5: Decoder spectral-response probe log slopes and half-gain locations (A2), presented as a summary table. Single-mode probe and real-field ratios are different estimators; full response curves are not shown.

Table 9: Decoder spectral-response probe spectral response, fitted over $k = 2$–43. The PCA-64 analytic-versus-measured maximum absolute gain discrepancy is 1.10×10^{-5}. "Not detected" preserves the missing half-gain location.

Representation	Measurement	Log slope	Half-gain k
ConvAE, seed 42	Single-mode probe	-3.0622309	7.0582684
ConvAE, seed 42	Real-field spectral ratio	$+9.7915934$	Not detected
PCA-64	Single-mode probe	-9.3653414	0.0000000
PCA-64	Real-field spectral ratio	$+3.2427866$	5.8449146

B ARCHITECTURES, ROUTES, AND TRAINING SPECIFICATION

The available records identify model families and parameter counts, but do not provide a complete architecture and optimizer specification. Table 10 preserves the available model information. Layer widths, kernel sizes, normalization, optimizer, learning-rate schedule, batch size, epoch count, checkpoint selection and exact seed-to-checkpoint mapping remain incompletely documented; these are required for reproduction from scratch.

Table 10: Recorded model sizes and training-seed coverage. The kNN-copy aggregate uses the shared $N = 500$ protocol; its supplementary stepwise record is omitted. Missing entries are not zero.

Model	Parameters	Reported seeds
Persistence	0	*not reported*
FNO	465,717	3
PCA relay, $d = 36$	0	*not reported*
kNN-copy	0	*not reported*
ConvAE-GRU	1,706,081	3
RSSM-style	5,274,337	3
ConvAE-Relay, latent	1,606,561	3
ConvAE-Relay, physical	1,606,561	Seed 42 in main comparison
U-Net	7,768,001	3

The matched λ sweep uses a PairDataset training-data setting. Its matched baseline is 38.535474%, compared with the canonical three-seed 38.343243%. The recorded difference is 0.192231 per-

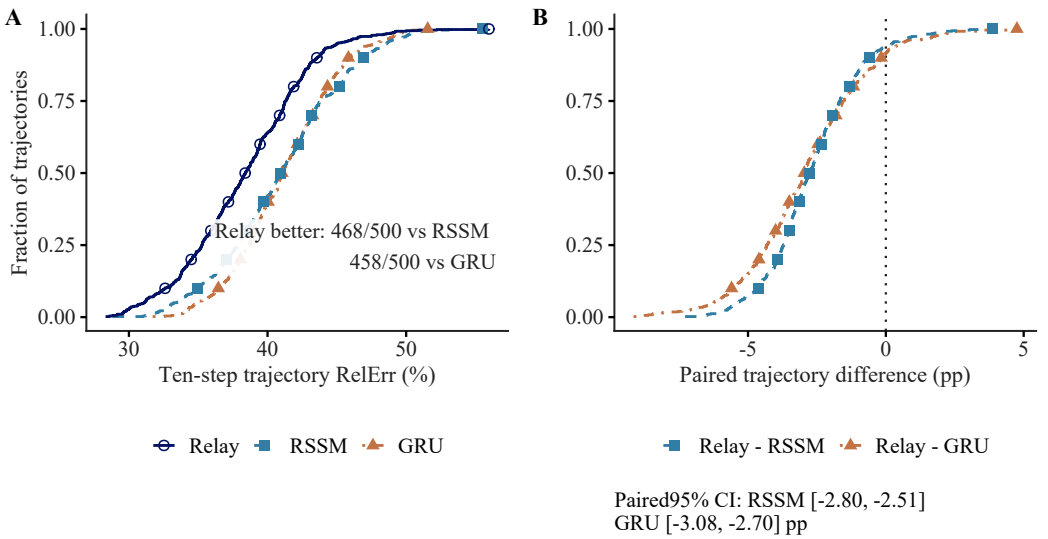


Figure 6: Shared-encoder, decoded-output distributions (Section 8). Each trajectory value averages ten scored steps and three seeds. A: all 500 trajectory values; relay is lower than RSSM for $468/500$ (93.6%) and GRU for $458/500$ (91.6%). B: paired trajectory differences. A new paired bootstrap uses 10,000 shared whole-trajectory draws, seed 12345: relay minus RSSM is -2.654862 pp $[-2.800876, -2.505606]$, and relay minus GRU is -2.887287 pp $[-3.077403, -2.697651]$. The displayed intervals use this paired trajectory-resampling protocol.

centage points. The complete dataset-construction difference and its rationale remain incompletely documented; the matched and canonical controls remain separate throughout the reporting.

The JEPA configurations are recorded as config1a, config1b, config2 and config2b. Their complete architectural definitions, training objectives, inference differences and readout-fitting protocols remain incompletely documented. The route-matched physical baseline and measured readout errors are reported in Appendix F.

B.1 Paired shared-encoder trajectory distributions

C Stepwise diagnostics and metric weighting

Figure 8 gives the measured latent-drift endpoints.

The second-log-difference diagnostic is unavailable. No curve, differentiated values or null result are inferred from rounded stepwise errors.

Spectral discriminating window. For U-Net, relay, GRU, and FNO, the spectral analysis uses between-over-pooled weights: $f(t) = b(t)/v(t)$, $w(t) = f(t)/\sum_s f(s)$, where $v(t)$ pools method–trajectory errors. Under this four-method definition, N_{eff} is 6.985763 at $\alpha = 0$. It rises to 8.237395 at $\alpha = 1$ and 8.800532 at $\alpha = 2$. Spectral weighting distributes discrimination more evenly over the ten scored steps in this comparison. The weighting definition differs from Section C; the difference between those diagnostics is principally definitional, not merely a change in method coverage.

Rollout detail. The six-method diagnostic uses 500 shared target trajectories. With between-method variance $b(t)$ and weights $w(t) = b(t)/\sum_s b(s)$, the measured $N_{\text{eff}} = 1/\sum_t w(t)^2$ is 4.901499 out of ten scored steps. This is a concentration measure of between-method variation, not a reduction in the number of evaluated frames.

U-Net starts at 5.5386% error, compared with 11.20% for relay, 14.31% for GRU, 17.81% for RSSM, and 32.71% for FNO. By step ten, U-Net reaches 66.7901%, close to GRU's 67.07% and

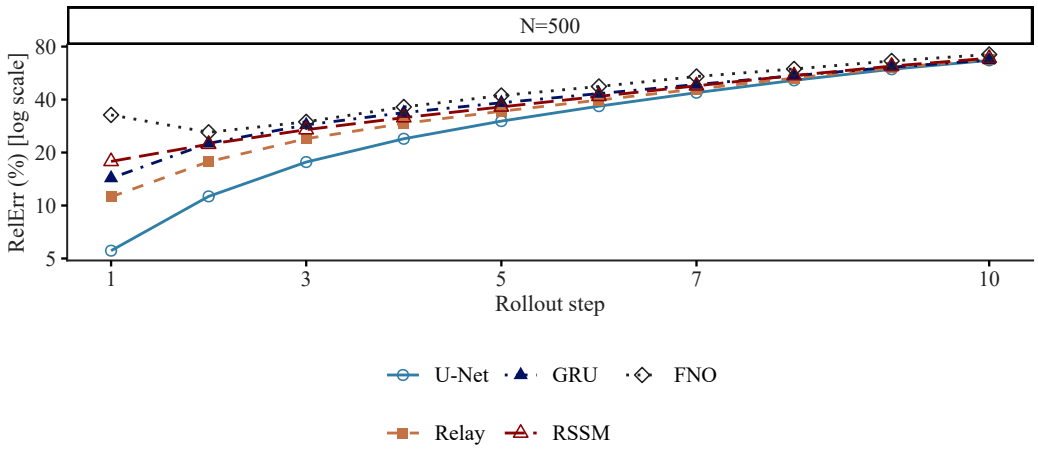


Figure 7: Stepwise RelErr for five $N = 500$ methods (F3). PCA-64 stepwise values and bootstrap bands are not reported here. The persistence reference is an aggregate, not a stepwise curve.

A1: latent target-cloud distance endpoints
Full curves and PCA endpoints are not displayed

Method	Step 1	Step 10
Relay	0.168426	0.452983
GRU	0.223587	0.384741
FNO	0.409036	0.298289
U-Net	0.073453	0.404303

Figure 8: Latent-drift endpoints (A1), presented as a summary rather than a full target-cloud-distance curve. Full profiles are not reported here, and their normalization remains incompletely specified in the comparison summary.

relay's 67.80%. Its advantage is concentrated in the early steps; these trajectories do not isolate an initial-projection mechanism. The full stepwise records are in Appendix C.

