A DISTRIBUTIONAL EVALUATION METRIC FOR CHAOTIC PHYSICAL SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon predictions of chaotic physical systems eventually decorrelate from reference trajectories, making pointwise errors unreliable for distinguishing simulation failures from plausible alternative evolutions. We address this evaluation gap by comparing distributions of predicted and reference states using a limited number of independent trajectories. We introduce PDE-KD, a distributional evaluation metric for neural physics simulators based on a kernel discrepancy between predicted and reference distributions of states embedded in a learned feature space. To evaluate feature spaces for this application, we assess their statistical discriminative power under controlled perturbations of the spatial field and the initial conditions. PDE-KD is sensitive to disturbances across a broad range of spatial scales on several physical systems; it is sample efficient, requiring few trajectories to reliably detect discrepancies, and it reliably distinguishes conditional distributions. On trained neural simulators, it orders the models of each family consistently with measures of rollout quality, whereas pixel-space errors favor over-smoothed predictions. We thus present PDE-KD as a general, sample-efficient distributional evaluation metric that complements pointwise errors and targeted physical diagnostics for model evaluations.

1 INTRODUCTION

Chaotic systems are challenging to predict over long horizons because small perturbations, arising from numerical and modeling errors, grow exponentially with time until trajectories decorrelate. After decorrelation, trajectory-by-trajectory comparisons are no longer informative, but the distribution of states generated from the initial-condition distribution is well defined and typically robust to small perturbations. Evaluation metrics for simulators of chaotic systems should thus compare distributions of states, including their dependence on the initial conditions at intermediate times, reliably and from a small number of independent simulations.

Existing evaluation metrics. Existing approaches assess several aspects of PDE surrogate predictions beyond pointwise trajectory agreement. Spectral and conservation errors measure the distribution of energy across scales and the preservation of conserved quantities, respectively (Takamoto et al., 2022; Kohl et al., 2026). These diagnostics remain interpretable after chaotic decorrelation, with each targeting a particular physical property. Probabilistic scoring rules assess the distribution of possible outcomes. The continuous ranked probability score (CRPS), widely used in weather forecasting (Gneiting & Raftery, 2007), has also been applied to probabilistic PDE emulators (Diaconu et al., 2026). Applied pointwise, it evaluates marginal predictive distributions, but does not directly assess spatial dependencies within fields or temporal dependencies across frames. These metrics are complementary, yet each characterizes targeted or low-order statistics rather than the joint structure of the physical state.

Feature-space metrics. Image generation faced a similar problem: pixel-level and likelihood-based measures correlate poorly with sample quality (Theis et al., 2016). Generative image models are now evaluated by comparing populations in learned feature spaces, most commonly with the Fréchet Inception Distance (FID) and the Kernel Inception Distance (KID) (Heusel et al., 2017; Bińkowski et al., 2018). These metrics were adopted after being compared with alternatives on controlled image distortions and at finite sample sizes (Heusel et al., 2017; Xu et al., 2018). Fréchet Poseidon Distance

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

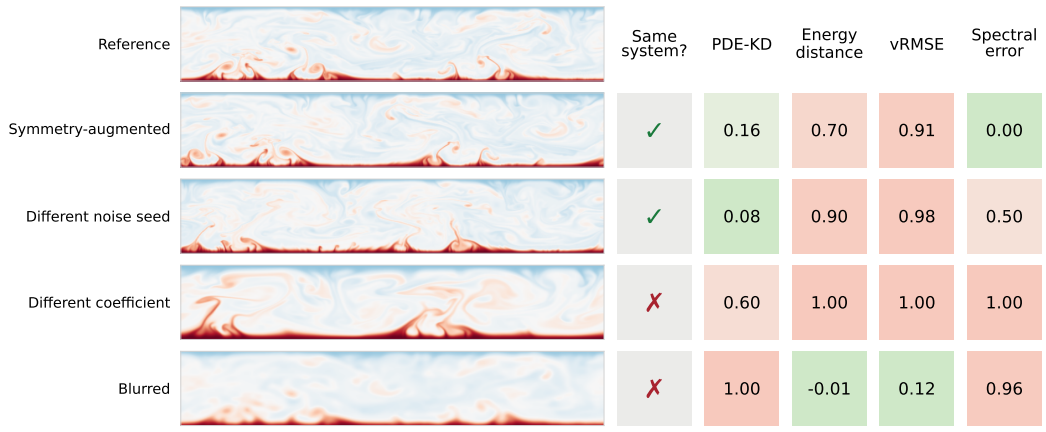


Figure 1: **Can evaluation metrics tell whether these are the same system?** One realization of chaotic Rayleigh–Bénard convection (top) against a symmetry-augmented copy, a different initial-noise seed, a different Ra (a training-split trajectory; Ra is a supervised target of the PDE-KD feature map) and a blurred copy. vRMSE: variance-normalized RMSE. Each column is scaled so its most severe response is 1.00 (more systems: Appendix A).

(FPD) (Cheng et al., 2025) first applied FID to PDEs, replacing Inception features with pretrained Poseidon features. To our knowledge, no comparable study of population metrics exists for PDEs.

Evaluation goal. Our goal is to distinguish distributions of simulated physical states through differences in their physically relevant statistical structure, without requiring pointwise agreement between individual chaotic trajectories. A useful metric should detect discrepancies in properties such as spatial organization and energy distribution across scales, while accounting for natural variability among realizations of the same distribution. The practical challenge is to identify such discrepancies reliably from a limited number of independent trajectories.

A "metric of metrics." We assess metrics by their ability to detect physically meaningful distributional differences at a fixed false-positive rate. We compare discriminative power at a given sample size and the number of independent trajectories needed to reach a target power, providing a common basis for evaluating sensitivity and sampling variability across metrics with different numerical scales.

Contributions.