Figure 9 summarises the six-method between-variance effective-step count.

Figure 13 gives the increment-discrepancy distributions.

U-Net's advantage is concentrated in the early prediction steps. Its first-step error is 5.538634%, compared with 11.20% for relay, 14.31% for GRU, 17.81% for RSSM, and 32.71% for FNO. At the final scored step, U-Net and GRU reach 66.790117% and 67.07%. These comparisons locate an observed advantage in time without isolating its cause.

Table 11: Stepwise RelErr on the reported $N = 500$ sets. U-Net errors are first computed separately from each seed's predicted fields and then averaged across seeds. The individual seed-by-step values were not saved.

Step	U-Net	Relay	GRU	RSSM	FNO
1	5.538634	11.20	14.31	17.81	32.71
2	11.268797	17.75	22.55	22.29	26.06

Step	U-Net	Relay	GRU	RSSM	FNO
3	17.654276	23.98	28.70	26.97	29.88
4	23.941517	29.26	33.70	31.56	36.30
5	30.169434	34.34	38.35	36.33	42.16
6	36.684382	39.73	43.24	41.70	47.46
7	43.759126	45.89	48.68	47.88	54.09
8	51.434108	52.95	54.68	54.76	59.69
9	59.445159	60.53	61.03	61.97	66.26
10	66.790117	67.80	67.07	68.70	72.13

A supplementary stepwise kNN-copy record exists but its trajectory alignment with the shared evaluation set is unresolved; it is not reported here.

C.1 Two definitions of discriminative time weights

For the six-method diagnostic, let $b(t)$ be the across-method variance of the mean error at step t. The normalized between-method variance weight is

$$w_{\text{between}}(t) = \frac{b(t)}{\sum_s b(s)}, \qquad N_{\text{eff}} = \frac{1}{\sum_t w(t)^2}. \tag{3}$$

The six-method set contains ConvAE relay, ConvAE-GRU, RSSM, FNO, U-Net, and PCA-64 relay, with the reported shared $N = 500$ arrays. This definition gives $N_{\text{eff}} = 4.901499$. On the four-method subset, it gives 4.613105. These values describe the concentration of discriminative weights across scored steps, rather than a count of independent dynamical observations.

For the four-method spectral diagnostic, let $v(t)$ denote the pooled variance over method-by-trajectory errors at step t. This analysis instead uses

$$f(t) = \frac{b(t)}{v(t)}, \qquad w_{\text{ratio}}(t) = \frac{f(t)}{\sum_s f(s)}. \tag{4}$$

On the four-method set comprising U-Net, ConvAE relay, ConvAE-GRU, and FNO, this between-over-pooled definition gives $N_{\text{eff}} = 6.985763$ at $\alpha = 0$. Applying this definition to the six-method $\alpha = 0$ arrays gives 6.525105. The definition of the weights, rather than method coverage alone, accounts for the principal difference between the two reported analyses.

Table 12: Spectral weighting and effective discriminative steps under the four-method analysis's between-over-pooled definition, using the four-method set. These values should not be directly compared with a six-method between-variance value without changing the weight definition.

Spectral weight α	N_{eff}
0	6.985763
1	8.237395
2	8.800532

The six-method analysis cannot be reconstructed at $\alpha = 1$ or $\alpha = 2$ from the retained artifacts. RSSM retains only the $\alpha = 0$ error arrays, without the $N = 500$ prediction fields. The available PCA relay fields contain missing tail entries, as documented in Appendix E. The separate source-fitted PCA capacity results remain usable because their source-only pipeline retained finite arrays.

C.2 Rollout-state mismatch and latent drift

The reported rollout-state ϵ_ϕ reaches a local minimum of 64.071931% at step 3. It rises to 100.811557% at step 10. At the latter step, it exceeds the true-state retrieval measurement by

F4: effective-step summary

Between-variance definition; full w(t) curve not displayed

Weight definition / coverage	N_eff
Between-variance; six methods; N=500	4.901499

Figure 9: Six-method between-variance effective-step summary (F4), $N_{\text{eff}} = 4.901499$. The full $w(t)$ curve is not displayed. This is not the four-method between-over-pooled diagnostic.

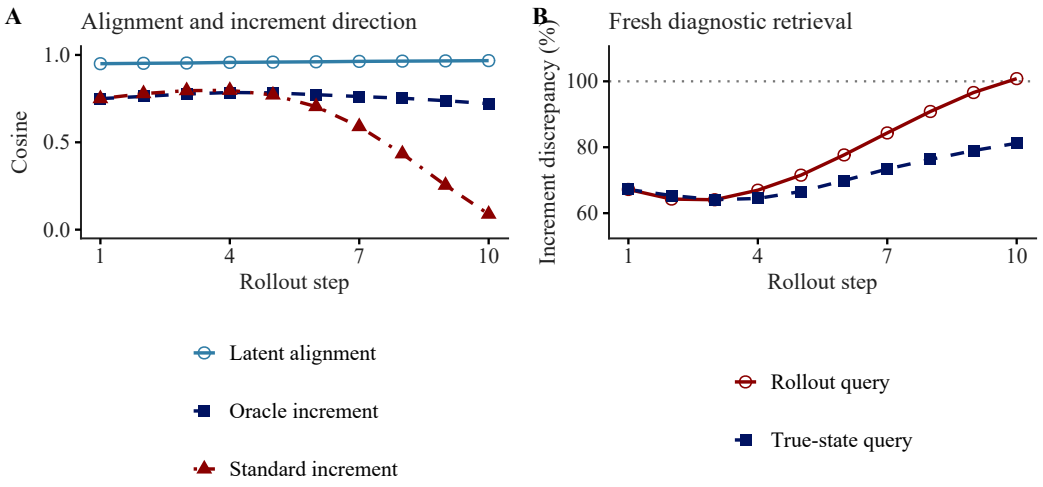


Figure 10: Alignment and retrieval-quality profiles. A: means of latent-state alignment and physical increment-direction cosines; these are different comparisons. B: relative physical-increment norm discrepancies after fresh diagnostic retrieval at each rollout or true-state query, distinct from three-step riding indices. The rollout minimum is 64.071931% at step 3; step 10 is 100.811557%, exceeding the true-state value by 19.597272 pp. A discrepancy of 100% means equal discrepancy and target-increment norms, not zero correlation. Curves show means; intervals are not displayed.

19.597272 percentage points. Here ϵ_ϕ is a relative norm discrepancy between the borrowed and true increments; its value does not by itself determine the angle between them. Figure 10 presents the full mean profiles without intervals. The precise latent-drift normalization remains incompletely specified in the comparison summary.

Table 13: Reported latent-drift endpoints.

Method	$d(1)$	$d(10)$
Relay	0.168426	0.452983
GRU	0.223587	0.384741
FNO	0.409036	0.298289
U-Net	0.073453	0.404303

C.3 The first-step comparison

The frame-matched audit aligns the reconstruction reference and prediction errors at the same frame. The seed-42 reconstruction reference is 16.195261% averaged over frames 10–19, but 7.829778%

at frame 10 alone. The corresponding same-frame relay error is 11.501738%. This audit and the rounded stepwise relay entry in Table 11 retain distinct evaluation labels; their exact first-step numerical reconciliation is not established here.

The raw step-1 Spearman coefficient is $+0.232912$ with interval $[+0.145447, +0.318889]$. Controlling for the true-increment norm and magnitude ratio gives a partial Spearman coefficient of -0.507191, with interval $[-0.568689, -0.432519]$. The paired change, controlled minus raw, is -0.740103 with interval $[-0.820303, -0.648382]$. This recorded adjustment changes the sign of the association; the exact variable definitions and partial-rank implementation are not fully specified in the archived comparison summary.

C.4 Weighted error comparisons

U-Net retains the lower reported error for each measured spectral weighting in Table 14. At $\alpha = 0$ its paired difference from relay is -3.6725 percentage points; at $\alpha = 1$ it is -3.5314 percentage points. The three reported intervals exclude zero, while the intervals at $\alpha = 0$ and $\alpha = 1$ overlap. The exact Fourier weighting operator, zero-mode handling and normalization are not fully reproduced in this comparison summary.

Table 14: Metric audit, U-Net minus relay. Error values are percentages; differences and confidence intervals are percentage points. These metric-audit aggregates retain their own precision and should not replace the main model-comparison aggregates.

α	U-Net	Relay	Difference	95% interval
0	34.6686	38.3411	-3.6725	$[-3.8779, -3.4723]$
1	29.6312	33.1626	-3.5314	$[-3.8017, -3.2596]$
2	30.9576	33.9186	-2.9610	$[-3.3176, -2.5970]$

D Boundary-regime scope

The boundary regime's resolution adequacy is marginal under the specified audit. At $\nu = 10^{-5}$, the ratio $Z_{\mathrm{Nyq}}/Z_{\mathrm{peak}} = 3.945078 \times 10^{-6}$ lies in the questionable interval, and the reported tail slope is -3.2163. Boundary-regime predictor errors, sample coverage and paired comparisons are not reported here. The record therefore supports a resolution warning for this regime but not an additional operating-regime performance conclusion.

E Numerical validation and available comparison data

The numerical checks reproduce the named scalar targets within the discrepancies in Table 15. The latent $\lambda = 0.1$ check differs by 1.75×10^{-7} percentage points. The canonical seed-42 reconstruction check differs by 1.03×10^{-4} percentage points. These validate the specified scalar measurements, not every component of an end-to-end implementation.

Table 15: Recorded reproduction checks. Error targets and measurements are percentages unless noted.

Check	Target	Reproduced	Discrepancy
$\lambda = 0.1$ latent relay	40.529167	40.529167	1.75×10^{-7} pp
Canonical seed-42 reference	16.195261	16.195158	1.03×10^{-4} pp
Physical-output comparison baseline	38.92515182495	38.92515217237	3.47×10^{-7} pp
Raw-field relay regeneration	Not reported	Not reported	4.8×10^{-8} pp
ϵ_ϕ norm reconstruction	Not reported	Not reported	-2.69×10^{-8} pp

Check	Target	Reproduced	Discrepancy
Angular ratio reproduction	18.353420	18.3534465	Relative 1.5×10^{-6}
PCA affine decoder check	$0/1/1$	Table 26	$\leq 1.9 \times 10^{-6}$

Additional numerical comparisons. The increment-discrepancy validation has error -2.69×10^{-8} pp, also reported at rounded precision as -2.7×10^{-8} pp. The angular ratio is 18.3534465. Recorded reference-error values are 34.17 and 34.169479; recorded manifold-separation values are 6.10 and 6.241392. The latter two records are not rounding-equivalent and are not substituted for one another. Source-fitted and target-fitted PCA-36 use different protocols, as specified in Appendix A.

We do not present a method: no variant beats the baseline. Here "variant" means the tested modifications to the diagnostic predictor, evaluated against the appropriate canonical route: the best small-penalty sweep mean is 38.388809% against 38.343243%, and JEPA config2 is 50.8477% against the physical-route 38.92515182%. This statement does not concern the external reference predictors: U-Net's reported 34.72% is lower than relay's. The multiscale explanation remains untested.

E.1 Non-finite PCA cache entries

The available PCA relay fields contain missing tails where the rollout ends at a source-trajectory boundary. The affected target trajectories are 42 at steps 5–10, 133 at steps 8–10, and 462 at step 10. The cache contains 40,960 NaN values and no infinities, corresponding to 10 missing fields; the target inputs at frames 9–19 remain finite. The retained protocol record gives `n_ok=497`, consistent with the shortened cache rows.

This recorded cache problem does not affect the independent source-only PCA pipeline. Its arrays are finite, and the $d = 64$ deployed error is 41.931767%. The separate PCA evaluation reports 41.931800%, with a recorded difference of -0.000033 percentage points. The capacity table can therefore retain PCA while the historical field cache remains unavailable for the six-method weighted reanalysis.

F Oracle comparisons and representation interventions

The nonadditive diagnostic ladder is shown in Figure 11; full-precision operation details remain below.

F.1 A consistent seed for the oracle ladder

The oracle ladder uses seed 42 throughout. Its reconstruction reference is 16.1953%. Its refreshed-base direction-oracle level is 21.8518%. Its free-running direction-oracle level is 35.8606%, compared with deployment at 38.3001%. These levels have conditional differences, not independent additive error sources. The first level evaluates the encoder-selected latent point, not an optimum over latent points.

Table 16: Seed-42 oracle ladder. Differences are percentage points.

Intervention level	RelErr (%)	Adjacent difference
Reconstruction reference	16.1953	Not applicable
Direction oracle, refreshed true base	21.8518	+5.6565
Direction oracle, free running	35.8606	+14.0088
Deployed	38.3001	+2.4395

The refreshed-base comparison remains 5.6565 pp above the reconstruction reference. Retaining the generated latent base adds 14.0088 pp with the same direction-oracle index sequence: this is a

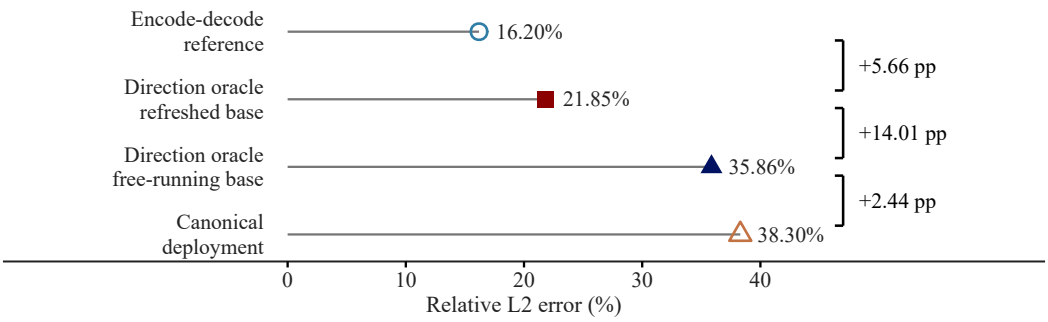


Figure 11: Nested diagnostic ladder. Successive levels use the same encoder, decoder and evaluation set. Annotated gaps are differences between levels, not independent additive error sources. Refreshed and free-running direction-oracle levels use identical indices; retaining generated rather than refreshed true-state bases produces the largest adjacent change. Attained selector improvements depend on the oracle objective, not an optimal-selector performance limit.

state-refresh contrast, not a query-selection contrast. The direction oracle maximizes cosine between each true target physical increment and all 156,000 source physical increments. It does not screen candidates with a latent query. Ties retain the first maximizer within each ordered block and replace the global winner only on strict improvement. The selected latent increment is added to either the true preceding-state encoding or the retained generated state, then decoded. A separate magnitude-labelled oracle minimizes stabilized relative squared physical increment discrepancy and does not rescale increments. Both use a new index each step; neither optimizes final decoded trajectory error. Their distinct improvements cannot certify an optimal-selector performance limit.

For the seed-42 latent relay, the encode–decode reference is 16.1953%. Resetting the latent base to the true preceding state before applying a direction-oracle increment gives 21.8518%. Applying the same oracle index sequence to an unrefreshed generated latent state gives 35.8606%, compared with deployment at 38.3001%. The successive gaps are 5.6565, 14.0088 and 2.4395 pp. These are differences between diagnostic levels, not independent additive error sources: relative L_2 is a ratio of norms. The 16.1953% reference evaluates the encoder-selected latent point, not an optimum over decoder outputs.

The oracle maximizes cosine between a source physical increment and the true target physical increment, without query-dependent screening. Its indices are therefore identical in the refreshed and free-running levels. The 14.0088 pp difference measures the effect of refreshing the latent state; it does not isolate erroneous queries selecting erroneous neighbours. The 5.6565 pp gap is likewise conditional on this particular oracle and output route. Complete implementation details and the numerical ladder are in Appendix F; Figure 1 illustrates the pipeline.

The two oracle levels also share every-step selection cadence, so query selection and cadence do not vary within the 14.0088 pp contrast. A full selection-by-refresh factorial comparison was not run: the free-running direction oracle exists, but canonical selection with state refresh does not. The cadence control in Section 4.2 changes rematching frequency without refreshing the base; none of these contrasts distinguishes accumulated increment error from dynamical amplification.

F.2 Scope of the same-metric diagnostic

The oracle ladder asks how this fixed pipeline's reported RelErr changes under each intervention. Spectral fidelity asks a different question about predicted fields; neither the ladder arithmetic nor an incomplete qualitative ranking establishes a metric-independent decomposition of physical fidelity. Field-level diagnostics are needed before assigning a mechanism.