- We introduce PDE-KD, a kernel discrepancy between generated and reference frame populations in a physics-supervised feature space (Section 3).
- We evaluate feature spaces by *discriminative power* and *sample efficiency*: test power at a fixed false-positive rate, and the trajectories needed to reach it (Section 4).
- On controlled disturbances across five physical systems, PDE-KD rejects most disturbance families, and more spectral families than the other feature spaces we compare (Section 4.1); it also distinguishes initial-condition populations and orders trained neural simulators consistently with measures of rollout quality (Sections 4.2 and 4.3).

2 Problem Setting and Related Work

2.1 Distributional Evaluation

We study the evaluation of physical simulations as a two-sample problem: given samples from distributions P and Q, can a metric reliably detect physically meaningful differences between them? We test $H_0 \colon P = Q$ against $H_1 \colon P \neq Q$, using each metric as a test statistic. We compare discriminative power at a fixed false-positive rate and the number of independent trajectories required to reach a target power. Controlled perturbations of the fields, initial conditions and coefficients probe which differences each metric is sensitive to.

Dynamics-induced distributions. We consider two-dimensional PDE systems with states $u \in \mathcal{U} \subseteq L^2(\Omega; \mathbb{R}^C)$, where $\Omega = [0,1]^2$ and C is the number of physical channels. For a fixed PDE family and parameter setting, let S_t and $\widehat{S}_t$ denote the reference and surrogate evolution maps. Hats denote surrogate quantities rather than empirical estimates.

Given an initial-condition distribution ν, the state distributions at lead time t are

$$\mu_t = (S_t)_{\#}\nu, \qquad \widehat{\mu}_t = (\widehat{S}_t)_{\#}\nu, \tag{1}$$

where $\#$ denotes pushforward: these are the distributions of states obtained by evolving initial conditions sampled from ν. We present deterministic dynamics for simplicity; stochastic surrogates additionally average over their internal randomness.

Frame populations. To evaluate a finite lead-time window, we sample observation times independently of initial conditions according to a distribution ρ, such as a uniform distribution over the saved times in that window. This gives

$$\mu = \int \mu_t \, \rho(dt), \qquad \widehat{\mu} = \int \widehat{\mu}_t \, \rho(dt). \tag{2}$$

The reference and surrogate use the same initial-condition distribution ν and observation-time distribution ρ. These choices specify the populations being compared.

Ergodicity. For dynamics ergodic with respect to an invariant probability measure, long-time averages of integrable observables equal their expectations under that measure for almost every initial state with respect to the same measure. Comparing reference and surrogate invariant measures therefore assesses agreement in long-time statistical behavior without requiring pointwise agreement between trajectories. This motivates distributional evaluation, although our finite-window comparisons require neither ergodicity nor an asymptotic limit.

Conditioning on initial conditions. We also compare different initial-condition populations evolved under the same dynamics. Let ν_A and ν_B denote two such distributions, for example obtained by conditioning on different initial-condition sets. At lead time t, they induce

$$\mu_t^A = (S_t)_{\#}\nu_A, \qquad \mu_t^B = (S_t)_{\#}\nu_B. \tag{3}$$

Comparing these populations tests whether the initial-condition differences remain detectable after evolution, while holding the dynamics and observation time fixed.

Spatiotemporal populations. To assess temporal structure, we also compare distributions of trajectory segments represented as fields $v \in L^2(\Omega \times [0,\tau]; \mathbb{R}^C)$, where τ is the segment duration. A reference sample is

$$v(x,s) = \big(S_{t+s}(u_0)\big)(x), \qquad s \in [0,\tau], \tag{4}$$

with the surrogate sample defined analogously using $\widehat{S}$. Initial conditions and segment-start times follow the same sampling distributions for both systems. These samples retain temporal ordering and dependence, which pooled frame populations discard. In practice, the fields are represented on discrete spatial and temporal grids.

Feature-space comparison. Writing μ and $\widehat{\mu}$ for the selected reference and surrogate populations, we compute

$$D(\phi_{\#}\widehat{\mu}, \phi_{\#}\mu), \tag{5}$$

where ϕ is a frozen frame or clip feature map and D is a distributional discrepancy. The same construction applies to the two initial-condition populations defined above. The feature map and discrepancy determine which differences the metric can detect. Sampling budgets and test calibration use independent trajectories as the sampling unit, accounting for dependence among observations within each rollout. Population agreement does not establish forecast fidelity for each individual initial condition.

2.2 Distributional and Feature-Space Metrics

Kernel two-sample discrepancies. The squared maximum mean discrepancy (MMD) between distributions P and Q under a kernel k is

$$\mathrm{MMD}_k^2(P,Q) = \mathbb{E}\big[k(X,X')\big] - 2\,\mathbb{E}\big[k(X,Y)\big] + \mathbb{E}\big[k(Y,Y')\big], \tag{6}$$

Table 1: Complementary approaches to evaluating physical simulations. Population comparisons assess distributions of states, while paired errors and probabilistic scores assess predictions against individual reference outcomes.

Metric	Space	Comparison	Captures
vRMSE	Pixel	Paired	Pointwise error
Spectral error	Fourier	Paired	Energy across scales
Energy score	Pixel	Ensemble vs. outcome	Multivariate predictive quality
PDE-KD (ours)	Features	Population	Learned field structure

where $X, X' \sim P$ and $Y, Y' \sim Q$ are mutually independent and the expectations are finite (Gretton et al., 2012). For two independent i.i.d. samples, its U-statistic estimator is unbiased and can be negative at finite sample size. Under the null hypothesis, exchangeability permits permutation tests with a controlled false-positive rate (Gretton et al., 2012). For trajectory data, calibration must additionally respect the dependence among observations from the same rollout.