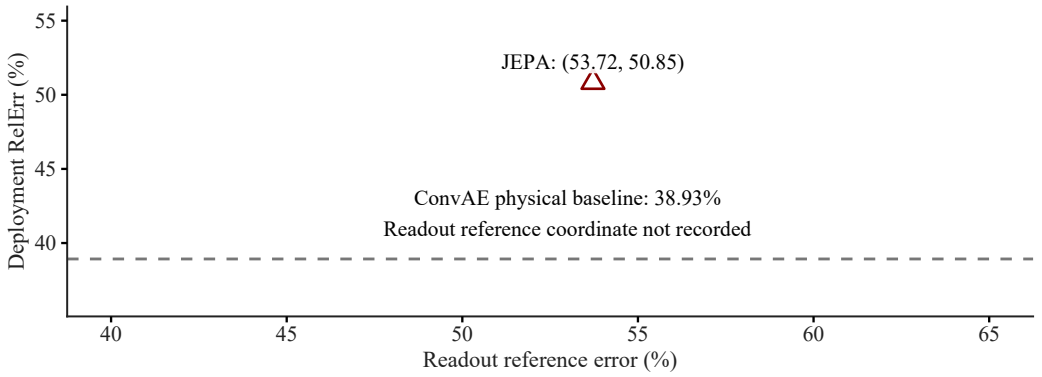


Figure 12: Physical-output readout-reference comparison, separated from the latent-output comparison in Figure 3. JEPA config2 uses its three-seed readout reference. The dashed ConvAE physical-output baseline is a deployment reference only: no corresponding readout-reference horizontal coordinate was recorded, so no such point is fabricated.

F.3 Every-step cadence control and numerical provenance

The every-step canonical control uses the same seed-42 model, 156,000 database entries, target trajectories 0–499 and frames 10–19, changing only $N_{\mathrm{RIDE}} = 3$ to 1. It gives 38.26602058143616%, while the same-code three-step run gives 38.30004704475403%. The paired cadence improvement is 0.03402646331787117 pp, interval $[-0.05314423240184746, 0.1174260830259324]$. The every-step canonical-minus-direction-oracle difference is 2.405381539821625 pp, interval $[2.243558347744942, 2.569481683809757]$. The two mean contrasts sum to 2.439408003139496 pp; intervals are not added. All $5{,}000$ regenerated three-step increment indices match the saved canonical sequence.

The diagnostic-cache baseline is 38.3000470416069% under float64 averaging, not the separately recorded 38.300117% aggregate. Comparing the every-step control to that older aggregate's trajectory array gives 0.034101305770873905 pp with interval $[-0.05307381787538532, 0.11750036029338824]$; these records are not silently interchanged. Decoder batching and reduction conventions differ, but the complete elementwise numerical attribution is not isolated. The new control uses $10{,}000$ paired trajectory resamples, seed 12345. Its newly reset-stream total interval is $[2.2753128169202808, 2.6003215388512606]$, while the earlier shared-stream interval is $[2.273871, 2.604401]$: preceding random draws differ, not the observed contrast. The pre-result absolute cadence prediction was below 0.2 pp, with a rewrite trigger above 0.5 pp; the observed absolute effect is below 0.2 pp, while its interval includes zero.

Figure 12 separates the physical-output readout-reference comparison from the latent-output reconstruction comparison.

Table 17: ConvAE-Relay seed-specific errors. The reconstruction reference is the target reference associated with each seed.

Seed	Latent route	Physical route	Reference
42	38.300117	38.925152	16.195261
123	38.287052	39.068680	13.888114
7	38.442635	39.107353	14.048768
Mean	38.343243	39.033728	14.710714

The three-seed reconstruction mean is 14.710714%, whereas the seed-42 reference is 16.195261%. The corresponding latent-route errors are 38.343243% and 38.300117%. The error ladder must keep the seed-42 values, rather than combine its intervention levels with the three-seed reference mean.

F.4 REPRESENTATION AND CAPACITY MEASUREMENTS

Capacity and reconstruction descriptors do not explain the transfer ordering; the covariance-spectrum participation rank and its scope are retained in Appendix F.

Capacity detail. The seed-42 ConvAE has covariance-spectrum participation rank 10.59921 within its 64-dimensional latent space, where participation rank is $(\sum_j \lambda_j)^2 / \sum_j \lambda_j^2$ for covariance eigenvalues λ_j. PCA-64 has reconstruction reference error 10.98727%, below ConvAE's 16.19526%. Its deployed error is nevertheless 41.93177% against 38.30012%. These measurements describe retained variation and reconstruction, but their ordering does not explain the transfer ordering. No nonlinear intrinsic-dimension measurement is supplied, and it is not inferred from these quantities.

PCA-64 reconstructs with reference error 10.987268% and transfers with error 41.931767%. ConvAE seed 42, at the same latent width, gives reference error 16.195261% and transfer error 38.300117%. The paired PCA-minus-ConvAE difference is $+3.6317$ percentage points, with interval $[3.2635, 4.0007]$. This equal-width comparison contradicts the ordering that would follow from treating reconstruction accuracy alone as a selector of transfer quality on this benchmark.

Table 18: Full representation measurements. The raw-field kNN-copy aggregate uses the shared $N = 500$ evaluation; its supplementary stepwise curve is omitted. JEPA uses a readout reference, rather than the ConvAE reconstruction reference.

Representation	d	Reference	Deployment	ϵ_ϕ
ConvAE, seed 42	64	16.195261	38.300117	70.748903
PCA-8	8	28.254406	43.467083	79.484000
PCA-16	16	22.772667	42.872742	75.258100
PCA-36	36	15.904285	42.250317	74.621200
PCA-64	64	10.987268	41.931767	74.188400
PCA-128	128	7.035465	41.969486	74.072900
PCA-256	256	4.208326	42.005287	74.063600
PCA-512	512	3.471558	42.012421	74.062400
Raw field / kNN-copy	4096	0 by definition	41.012695	70.732704

Widening source-fitted PCA from 64 to 512 reduces its reference error from 10.987268% to 3.471558%, a recorded 3.16-fold improvement. Deployment changes from 41.931767% to 42.012421%. The deployed errors increase slightly beyond $d = 64$; the recorded series is not monotone across the full width range. No nonlinear intrinsic-dimension estimate is supplied, and the covariance participation ranks below measure effective variance directions rather than manifold dimension.

Table 19: Additional matched-width intervention records, $d = 64$. All errors are percentages.

Representation	Reference	Deployment	Participation rank
ConvAE, seed 42	16.195261	38.300117	10.599211
ConvAE, three seeds	Not reported	38.343243	Not reported
Random projection, three seeds	99.27	99.42 ± 0.07	Not reported
JEPA, three seeds	53.7193 (readout)	50.8477	18.29–19.64

Random projection has mean reference error 99.27% and deployed error $99.42 \pm 0.07\%$. Its seed-specific reference errors are 99.2683%, 99.1809%, and 99.3779%. Because the reference error is already near saturation, this control does not isolate geometry from severe loss of recoverable information.

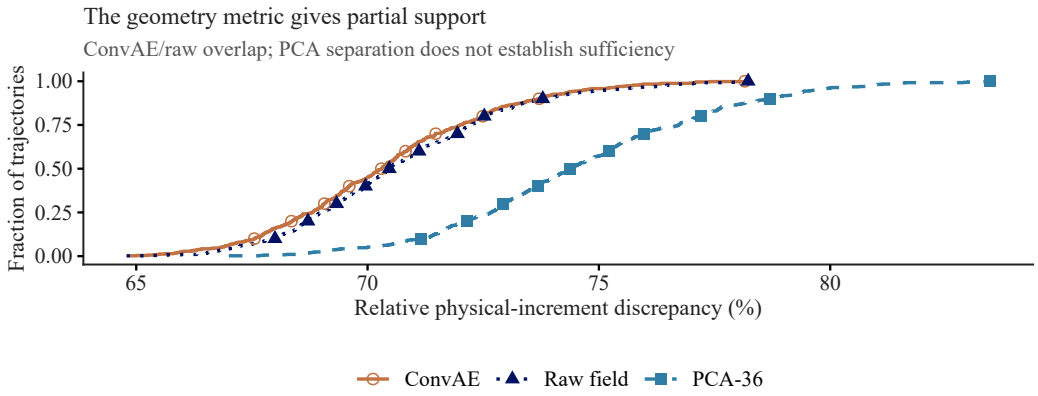


Figure 13: Increment-discrepancy distributions. Each curve uses all 500 ten-step trajectory means; ConvAE first averages the three seeds within each trajectory. Mean discrepancies are ConvAE 70.428636% (seed SD 0.231604 pp), raw fields 70.732704%, and source-fitted PCA-36 74.621200%. PCA uses the source-only fitted basis, not the legacy target-fitted array. The PCA separation and ConvAE/raw overlap give partial support, not a sufficient transfer-ordering metric.

F.5 Angular distances and increment discrepancy

The angular target-to-source nearest-neighbour distance ratio is 18.3534465. The Euclidean ratio is 7.8144152. These ratios compare target-to-source distances with source-internal distances and refer to different distance definitions; the angular value must not be described as Euclidean.

ConvAE's three-seed mean ϵ_ϕ is $70.428636 \pm 0.231604\%$, compared with the seed-42 value 70.748903%. The raw-field value is 70.732704%, while the source-fitted PCA-512 value is 74.062400%. JEPA's true-state values are 91.5431%, 91.4700%, and 93.4423%. These are norm discrepancies of borrowed increments, not correlation coefficients.

F.6 JEPA configurations and the rollout exception

JEPA config2 gives mean transfer error 50.8477%. On the physical-increment route, its paired disadvantage against the 38.92515182% ConvAE baseline is $+11.9226$ percentage points, with interval $[11.3002, 12.5385]$. This fourth intervention runs in the opposite direction to an account in which moving away from a reconstruction objective is sufficient to improve transfer. Because the representations differ, this route-matched comparison does not isolate learned dynamics from retrieved dynamics within one representation.

Table 20: JEPA transfer error by configuration and training seed.

Configuration	Seed 42	Seed 123	Seed 7	Mean	SD
config1a	54.0719	53.3360	52.9498	53.4526	0.4654
config1b	51.1246	50.1430	50.3779	50.5485	0.4185
config2	50.7711	50.4807	51.2913	50.8477	0.3353
config2b	52.2898	52.0937	52.3690	52.2508	0.1157

Table 21: Paired JEPA and matched-width comparisons. Differences and intervals are in percentage points.

Comparison	Difference	95% interval
PCA-64 minus ConvAE seed 42	$+3.6317$	$[3.2635, 4.0007]$
config2 minus physical baseline	$+11.9226$	$[11.3002, 12.5385]$
config2b minus physical baseline	$+13.3257$	$[12.6483, 13.9913]$
config2b minus config2	$+1.4031$	$[1.1772, 1.6236]$
config1a minus readout reference	-0.2667	$[-0.6514, 0.1224]$
config1b minus readout reference	-3.1708	$[-3.6210, -2.7204]$

JEPA's readout references are 53.5862%, 53.5737%, and 53.9981%, compared with the separately measured ConvAE seed-42 readout reference of 29.7748%. The encoder-collapse checks pass, with participation ranks 18.2930, 19.6389, and 18.5224 against a recorded ConvAE reference of 10.5994. These ranks characterize encoder variance and do not rule out changes in the distribution of rollout queries.

Config1b is below its readout reference by 3.1708 percentage points, with paired interval $[-3.6210, -2.7204]$. Its rollout-query angular ratios are 1.7700, 1.7203, and 1.8337, whereas the same encoders on true target states give 22.2924, 22.3709, and 26.1026. The rollout queries are therefore much nearer the source distribution under this recorded comparison. This association is consistent with a change in rollout-state distribution, but it does not establish that change as the cause of the lower error relative to the readout reference.

F.7 The matched dynamics-objective sweep

The lowest nonzero λ improves numerically on the matched sweep control, while remaining above the canonical baseline. Table 22 keeps the matched 38.535474%, selected 38.388809%, and canonical 38.343243% results together. The paired interval excludes zero for seed 42 and includes zero for seeds 123 and 7. These are distinct comparisons and must be reported together.

Table 22: Dynamics-objective scan and both controls. The canonical row is included to prevent interpreting a matched-control improvement as an improvement over the principal baseline. Blank measurements were not reported.

λ / control	Reference (%)	Latent relay (%)	Direction cosine	Rank
Canonical	14.710714	38.343243	Not reported	Not reported
0, matched	14.911939 $\pm$0.682370	38.535474 $\pm$0.044158	0.778	10.09
0.0006	16.437883 $\pm$0.358238	38.388809 $\pm$0.072940	0.804176	11.21
0.002	18.618051	38.927467	0.732507	11.72
0.006	21.066564	39.331062	0.608618	8.59
0.02	22.369968	38.448770	0.677150	5.19
0.06	35.555697	40.851409	0.790999	1.82
0.1	33.270302	40.089286	0.859523	2.96
Earlier 0.1	34.169479	40.529167	0.887761	2.73

Table 23: Paired error difference for $\lambda = 0.0006$ minus matched $\lambda = 0$, in percentage points.

Seed	Difference	95% interval	Excludes zero
42	-0.273896	$[-0.412666, -0.135398]$	Yes
123	-0.095384	$[-0.238155, 0.046700]$	No
7	-0.070721	$[-0.220650, 0.081009]$	No

The earlier $\lambda = 0.1$ representation has participation rank 2.73, alongside reference error 34.169479% and relay error 40.529167%. The scan reaches rank 1.82 at $\lambda = 0.06$ and rank 2.96 at $\lambda = 0.1$. The recorded gradient ratios at $\lambda = 0.1$ are 160.836885, 714.186066, and 3.023386. These observations expose a collapsed-representation confound in the earlier comparison; they do not isolate the causal contribution of collapse from the training changes.

Stage 3 was *not run*: its pre-specified reference-error gate was 15.411939%, and the lowest nonzero-λ reference error was 16.437883%. No nonzero λ qualified. The absence of Stage 3 outcomes is an outcome of the stopping rule, not a null experimental result.

F.8 Paired diagnostic contrasts

The named diagnostic contrasts are reported in Table 24. The canonical-minus-direction-oracle contrast is $+2.439408$ percentage points, with interval $[+2.273871, +2.604401]$, agreeing numerically with the ladder's rounded selection gap. The canonical-minus-magnitude-oracle contrast is $+3.412029$ percentage points, with interval $[+3.256380, +3.570465]$. Their distinct intervention objectives remain explicit.

Table 24: Paired diagnostic contrasts over 500 shared trajectories, with 10,000 bootstrap replicates and seed 12345. Database and retrieval-error contrasts are percentage points; the angle contrast is in degrees, and the last row is a difference of Spearman coefficients.

Contrast	Mean difference	95% interval
Standard minus oracle angle	$+11.155441$	$[+10.723831, +11.589677]$
Database 10% minus 100%	$+0.748518$	$[+0.561617, +0.939335]$
Database 25% minus 100%	$+0.274218$	$[+0.127959, +0.422849]$
Database 50% minus 100%	$+0.189335$	$[+0.076240, +0.305487]$
Canonical minus direction oracle	$+2.439408$	$[+2.273871, +2.604401]$
Canonical minus magnitude oracle	$+3.412029$	$[+3.256380, +3.570465]$
Controlled minus raw Spearman	-0.740103	$[-0.820303, -0.648382]$

F.9 Decoder-increment measurement and affine validation

Untested multiscale hypothesis. Global elliptic coupling in the vorticity formulation motivates an untested hypothesis that a skip-connected hierarchy approximates multigrid. This hypothesis is not a tested explanation of the representation ordering.

The affine PCA decoder gives relative mismatch 5.893121×10^{-7}. Its direction cosine and magnitude ratio are both 1.000000008, within 1.9×10^{-6} of their theoretical values, validating the implementation against the affine reference. ConvAE gives direction cosine 0.988060 and magnitude ratio 0.998800, alongside relative mismatch 0.150035. Its incremental mapping is thus close to affine in direction and scale over the tested source-to-target displacements; this does not identify the encoder as the cause of its advantage. Our magnitude-mismatch expectation and the measurement are compared in Appendix G. The ConvAE measurement gives relative mismatch 0.150035. Its direction cosine is 0.988060 and magnitude ratio is 0.998800. The corresponding 5%–95% ranges are given in Table 25. The measured direction and magnitude are close to the affine reference behavior

Table 27: Seven retrieval-coordinate conditions on the seed-42 latent route, with target trajectories 0–499 and scored frames 10–19. Reconstruction differences from the canonical 16.195261% reference are at most 7.851874×10^{-6} pp. Forced-A0-index deployment differences are at most 9.452343×10^{-6} pp. Orthogonal controls retain all indices; finite-precision differences are retained rather than asserted to be bitwise zero.

A	$\kappa(A)$	Agreement	Ref. (%)	Deployed (%)	Δ (pp)
A0 identity	1	100%	16.195258494	38.300047036	0
A1 orthogonal	1	100%	16.195257010	38.300047751	$+7.15 \times 10^{-7}$
A2 orthogonal	1	100%	16.195258325	38.300044697	-2.34×10^{-6}
A3	2	62.14%	16.195254093	38.309710819	+0.009664
A4	5	38.04%	16.195253148	38.386475960	+0.086429
A5	20	20.80%	16.195262445	38.926753405	+0.626706
A6 source whitening	56.3492	7.50%	16.195254243	43.053873771	+4.753827

at the tested source-to-target displacement scale; this does not identify an encoder contribution to transfer.

Table 25: ConvAE decoder-increment measurements.