FID and KID. FID (Heusel et al., 2017) fits Gaussians to Inception-v3 features (Szegedy et al., 2016), and its plug-in estimator is biased at finite sample size; KID (Bińkowski et al., 2018) instead estimates MMD_k^2 in the same features using the cubic kernel

$$k(z, z') = \left(1 + \frac{z^\top z'}{d}\right)^3, \tag{7}$$

where d is the feature dimension. This discrepancy compares feature moments up to order three without a Gaussian assumption; equality of these moments does not imply equality of the full distributions. For an arbitrary feature map, we call MMD_k^2 with this kernel the **kernel discrepancy** (KD), and its U-statistic estimate the KD score.

2.3 Pointwise, Physical, and Probabilistic Metrics

vRMSE. Variance-normalized RMSE measures error against a paired reference relative to its spatial standard deviation (Ohana et al., 2024). After chaotic decorrelation, a large pointwise error alone cannot distinguish physical failure from a plausible alternative evolution: in Figure 1, realizations of the same system incur errors comparable to a change in the PDE coefficients, while a blurred field incurs little error. Distributional comparisons instead assess whether populations retain the statistical properties of reference simulations.

Spectral error. Spectral diagnostics compare the power $|(\mathcal{F}u)(k)|^2$ of each predicted field with that of its paired reference field across spatial scales, where $\mathcal{F}$ denotes the Fourier transform and k the wavenumber (Rozet et al., 2025). We report the relative power error in four radial wavenumber bands, averaged over frames. They reveal excess or missing energy at particular scales, but matching power spectra does not ensure matching spatial structure.

CRPS and energy score. CRPS evaluates univariate predictive distributions, while the energy score evaluates multivariate outcomes (Gneiting & Raftery, 2007). Both have the form

$$\mathcal{S}(Q, u) = \mathbb{E}_{X\sim Q}\big[d(X, u)\big] - \tfrac{1}{2}\,\mathbb{E}_{X,X'\sim Q}\big[d(X, X')\big], \tag{8}$$

where Q is a predictive distribution, u is the observed reference outcome, and X, X' are independent. The distance d is, respectively, absolute distance for scalar outcomes or Euclidean distance for multivariate outcomes. Gridwise CRPS assesses marginal predictive distributions but does not directly assess their spatial dependence. It is widely used in weather forecasting (Price et al., 2025; Lang et al., 2026) and probabilistic PDE modeling (Diaconu et al., 2026). These scores and targeted physical diagnostics provide complementary information to our evaluation of distributional discriminative power and sample efficiency (Table 1).

3 Approach

We compare reference and model-generated populations in a learned feature space: PDE-KD is the discrepancy $D(\phi_{\#}\widehat{\mu}, \phi_{\#}\mu)$ of Section 2.1, with D the KD of Section 2.2 and ϕ the feature map of a

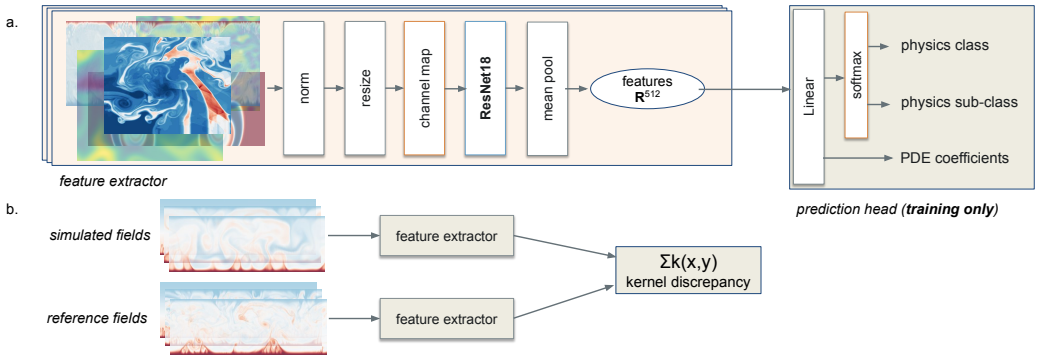


Figure 2: **The PDE-KD feature map and score. (a, left)** Single frames are normalized, resized and mapped to channel slots, then embedded by a ResNet-18 (mean-pool features). **(a, right)** Training supervises the physics class (PDE family), sub-class (within-family regime) and PDE coefficients. **(b)** KD between the feature distributions of simulated and reference trajectories (Appendix B).

small convolutional model trained to predict the *PDE family*, coefficients and *within-family regime* of *single time-step snapshots* from nine PDE families, dataset classes that span six PDE systems (Figure 2). The feature map determines which physical differences this discrepancy is sensitive to.

3.1 Feature-Map Design

We use a ResNet-18 (He et al., 2016) as a feature map and train it with three joint supervised tasks: PDE-family classification, PDE-coefficient regression, and categorical within-family regime classification. We obtain the *features*, used for KD, directly from the mean-pool layer of the shared pre-head embedding, $\phi(x) \in \mathbb{R}^{512}$.

Training dataset. We use nine PDE families from The Well (Ohana et al., 2024) and PDEgym (Herde et al., 2024), detailed in Appendix B.1. We present frames from all families in mixed batches. Every frame contributes to a global nine-way family-classification task and to one family-specific task, either regression on PDE coefficients or classification on categorical regimes, depending on the PDE family.

Inputs and prediction head. Each frame is z-scored channel by channel with statistics of its own PDE family, resized to 512×512, and written into a 13-slot channel dictionary whose slots are generic capacity roles rather than shared physical variables (Appendix B.2). A single linear head on $\phi(x)$ produces the outputs of all ten tasks; each frame's family label selects its own family's outputs, and the other families' outputs are masked out of the loss (Appendix B.3).

Objective. We use the batch loss function

$$\mathcal{L} = \mathcal{L}_{\text{family}} + \beta\, \mathcal{L}_{\text{regime}} + \alpha\, \mathcal{L}_{\text{coeff}}, \tag{9}$$

with tunable hyperparameters α and β. Within a batch, $\mathcal{L}_{\text{coeff}}$ is a mean-squared-error loss masked to families with swept coefficients, and $\mathcal{L}_{\text{regime}}$ is a softmax cross-entropy loss masked to families with categorical regimes. We use $\alpha = 2$ and $\beta = 0.5$.