Quantity	Mean	5%–95% range
Relative mismatch	0.150035	0.079080–0.232945
Direction cosine	0.988060	0.973601–0.997004
Magnitude ratio	0.998800	0.947213–1.046142

PCA-64 provides an affine-decoder implementation check because its Jacobian does not depend on the latent point. The observed mean relative mismatch is 5.893121×10^{-7}, with maximum absolute deviation 1.856877×10^{-6} from the theoretical value. The observed direction-cosine and magnitude-ratio means are both 1.000000008, with maximum absolute deviations 2.384186×10^{-7}. The validation gate is recorded as passed.

Table 26: PCA-64 affine-decoder validation against known analytic targets.

Quantity	Theory	Measured mean	Maximum deviation
Relative mismatch	0	5.893121×10^{-7}	1.856877×10^{-6}
Direction cosine	1	1.000000008	2.384186×10^{-7}
Magnitude ratio	1	1.000000008	2.384186×10^{-7}

F.10 Fidelity-preserving retrieval-coordinate measurements

We operationalise organisation as the retrieval coordinate frame: the geometry in which cosine matching selects source entries. Section 5.4 intervenes on this frame directly, preserving reconstruction within 7.851874×10^{-6} pp while deployment changes by up to 4.75 pp across seven tested frames. This tests an operational aspect of organisation, not merely a residual explanation after eliminating alternatives; the responsible frame property remains unidentified. For an invertible linear map A, the conjugate encoder and decoder are $E_A = AE$ and $D_A = D \circ A^{-1}$, with source increments $A\Delta z$. Their composition is unchanged analytically. Table 27 records the measured reconstruction and deployment values for the seven fixed coordinate frames. Top-1 agreement counts all $5{,}000$ applied target transitions, including indices carried by the unchanged three-step riding rule. This is distinct from the separate every-step true-state queries used for geometric diagnostics.

Differences and intervals use the saved identity-transform A0 array, whose measured mean is 38.300047036%. The separately evaluated native path gives 38.300047055%; both are retained and satisfy the revised 2×10^{-4} pp deployment tolerance against the recorded 38.300117%.

Table 28: Paired deployment changes relative to saved A0. Whole trajectories are resampled 10,000 times with seed 12345, after averaging their ten scored steps.

Condition	Same / 5,000	Difference (pp)	95% interval (pp)
A0	5000	0	[0, 0]
A1	5000	+0.000000715	[−0.000003100, +0.000004574]
A2	5000	−0.000002339	[−0.000005839, +0.000001187]
A3	3107	+0.009663783	[−0.101900739, +0.125096693]
A4	1902	+0.086428924	[−0.069800541, +0.242212591]
A5	1040	+0.626706369	[+0.428116232, +0.835235767]
A6	375	+4.753826735	[+4.306199421, +5.214601232]

F.11 Additional intervention detail

The four interventions test reduction, reconstruction fidelity, width, and a predictive training objective (Table 2). They establish a decoupling between reconstruction fidelity and deployment ordering. At width 64, random projection gives 99.42% deployed error. Its reconstruction reference error is 99.27%. Their proximity shows that dimension reduction alone does not recover ConvAE's result; because the reference error is already near 100%, this control does not isolate latent geometry.

At equal width, PCA-64 reconstructs the target regime better than ConvAE (10.9873% versus 16.1953%) and transfers worse (41.9318% versus 38.3001%). The paired deployment difference is 3.6317 pp [3.2635, 4.0007]. Equal width and the common latent-output route remove dimensionality and output-route differences from this comparison; the reversal rejects reconstruction fidelity as a sufficient ordering criterion among these encoders.

Increasing PCA width from 64 to 512 improves its reconstruction reference error 3.16-fold, from 10.98727% to 3.471558%. Deployment error instead changes from 41.93177% to 42.01242%, a slight increase. Thus additional retained variation within this linear family does not repair transfer in the tested regime; the deployment curve is not monotone over the full width sweep.

The fourth intervention runs opposite to the inference that any departure from reconstruction training should help: JEPA config2 gives 50.8477%, exceeding the physical-output ConvAE baseline by 11.9226 pp [11.3002, 12.5385]. Its encoder passes the recorded collapse gates, with participation ranks 18.29–19.64 versus the ConvAE reference of 10.60. Encoder collapse therefore does not account for this measured difference under those gates, but the result does not isolate an effect of the predictive objective from the rest of the representation.

JEPA config1b scores 3.1708 pp below its own readout reference [−3.6210, −2.7204], which is possible because that reference is not a lower bound. Its rollout-query angular ratios are 1.7203–1.8337. Its true-target-state ratios are 22.2924–26.1026. Encoder non-collapse and rollout queries approaching the source distribution can coexist. This association does not establish why config1b lies below its readout reference.

G Prior expectations and measured outcomes

We compare our expectations with the measured outcomes. Four expectations were not supported. For spectral weighting, the literal inequality was satisfied; this is not a failed prediction. All comparisons are exploratory, not confirmatory tests.

G.1 Spectral weighting

We expected U-Net's lead at $\alpha = 1$ to fall below 3.62 pp. The measured margin is 3.5314 pp, satisfying that literal statement. U-Net-minus-relay is -3.6725 pp at $\alpha = 0$ and -3.5314 pp at $\alpha = 1$, a recorded reduction of 0.141 pp.

This small reduction does not establish a substantial change. Overlap of the two intervals is not a paired test of their margin change; that paired change interval is unavailable.

G.2 JEPA: A NON-RECONSTRUCTION OBJECTIVE

We expected a representation organised by prediction without reconstruction to transfer better. JEPA config2's measured physical-route disadvantage is $+11.9226$ pp $[11.3002, 12.5385]$; encoder non-collapse and rollout behavior remain separate observations.

G.3 THE LINEAR RECONSTRUCTION REFERENCE

We expected PCA's reconstruction reference error to exceed ConvAE's. The measured PCA-64 reference is 10.987268%, below ConvAE's 16.195261%, while its paired deployment disadvantage is $+3.6317$ pp $[3.2635, 4.0007]$.

G.4 DECODER-INCREMENT MAGNITUDE MISMATCH

We expected magnitude mismatch to explain the reported 41% deficit. ConvAE's measured direction cosine is 0.988060 and its magnitude ratio is 0.998800 at the tested displacement scale. PCA-64's maximum relative-mismatch deviation is 1.856877×10^{-6}, while the maximum direction and magnitude deviations are 2.384186×10^{-7}. The analytic affine check passes; these measurements do not identify the cause of the transfer advantage.

G.5 SOURCE WHITENING

We expected source whitening to improve transfer. The measured outcome is a degradation of 4.753827 pp.

H DIAGNOSTIC RULES, VALIDATION CONTROLS, AND EVIDENCE

Diagnostic eligibility rules, analytic validation controls and interpretations of measured outcomes are distinguished below. The diagnostics remain exploratory.

Query-refresh angle diagnostic. The standard-minus-oracle angle contrast is 11.155441 degrees, with paired interval $[10.723831, 11.589677]$. For physical increment cosine c_{nt}, the diagnostic is $\theta_{nt} = (180/\pi)\arccos(\mathrm{clip}(c_{nt}, -1, 1))$, averaged over trajectories before fitting. A bounded least-squares fit over steps 1–10 uses $at + b\exp(\lambda t)$ with $a, b \geq 0$ and $0 \leq \lambda \leq 2$, no intercept, and the minimum residual over multiple initializations. Full-model rates are 0.1114250485481617 for standard queries and 0.008590823707720109 for true-state rematching. Their fit uncertainty uses 300 trajectory resamples and seed 20260906, not the primary paired-bootstrap protocol. Restricted exponential fits win the recorded AIC comparison in both conditions; these descriptive growth rates of a bounded angle are not measured Lyapunov separation rates. This true-query latent-cosine diagnostic is not the physical-direction optimizer used in the state-refresh ladder.

Curvature estimator and samples. The measured proxy has no statistically significant association with error; this does not reject intrinsic curvature. For 20 Euclidean nearest neighbours excluding the query, center the vectors on their centroid and use singular values σ_j to define $C(x) = \sqrt{\sum_{j>5}\sigma_j^2/20}/\max(r_{20}, 10^{-15})$, where r_{20} is the furthest-neighbour distance. Other frames from the query trajectory are not excluded. This is a rank-5 local affine-residual proxy; negative Gram eigenvalues are clipped to zero. ConvAE source, target and boundary means are 0.19022376858698134, 0.2601859630487629 and 0.30179986539645953. Corresponding raw-field means are 0.15585383134261943, 0.28036540610110217 and 0.3243032566954008; PCA-36 means are 0.15496838158544765, 0.2358299050616054 and 0.26822566750610427. On target trajectories, the ten-step mean proxy and prediction error have Spearman $\rho = -0.07819653678614714$, $p = 0.08066736585254607$, $N = 500$; other condition-matched correlations are not reported. The qualitative expectation of markedly greater boundary curvature is not a quantitative rejection test.