Model size and training budget. We use the ResNet-18 implementation in Flax, with 11.2 million trainable parameters. We train on Google Cloud TPU-v5e over $1,298$ TPU-hours, sampling the nine PDE families with equal weight. We use the AdamW optimizer (Loshchilov & Hutter, 2019) for $22,176$ steps, with a warmup–steady–decay learning-rate schedule and peak learning rate 10^{-4}. Each realized unique train frame is seen about 3.2 times on average (3.17–3.18; Table 8).

Spatiotemporal feature map. To score trajectory segments (Section 2.1), we inflate a longer-trained feature-map checkpoint to 3D (Carreira & Zisserman, 2017): every convolution except the 1×1 shortcut projections gains a temporal width of three frames, initialized so that the network reproduces that checkpoint frame by frame. We fine-tune it on 20-frame clips with the same training data and objective (equation 9) and use the mean-pool layer as its feature, $\phi(v) \in \mathbb{R}^{512}$ (Appendix C.3).

3.2 COMPUTING KD

To evaluate a model on a PDE family, we compare a *reference* arm, the frames of reference simulations from a set of initial conditions (ICs), with a *generated* arm, the frames of the model's rollouts from the same ICs and coefficients. Each arm pools all frames in a lead-time window, across ICs and across the family's coefficient settings. We embed every frame with the frozen feature map ϕ and report $\mathrm{KD}(\phi_{\#}\widehat{\mu}, \phi_{\#}\mu)$, where $\widehat{\mu}$ and μ are the pooled generated and reference frame distributions.

Frames of one trajectory are strongly correlated, so we treat the IC, not the frame, as the unit of independence: the U-statistic of Section 2.2 omits every pair of frames that share an IC. With ICs drawn at random, the score is unbiased, and it can be slightly negative when the two arms match. Estimator details and the variant each experiment uses are in Appendix D.

4 METRIC VALIDATION

We validate our metric, and its underlying learned features, by assessing its discriminative power and sample efficiency on controlled on-manifold and off-manifold disturbances, in comparison with other alternative feature spaces. *Discriminative power* is the power of the KD two-sample test (Sections 2.2 and 3.2); *sample efficiency* is the number of independent trajectories needed for 80% power. The kernel and KD estimator are fixed across representations.

Our experiments take inspiration from methods used to validate evaluation metrics for image generation. The authors of FID (Heusel et al., 2017) validated their methods in ways resembling how photographs degrade ("perceptual" disturbances): sensor noise, defocus, occlusion, and dead pixels. Meanwhile, physical fields are characterized by the distribution of energy across spatial scales. The characteristic errors of numerical and learned simulators, such as over-smoothing and spurious small-scale energy, appear as distortions of this distribution. Section 4.1 therefore adopts FID's protocol of controlled disturbances at graded severity, adapting its distortions as perceptual families and adding spectral families that apply the protocol in Fourier space; these disturbances take fields off the manifold of reference solutions. By contrast, Section 4.2 tests an on-manifold difference, analogous to mode dropping in image generation (Xu et al., 2018): every state is a valid solution, but populations evolved from different initial conditions differ, and we measure how long they remain distinguishable. Section 4.3 scores trained neural simulators against an expected order of quality, and Appendix E.5 tests whether KD grows with the size of a difference, as FID does with disturbance level.

Comparison spaces. We compare PDE-KD with several alternative feature spaces. The learned frame representations are our physics-supervised ResNet-18 (PDE-KD), an identically structured ResNet-18 (He et al., 2016) trained on ImageNet, and mean-pooled Poseidon features (Herde et al., 2024). The Poseidon comparison is inspired by FPD (Cheng et al., 2025), but uses a different construction (Appendix C.1). Non-learned representations include a fixed 512-dimensional pixel subsample (Pixels), isotropic log-power spectra (Spectrum), and a whitened principal-component basis (PCA-white). To isolate the effect of training, we also score the untrained PDE-KD network (Random init); all its values are means over three random initializations (Appendix C.2). All representations are compared using KD.

For spatiotemporal populations, we compare our 3D feature map, PDE-KD (3D), with mean-pooled features from the JEPA architecture of Qu et al. (2026), which we retrain on Rayleigh–Bénard (RB) following the published training objective (Appendix C.4).

4.1 CONTROLLED DISTURBANCES: DISCRIMINATIVE POWER AND SAMPLE EFFICIENCY

We perturb post-transient RB fields (frames 61 onward) with four perceptual, eleven spectral, and seven temporal disturbance families at graded severity levels, numbered L1 through L5 in order of increasing severity (L5 examples in the top row of Table 2; Appendix E.1).

For each total IC count n between 8 and 175, we run 200 replicates, each drawing n ICs and randomly assigning them to disjoint clean and disturbed groups of approximately equal size. Each group pools frames, or clips for PDE-KD (3D) and JEPA, across its ICs and coefficient settings.

Table 2: **Discriminative power and sample efficiency on controlled disturbances.** Each cell shows n_{80} (lower better) | power at $n = 175$ (higher better; bold when at least .80). Sample efficiency uses L5 (the most severe disturbance); power uses L2. 8– and 175+ are censored at the sampled range. Top row: each disturbance at L5 on one RB buoyancy field (first column: the undisturbed reference).

	perceptual				spectral					temporal	
	Occlusion	Gaussian noise	Gaussian blur	Phase scrambling	Low-k inject	High-k inject	Low-k remove	High-k remove	Local swap (4–20)	Global swap (1–5, low)
										
PDE-KD	8– | **1.00**	8– | **1.00**	42 | **.89**	26 | .21	8– | **1.00**	8– | **1.00**	9 | **1.00**	41 | **1.00**	–	–
ImageNet	9 | **1.00**	8– | **1.00**	8– | **1.00**	12 | **1.00**	8– | .23	13 | **.94**	175+ | .08	21 | **1.00**	–	–
Poseidon	9 | .26	22 | **.82**	175+ | .07	175+ | .08	8– | **1.00**	8– | .08	8 | **1.00**	175+ | .11	–	–
Random init	22 | .10	175+ | .20	175+ | .12	175+ | .04	8– | **1.00**	8– | .68	33 | **1.00**	175+ | .13	–	–
Pixels	23 | .12	97 | .16	175+ | .07	61 | .09	10 | **1.00**	8– | .28	10 | **1.00**	175+ | .17	–	–
Spectrum	11 | **1.00**	8– | **1.00**	51 | **.98**	175+ | .07	175+ | .07	18 | .60	63 | .07	8– | **1.00**	–	–
PCA-white	8– | **1.00**	16 | **.96**	175+ | .06	11 | **.81**	8– | **1.00**	8– | **1.00**	8– | **1.00**	105 | .77	–	–
PDE-KD (3D)	8– | **1.00**	8 | **1.00**	69 | .28	8– | **1.00**	8– | **1.00**	8– | **1.00**	8– | **1.00**	33 | **1.00**	97 | .22	12 | **1.00**
JEPA	8– | .61	14 | **1.00**	175+ | .07	10 | **1.00**	8– | **1.00**	8– | **1.00**	9 | **1.00**	34 | **1.00**	175+ | .06	175+ | .10

Table 3: **Rejection coverage across physical systems; higher is better.** Each entry counts the disturbance *families* for which a single KD test rejects the null hypothesis. We test perceptual disturbances (at L2; 4 families) and spectral disturbances (at L5; 11 families) on five physical systems.

	RB		Shear flow		Euler-periodic		Gray–Scott		HRRR	
representation	perc.	spec.	perc.	spec.	perc.	spec.	perc.	spec.	perc.	spec.
PDE-KD	**4**	**11**	**4**	9	**4**	10	3	**11**	**4**	**11**
ImageNet	**4**	8	**4**	7	**4**	9	**4**	9	**4**	9
Poseidon	3	9	1	4	2	4	0	1	2	8
Random init (average of 3 seeds)	3.0	6.0	1.0	5.0	1.0	2.0	**4.0**	8.0	**4.0**	9.0
Pixels	2	7	0	4	1	0	0	0	1	8
Spectrum	**4**	6	**4**	6	**4**	7	3	5	**4**	8
PCA-white	3	9	0	7	1	3	0	3	1	8

We compute KD as in Section 3.2 and use a one-sided permutation test with 999 IC-label permutations at $\alpha = 0.05$. Power is the fraction of replicates that reject.

Table 2 reports discriminative power at L2 with $n = 175$ (87 versus 88 ICs) and n_{80}, the estimated total IC count needed for 80% power at L5. We estimate n_{80} by fitting power against $\log n$ with a logistic curve and censor crossings outside the sampled range (Appendix E.1).

Spectral disturbances distinguish representations clearly (Table 2): PDE-KD is the only frame representation with at least 80% L2 power on every reported band injection and removal, and the only one that reaches 80% power at L5 with at most 42 ICs on all eight disturbances. With all available ICs, it reaches 80% power by L5 on 10 of 11 spectral families, compared with 9 for PCA-white, 7 for ImageNet and Pixels, and 5 for Poseidon (Appendix E). For temporal disturbances, PDE-KD (3D) reaches 80% power on both frame-order swap schedules in Table 2; JEPA does not within the sampled range. We conclude that PDE-KD has high discriminative power across a range of spectral and perceptual disturbances, most notably with broad spectral coverage.

Generalization. We assess the effect of off-manifold disturbances on other physical systems: four systems from the training families of PDE-KD (including Rayleigh–Bénard), and HRRR (Dowell et al., 2022), a surface weather dataset not seen during PDE-KD's training. On every system, a single full-pool test with PDE-KD rejects at least three of the four perceptual and nine of the eleven spectral families, the most spectral families of any feature space (Table 3; protocol in Appendix E.4). See Appendix I.2 for more evaluations on out-of-distribution datasets.

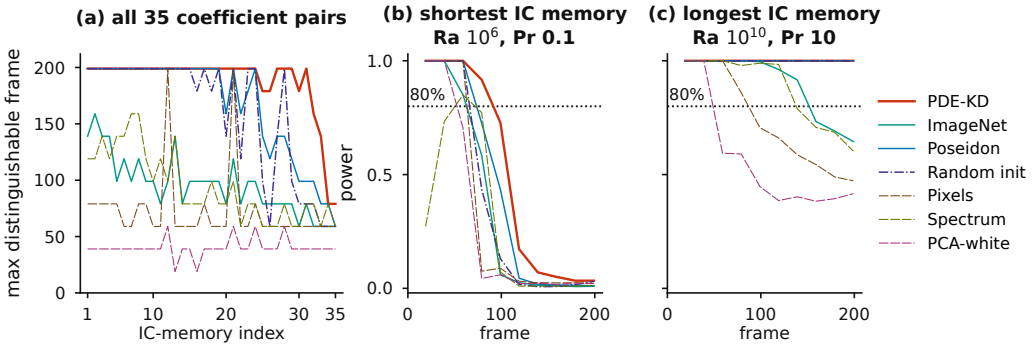


Figure 3: **Distinguishing initial-condition populations in RB. (a)** Max distinguishable frame of each coefficient pair, against its IC-memory index (1: longest memory). **(b, c)** Power over time at the pairs with the shortest and longest IC memory.

4.2 Distinguishing Initial-Condition Populations

Although chaos tends to erase memory about initial conditions, at intermediate time scales, the distribution of PDE solutions depends on how the system is initialized. We therefore assess the power to distinguish on-manifold differences that arise from different initial conditions in RB. We hold the dynamics and coefficients fixed and compare $\mu_t^A = (S_t)_\# \nu_A$ with $\mu_t^B = (S_t)_\# \nu_B$ (Section 2.1), as a function of lead time.