Stage 3 has a recorded numerical eligibility gate and a recorded decision not to run. The gate requires an eligible nonzero λ at reference error 15.411939%, whereas the lowest observed nonzero-λ reference error is 16.437883%. No eligible setting was recorded. Table 29 summarises diagnostic rules and their measured outcomes.

Table 29: Diagnostic rules, numerical validation controls and measured outcomes. These exploratory comparisons do not establish causal exclusions.

Diagnostic	Rule or validation target	Outcome and scope
Stage 3	largest nonzero λ within 0.50 pp of matched zero, giving reference gate 15.411939%.	Lowest nonzero reference 16.437883%; no eligible setting; *not run*.
Resolution audit	Sufficient below 10^{-6}; questionable from 10^{-6} to 10^{-4}; under-resolved above 10^{-4}.	Source and target sufficient; boundary questionable.
Spectral weighting	Predicted smaller U-Net lead at $\alpha = 1$ than the stated 3.62-point $\alpha = 0$ lead.	Measured 3.5314 pp satisfies the literal inequality; recorded shrinkage 0.141 pp. The literal expectation is satisfied, not a failed prediction; substantial change is not established.
JEPA collapse checks	Checks use angular, variance/rank and increment diagnostics; nonfinite/degenerate results require stop.	Participation ranks 18.2930, 19.6389, 18.5224. No fixed quantitative cutoff is specified for this diagnostic inventory.
JEPA seed spread	range over three seeds exceeding 2.0 pp requires stop/report.	This rule is not an additional seed-spread measurement.
JEPA reproduction	physical-route 38.92515182495117%, tolerance 0.01 pp.	Measured reproduction retained.
Historical λ reproduction	40.529167%, tolerance 0.05 pp.	Scalar reproduction retained.
PCA-64 affine validation	Known theoretical values $0/1/1$ for mismatch, direction cosine, and magnitude ratio.	Recorded PASS; deviations at most 1.9×10^{-6}.
Eight-explanation inventory	Evidence forms and candidate-specific controls.	Benefits, adverse effects, inconclusive association and an incomplete phase test; no common causal-exclusion grade.
Similarity-metric cancellation	Oracle performance-limit reasoning.	The performance-limit premise is withdrawn; attained oracle improvements do not bound the best possible selector.
Retrieval-coordinate implementation controls	Validation tolerance: 2×10^{-4} pp and exact orthogonal indices in amended rule.	Measured reconstruction and deployment agree within tolerance; orthogonal conditions retain all indices.
Retrieval-coordinate outcome categories	Operable / weak / ineffective / reverse used after measurement.	Descriptive classification of measured outcomes, not a predictive criterion.

Retrieval-coordinate outcome framework. The four-way partition (operable / weak / ineffective / reverse) is a descriptive classification of measured outcomes. It does not distinguish an isolated anomaly from a systematic monotone trend. The observed nonorthogonal conditions are described under "reverse"; this classification is not a predictive criterion.

H.1 EIGHT CANDIDATE EXPLANATIONS AND EVIDENCE FORMS

The four descriptive evidence families are observed changes with their sign, no observed benefit from tested variants, no significant measured-proxy association, and an incomplete test. Table 30 retains candidate-specific scope rather than treating these families as equal causal identification.

Table 30: Eight candidate explanations and measured evidence. Missing tests are not negative results.

	Candidate	Evidence form	Finding and scope
1	Mismatch growth	Observed intervention effect	Standard minus true-query angle is $+11.155441$ degrees, interval $[+10.723831, +11.589677]$. Not a Lyapunov measurement or a mechanism exclusion.
2	Insufficient context	No beneficial tested effect	Three variants are worse, with paired intervals excluding zero; only these variants were tested.
3	Database coverage	Positive effect; diminishing gains	10%, 25%, 50% subsets are worse than 100% by 0.748518, 0.274218, 0.189335 pp; all three intervals exclude zero. No saturation beyond tested fractions established.
4	Similarity metric	Objective-dependent improvement	Canonical minus direction and magnitude-labelled mismatch oracles: $+2.439408$ and $+3.412029$ pp. Larger attainable improvement unknown; cadence control in Appendix F.
5	Curvature	No significant proxy association	Spearman $\rho = -0.0782$, $p = 0.0807$ for the rank-5 affine-residual proxy; not a rejection of intrinsic curvature.
6	Phase coordinates	Test incomplete	No phase-aligned relay-control array; no substitute constructed.
7	Dynamics objective	Adverse high-penalty effect	Participation rank 10.09 to 1.82; small matched improvement remains, and Stage 3 was not run.
8	Step-1 anomaly	Adjusted association reversal	Raw $+0.232912$ becomes partial -0.507191; controlled-minus-raw -0.740103.

Phase records exist but no phase-aligned relay-control outcome arrays were retained; no substitute control was constructed. Distance concentration is a contextual issue, not a ninth tested candidate.

H.2 FULL CANDIDATE-EXPLANATION RECORDS

Larger databases improve performance within the tested range; their diminishing gains do not make coverage irrelevant. Stable phase coordinates remain untested as a performance intervention because no phase-aligned relay-control array exists and no substitute was constructed.

Mismatch growth and query refresh Replacing generated-state cosine queries and three-step riding with true-state queries rematched every step reduces the standard-minus-oracle angle by 11.155441 [10.723831, 11.589677]. This query/cadence diagnostic is distinct from the 14.0088 pp state-refresh contrast, whose physical-direction indices are fixed. Neither comparison distinguishes accumulated error injection from dynamical amplification; the angle fit is not a Lyapunov separation measurement.

Insufficient temporal context If missing history were the operative limitation, adding temporal context should have improved this relay. Three tested variants are worse, with paired intervals excluding zero. Concatenated keys $[z_t, z_{t-1}]$, $[z_t, z_{t-1}, z_{t-2}]$ and $[z_t, z_t - z_{t-1}]$ give RelErr 38.539794921875%, 38.80649185180664% and 38.598026275634766%, respectively. Their paired differences from baseline are 0.19655372202396393, 0.46324869990348816 and 0.2547825276851654 pp, with intervals

$[0.10411133579909801, 0.2928044386208057 4]$, $[0.3301439702510834, 0.5943509340286254]$ and $[0.17726647965610026, 0.33382555916905404]$. Keys are concatenated then unit-normalized; stored increments, decoder and $N_{\mathrm{RIDE}} = 3$ remain unchanged. These intervals use 2,000 trajectory resamples with seed 20260906, not the primary $10{,}000/12345$ protocol. The supported disposition is limited to the tested variants, not temporal information in general.

Insufficient database coverage Increasing database coverage might have removed the transfer deficit. Relative to the full database, errors at 10%, 25%, and 50% coverage are higher by 0.748518, 0.274218, and 0.189335 pp, with respective intervals [0.561617, 0.939335], [0.127959, 0.422849], and [0.076240, 0.305487]. Gains diminish over these tested fractions, but every interval excludes zero. Larger databases provide a beneficial performance effect, with diminishing gains over the tested fractions; calling coverage irrelevant would erase those improvements.

A better similarity metric Two specific oracle objectives improve over canonical: direction cosine by 2.439408 pp [2.273871, 2.604401], and relative squared physical increment mismatch by 3.412029 pp [3.256380, 3.570465]. Whether larger gains are attainable is not established. Both change selection cadence as well as objective; the same-cadence contrast and its provenance are given in Appendix F.

Curvature as a missing second-order quantity The measured local affine-residual proxy has target-trajectory Spearman $\rho = -0.0782$, $p = 0.0807$. This is no statistically significant association under the measured proxy, not evidence that intrinsic curvature is irrelevant. Regime means and the estimator appear above.

Stable phase coordinates No phase-aligned relay-control outcome array exists, so the performance test is incomplete. Source and target phase measurements alone are not that control, and no substitute was constructed.

A dynamics-aware encoder We initially read apparently favourable increment geometry as evidence that the dynamics penalty was helping. The earlier $\lambda = 0.1$ run has direction cosine 0.887761 but participation rank only 2.73. Its deployed error is 40.529167%. The matched $\lambda = 0$ run gives rank 10.09 and error 38.535474%. That geometric improvement occurred in a collapsed representation and therefore cannot by itself establish more transferable dynamics. The small-penalty matched improvement in Table 3 is a local exception: its canonical comparison and seed intervals remain part of the same account.

A first-step anomaly A positive raw association suggested a first-step anomaly. Raw Spearman correlation is 0.232912 [0.145447, 0.318889], but controlling true-increment norm and magnitude ratio yields -0.507191 [-0.568689, -0.432519], a controlled-minus-raw difference of -0.740103. Independently, the frame-matched reconstruction reference is 7.829778% and relay error is 11.501738%, whereas averaging the reference over scored frames gives 16.195261%. Both checks show why the unadjusted association and the unmatched time average cannot sustain the original anomaly.