In the RB data of The Well (Ohana et al., 2024), each initial state consists of a linear temperature profile, ΔT_{init}, plus Gaussian noise of standard deviation 10^{-3}, damped toward the walls. The dataset provides five temperature profiles, each with ten noise seeds, at each of 35 $(\mathrm{Ra}, \mathrm{Pr})$ coefficient pairs; most trajectories come from the training split of PDE-KD (Appendix F.1). The Rayleigh number Ra measures the strength of buoyant forcing relative to viscous and thermal diffusion, and the Prandtl number Pr is the ratio of kinematic viscosity to thermal diffusivity.

For each coefficient pair, pair of initial temperature profiles $T_{\text{init}}^A, T_{\text{init}}^B$, and 20-frame window W, we test the null hypothesis $H_0 \colon \mu_W^A = \mu_W^B$ (Section 2.1). We view two initializations as producing the same distribution if they only differ by a noise seed. We compare five trajectories per temperature profile, with disjoint noise seeds, using the permutation test of Section 4.1, and average the power over the ten pairs of temperature profiles. The *max distinguishable frame* is the last frame before this power falls below 80%. How long a physical system "remembers" its initial condition depends on its coefficients; we therefore rank the coefficient pairs by an *IC-memory index* using a data-driven approach (Appendix F.1), which runs from 1, where the initial condition is easiest to recover, to 35.

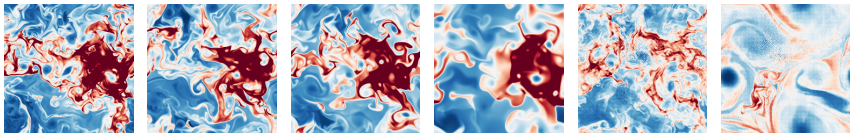
In the first window, which contains the initial state, every representation except Spectrum separates the initial conditions with at least 80% power at every coefficient pair (Figure 3). How long the separation lasts then depends strongly on the coefficients and generally shortens as the IC-memory index grows (panel a). PDE-KD keeps 80% power to the final frame at 28 pairs, against 21 for Poseidon and for Random init.

Averaged over all pairs, the final power is 0.90 for PDE-KD, 0.79 for Random init, and 0.73 for Poseidon. We conclude that, in RB, PDE-KD retains initial-condition differences for at least as long as every other tested feature space.

4.3 Scoring Trained Neural Simulators

We assess how PDE-KD ranks real-world neural simulators, and compare this ranking with the differences in rollout quality measured by conventional metrics and implied by the model design. We score two families of neural simulators whose quality has an expected order on inviscid Euler flow (RB: Appendix G.5). The Lola latent diffusion models (Rozet et al., 2025) differ only in the number of latent channels (64, 16 and 4). Distribution-level physical diagnostics support ordering them by compression: total entropy decreases, violating the second law, in 1.2%, 14.4% and 57.6%

Table 4: **Trained neural simulators against their expected order of quality** (Euler, periodic boundaries; one snapshot per model). Each row is rescaled to $[0, 1]$ (raw values: Table 50); parentheses: KD within two s.d. of the reference-versus-reference floor; green and red: best and worst in the row. "Expected order": within-family pairs ordered as expected (†ties excluded).

		Lola, best → worst			Walrus, best → worst		
	reference	c64	c16	c4	FT	PT	expected order
lead 88							
KD, learned features							
PDE-KD		(**0.00**)	0.03	0.25	0.03	1.00	4/4
ImageNet		**0.00**	0.06	0.61	0.29	1.00	4/4
Poseidon		(<0.01)	(**0.00**)	(<0.01)	(0.01)	(1.00)	3/4
KD, non-learned features							
Pixels		(0.02)	(**0.00**)	(0.03)	(0.27)	(1.00)	3/4
Spectrum		(**0.00**)	(<0.01)	0.06	1.00	0.82	3/4
error against the paired reference							
vRMSE		<0.01	**0.00**	0.05	0.44	1.00	3/4
CRPS		0.05	**0.00**	1.00	–	–	2/3
physical consensus (mean rank over distribution-level diagnostics)							
mean rank		**1.25**	2.00	3.25	4.25	4.25	3/3†

of trajectories from least to most compressed; the equation-of-state error and smoothing of local density grow likewise (Table 4; Appendix G.5). The Walrus model (McCabe et al., 2026) is either pretrained on a multi-physics mixture or fine-tuned on the target physics, and we expect fine-tuning to improve it.

For each metric, we count the within-family pairs ordered as expected (Table 4; Appendices G.5, G.3). On Euler, only KD in PDE-KD and ImageNet orders all four pairs. The snapshots show the degradation the expected order encodes: compression smooths the density field, and the pretrained Walrus loses the reference structure. On RB, KD in every learned feature space orders all four pairs, whereas vRMSE orders one and CRPS none, because both prefer the most compressed Lola model (Table 51). As Blau & Michaeli (2018) show, pixel-space metrics penalize a smoother field less for being out of phase with a decorrelated reference.

5 Discussion and Conclusion

We introduce PDE-KD, a kernel discrepancy between populations of simulated states, and supervise its feature space to identify PDE properties. PDE-KD detects a range of controlled off-manifold disturbances with high power and few trajectories, suggesting that its features retain information relevant to plausible model failures across physical systems within and outside its training families. It retains initial-condition differences in RB at least as long as every other tested feature space, supporting its use not only after decorrelation but also at intermediate time scales, where fidelity to the initial conditions matters. On trained neural simulators, PDE-KD orders the models of each family consistently with measures of rollout quality, according to violations of physical constraints and the loss of fine-scale structure, whereas pixel-space errors favor over-smoothed predictions (Section 4.3). This supports the use of PDE-KD as a general, physics-oriented evaluation metric for chaotic PDE simulations, complementing traditional evaluation metrics.