Distance-concentration context. Ambient width alone does not establish a nearest-neighbour concentration failure. The representation comparisons above motivate examining task-relevant geometry, but no result here excludes every distance-concentration effect. This discussion supplies context for the instrument rather than another counted diagnostic.

A3: resolution audit

Summary only; full enstrophy spectra not displayed

Regime	Z_Nyq/Z_peak	Tail slope	Assessment
Source (nu=1e-3)	8.731661e-18	-3.6700	Adequate
Target (nu=1e-4)	2.201566e-11	-12.9202	Adequate
Boundary (nu=1e-5)	3.945078e-06	-3.2163	Marginal

Figure 14: Three-regime resolution-audit summaries (A3). Full enstrophy spectra are not shown; this is not a grid-convergence experiment.

A4: distance-ratio summary

Full empirical distance distributions are not displayed

Metric (target/source)	Ratio
Angular nearest-neighbor distance	18.3534465
Euclidean nearest-neighbor distance	7.8144152

Figure 15: Angular and Euclidean nearest-neighbour ratios (A4). These metrics are distinct; full empirical distance distributions are not shown.

Figures 14, 15, 16 and 17 give the supplementary resolution, distance, dynamics-objective and JEPA summaries, respectively.

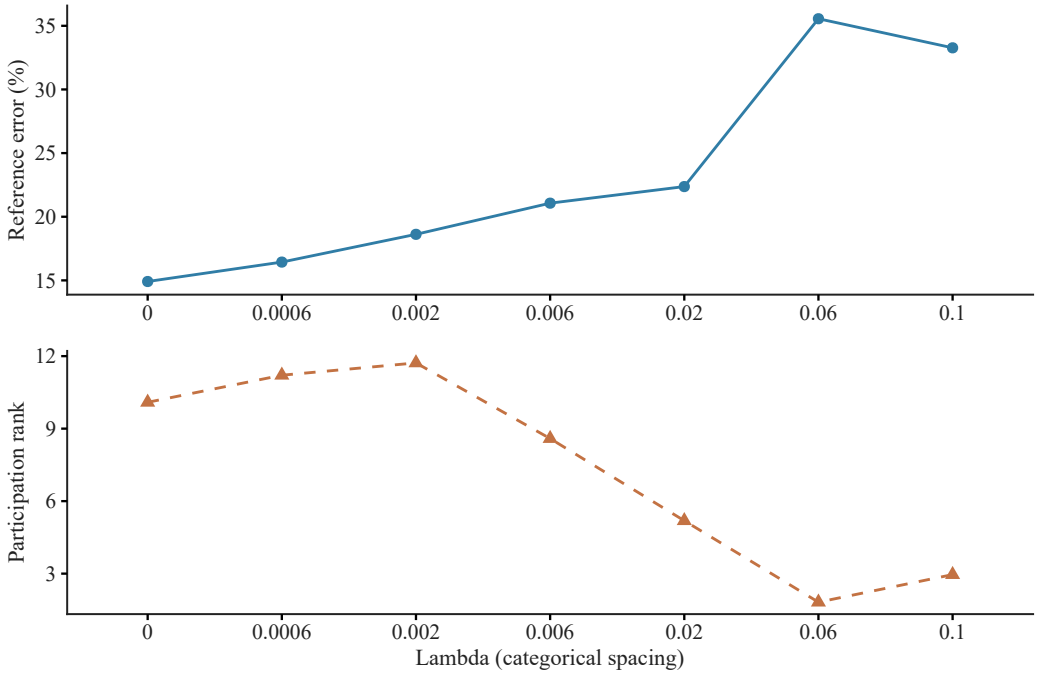


Figure 16: Matched λ sweep reconstruction reference and participation-rank summaries (A6). Matched zero is explicit. Deployment values remain in Table 22; that series is non-monotonic. No unavailable uncertainty is supplied.

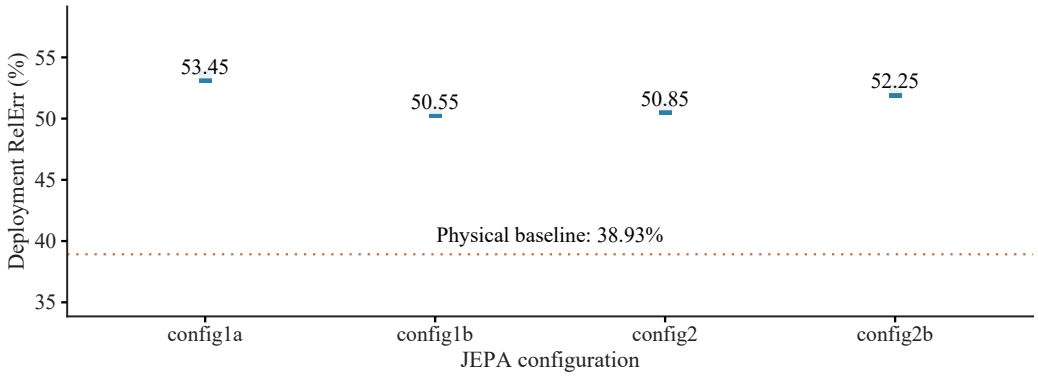


Figure 17: JEPA configuration deployment means (A7), with the separate physical-output ConvAE baseline. Readout references are reported separately in Table 19; they are not lower bounds. No unavailable uncertainty is supplied.

I ADDITIONAL LITERATURE CONTEXT

Parameter transfer and temporal models. Transfer studies examine scaling and adaptation across physical systems (Subramanian et al., 2023), while symmetry-based augmentation changes the training distribution through equation symmetries (Brandstetter et al., 2022a). A recent bias-aware benchmark explicitly separates physical regimes and distribution shifts (Chu et al., 2026). Structured state-space sequence models (Gu et al., 2022) and latent-space control models (Hansen et al., 2024) provide additional temporal-model context; neither citation identifies our RSSM-style implementation with those architectures.

Analogs and nonlinear dynamics. Local prediction of chaotic time series (Farmer & Sidorowich, 1987), learned analog forecasting of weather patterns (Chattopadhyay et al., 2020), and analog-based turbulent-velocity forecasting (Frogé et al., 2025) concern distinct observations and forecasting protocols. They establish related uses of state similarity, not directly comparable scores for the present field benchmark. Learned dissipative-system dynamics likewise require attention to long-time behavior (Li et al., 2022).

Predictive representations in physical tasks. Recent preprints extend joint-embedding prediction to PDE control and physics generalization (Gallagher & Guglielmi, 2026; Liu et al., 2026); general self-supervised objectives are also studied in LeJEPA (Balestriero & LeCun, 2025). These references concern their stated tasks and versions, rather than validation of our particular intervention. Work separating reconstruction from perceptual representation quality (Balestriero & LeCun, 2024) supplies a related distinction, not a theorem about this PDE benchmark.

Metric geometry and dimension. Metric learning changes the distance used for retrieval (Kulis, 2012), and hubness concerns the uneven frequency of nearest-neighbour selection (Radovanović et al., 2010). Random-projection results and applications address distance preservation under their specified assumptions (Johnson & Lindenstrauss, 1984; Bingham & Mannila, 2001); they do not promise field reconstruction fidelity for our control. Maximum-likelihood and two-neighbour estimators measure intrinsic dimension through different constructions (Levina & Bickel, 2004; Facco et al., 2017). Covariance-spectrum participation ratios instead summarize variance directions (Gao et al., 2017); none of those nonlinear estimators was measured here.

Evaluation and spectral diagnostics. The Well and RealPDEBench broaden the range of physical systems and data sources used in evaluation (Ohana et al., 2024; Hu et al., 2026). WeatherBench 2 discusses evaluation design and caveats for weather prediction (Rasp et al., 2024); it is not the dataset used here. Spectral representation and spectral bias have also been examined in turbulent-flow prediction and operator learning (Oommen et al., 2025; Khodakarami et al., 2026). Together with the reporting-bias and spectral-audit references in Section 2, these precedents preclude a blanket claim that evaluation auditing is absent from the literature.

Physical and visual context. Classical two-dimensional turbulence studies establish the relevance of scale-resolved dynamics (Kraichnan, 1967; Boffetta & Ecke, 2012); they do not establish the resolution adequacy of this benchmark. Figures use a fixed five-colour palette with redundant line styles, symbols, direct labels, and independently checked grayscale outputs, following colour-accessibility guidance (Okabe & Ito, 2008). The reused illustrative fields retain their original continuous colormap. The customized palette is not claimed to be perceptually uniform.