Future work. Beyond evaluation, the frozen, differentiable feature map could serve as a physics-aware training loss, as a batchwise KD objective complementing CRPS in probabilistic forecasting, or as an inference-time critic that flags rollout failures and ranks ensemble members; these use cases are not evaluated here.

AI USE STATEMENT

We used generative AI tools to assist code implementation, figure generation, manuscript editing, and consistency review. The authors checked AI-assisted code with automated tests, traced numerical claims to stored artifacts, and visually inspected generated figures and the compiled manuscript. Scientific decisions and interpretations were made and reviewed by the authors, who take responsibility for the final content.

ETHICS STATEMENT

This work studies evaluation methodology using simulated physical systems and publicly documented weather products. It involves no human subjects or personal data. As with any aggregate metric, misuse as a universal quality score could hide application-specific failures; we therefore report representation- and diagnostic-specific limitations and recommend using KD alongside targeted physical checks.

REPRODUCIBILITY STATEMENT

The problem formulation and estimator definitions are given in Sections 2–3, while Appendix E.1 records disturbance levels, sample support, hypothesis tests, matched lead grids, and comparator variants. Paper-facing tables and figures are produced by versioned generators with source hashes, commits, and executable environments recorded beside their outputs. The corresponding code, configurations, and generated provenance artifacts will accompany the supplementary material.

REFERENCES

Guenter Ahlers, Siegfried Grossmann, and Detlef Lohse. Heat transfer and large scale dynamics in turbulent Rayleigh–Bénard convection. *Reviews of Modern Physics*, 81(2):503–537, 2009. doi: 10.1103/RevModPhys.81.503.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970):533–538, 2023. doi: 10.1038/s41586-023-06185-3.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, 2018.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6228–6237, 2018. doi: 10.1109/CVPR.2018.00652.

Keaton J. Burns, Geoffrey M. Vasil, Jeffrey S. Oishi, Daniel Lecoanet, and Benjamin P. Brown. Dedalus: A flexible framework for numerical simulations with spectral methods. *Physical Review Research*, 2(2):023068, 2020. doi: 10.1103/PhysRevResearch.2.023068.

João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the Kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4724–4733, 2017. doi: 10.1109/CVPR.2017.502.

Lei Chen, Xiaohui Zhong, Feng Zhang, Yuan Cheng, Yinghui Xu, Yuan Qi, and Hao Li. FuXi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, 6(1):190, 2023. doi: 10.1038/s41612-023-00512-1.

Chaoran Cheng, Boran Han, Danielle C. Maddix, Abdul Fatir Ansari, Andrew Stuart, Michael W. Mahoney, and Yuyang Wang. Gradient-free generation for hard-constrained systems. In *International Conference on Learning Representations*, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Cristiana Diaconu, Miles Cranmer, Richard E. Turner, Tanya Marwah, and Payel Mukhopadhyay. Probabilistic retrofitting of learned simulators. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

David C. Dowell, Curtis R. Alexander, Eric P. James, Stephen S. Weygandt, Stanley G. Benjamin, Geoffrey S. Manikin, Benjamin T. Blake, John M. Brown, Joseph B. Olson, Ming Hu, Tatiana G. Smirnova, Terra Ladwig, Jaymes S. Kenyon, Ravan Ahmadov, David D. Turner, Jeffrey D. Duda, and Trevor I. Alcott. The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part I: Motivation and system description. *Weather and Forecasting*, 37(8):1371–1395, 2022. doi: 10.1175/WAF-D-21-0151.1.

Christopher A. T. Ferro. Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1917–1923, 2014. doi: 10.1002/qj.2270.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Maximilian Herde, Bogdan Raonić, Tobias Rohner, Roger Käppeli, Roberto Molinaro, Emmanuel de Bézenac, and Siddhartha Mishra. Poseidon: Efficient foundation models for PDEs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 72525–72624, 2024. doi: 10.52202/079017-2311.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: 10.1002/qj.3803.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6626–6637, 2017.

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*, 2026. URL https://openreview.net/forum?id=y3oHMcoItR.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456, 2015.

Aly-Khan Kassam and Lloyd N. Trefethen. Fourth-order time-stepping for stiff PDEs. *SIAM Journal on Scientific Computing*, 26(4):1214–1233, 2005. doi: 10.1137/S1064827502410633.

Georg Kohl, Li-Wei Chen, and Nils Thuerey. Benchmarking autoregressive conditional diffusion models for turbulent flow simulation. *Neural Networks*, 199:108641, 2026. doi: 10.1016/j.neunet.2026.108641.

Jean Kossaifi, Nikola Kovachki, Kamyar Azizzadenesheli, and Anima Anandkumar. Multi-grid tensorized Fourier neural operator for high-resolution PDEs. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=AWiDlO63bH.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, Alexander Merose, Stephan Hoyer, George Holland, Oriol Vinyals, Jacklynn Stott, Alexander Pritzel, Shakir Mohamed, and Peter Battaglia. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023. doi: 10.1126/science.adi2336.

Simon Lang, Mihai Alexe, Mariana C. A. Clare, Christopher Roberts, Rilwan Adewoyin, Zied Ben Bouallègue, Matthew Chantry, Jesper Dramsch, Peter D. Dueben, Sara Hahner, Pedro Maciel, Ana Prieto-Nemesio, Cathal O'Brien, Florian Pinault, Jan Polster, Baudouin Raoult, Steffen Tietsche, and Martin Leutbecher. AIFS-CRPS: Ensemble forecasting using a model trained with a loss function based on the continuous ranked probability score. *npj Artificial Intelligence*, 2(1): 18, 2026. doi: 10.1038/s44387-026-00073-7.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11966–11976, 2022. doi: 10.1109/CVPR52688.2022.01167.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Michael McCabe, Bruno Régaldo-Saint Blancard, Liam Holden Parker, Ruben Ohana, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Siavash Golkar, Géraud Krawezik, François Lanusse, Mariel Pettee, Tiberiu Tesileanu, Kyunghyun Cho, and Shirley Ho. Multiple physics pretraining for spatiotemporal surrogate models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 119301–119335, 2024. doi: 10.52202/079017-3791.

Michael McCabe, Payel Mukhopadhyay, Tanya Marwah, Bruno Régaldo-Saint Blancard, François Rozet, Cristiana Diaconu, Lucas Meyer, Kaze W. K. Wong, Hadi Sotoudeh, Alberto Bietti, Irina Espejo, Rio Fear, Siavash Golkar, Tom Hehir, Keiya Hirashima, Géraud Krawezik, François Lanusse, Rudy Morel, Ruben Ohana, Liam Parker, Mariel Pettee, Jeff Shen, Kyunghyun Cho, Miles Cranmer, and Shirley Ho. Walrus: A cross-domain foundation model for continuum dynamics. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026.

Rudy Morel, Jiequn Han, and Edouard Oyallon. DISCO: Learning to DISCover an evolution Operator for multi-physics-agnostic prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 44750–44774, 2025.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity metrics for generative models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7176–7185, 2020.

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 (Datasets and Benchmarks Track)*, volume 37, pp. 44989–45037, 2024. doi: 10.52202/079017-1430.

Belinda Phipson and Gordon K. Smyth. Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1):Article 39, 2010. doi: 10.2202/1544-6115.1585.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R. Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, Remi Lam, and Matthew Willson. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044): 84–90, 2025. doi: 10.1038/s41586-024-08252-9.

Helen Qu, Rudy Morel, Michael McCabe, Alberto Bietti, François Lanusse, Shirley Ho, and Yann LeCun. Representation learning for spatiotemporal physical systems. In *ICLR 2026 Workshop on AI & PDE*, 2026. URL https://arxiv.org/abs/2603.13227.

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 Bouallègue, 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.

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

François Rozet, Ruben Ohana, Michael McCabe, Gilles Louppe, François Lanusse, and Shirley Ho. Lost in latent space: An empirical study of latent diffusion models for physics emulation. In *Advances in Neural Information Processing Systems*, volume 38, pp. 149258–149303, 2025. doi: 10.52202/085713-4492.

Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics*, 41(5): 2263–2291, 2013. doi: 10.1214/13-AOS1140.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826, 2016. doi: 10.1109/CVPR.2016.308.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013. doi: 10.1016/j.jspi.2013.03.018.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Dan MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An extensive benchmark for scientific machine learning. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, volume 35, pp. 1596–1611, 2022. doi: 10.52202/068431-0117.

Ronak Tali, Ali Rabeh, Cheng-Hau Yang, Mehdi Shadkhah, Samundra Karki, Abhisek Upadhyaya, Suriya Dhakshinamoorthy, Marjan Saadati, Soumik Sarkar, Adarsh Krishnamurthy, Chinmay Hegde, Aditya Balu, and Baskar Ganapathysubramanian. FlowBench: A large scale benchmark for flow simulation over complex geometries, 2024. URL https://arxiv.org/abs/2409.18032.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations*, 2016.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, volume 35, pp. 10078–10093, 2022. doi: 10.52202/068431-0732.

Eleuterio F. Toro. *Riemann Solvers and Numerical Methods for Fluid Dynamics: A Practical Introduction*. Springer, Berlin, Heidelberg, third edition, 2009. ISBN 978-3-540-25202-3. doi: 10.1007/b79761.

Qiantong Xu, Gao Huang, Yang Yuan, Chuan Guo, Yu Sun, Felix Wu, and Kilian Q. Weinberger. An empirical study on evaluation metrics of generative adversarial networks, 2018. URL https://arxiv.org/abs/1806.07755.

Appendix

Organization. Appendices A–D document the evidence base: the protocol of Figure 1 and its repetition on three more systems, the training corpus, feature map and evaluation data, the comparison feature spaces and metrics, and the statistical methodology. Appendices E–G give the protocols and full results of Section 4 for controlled perturbations, IC-conditional laws, and physical diagnostics and model ladders. Appendices H–I cover probes of the feature map and out-of-distribution experiments, and Appendix J gives qualitative examples.

Three conventions hold throughout. (i) KD values are *scores*: unbiased estimates that can be negative, never compared in raw magnitude across feature spaces (Appendix C.7). (ii) Unless stated otherwise, "PDE-KD" denotes KD in the 512-dimensional feature space of the feature map of Appendix B. (iii) Disturbance suites on reference data are scored at native frames 61 onward, and RB rollouts at leads 61 onward (native frames 77–199); Euler rollouts use the shared stride-4 leads 64–88. Departures from these windows are stated where they occur.

A Figure 1 across physics 16
A.1 Protocol of Figure 1 (p. 16); A.2 Three more physical systems (p. 16)
B Datasets and feature map 22
B.1 Training corpus (p. 22); B.2 Channel dictionary and preprocessing (p. 23); B.3 Objective and training budget (p. 24); B.4 Supervised performance (p. 25); B.5 Feature-map design choices and ablations (p. 27); B.6 Evaluation data (p. 27)
C Comparison feature spaces and metrics 29
C.1 Poseidon features and Fréchet Poseidon Distance (p. 29); C.2 Untrained-network control (Random init) (p. 30); C.3 The clip feature map, PDE-KD (3D) (p. 32); C.4 The JEPA comparison encoder (p. 32); C.5 Non-learned spaces (p. 32); C.6 Pointwise, ensemble and spectral metrics (p. 33); C.7 Comparing spaces (p. 33)
D Statistical methodology 33
D.1 The KD estimator (p. 33); D.2 Unit of independence (p. 34); D.3 Sample-count response (p. 35); D.4 Error bars (p. 35); D.5 Pooling over lead time and regime (p. 35); D.6 Hypothesis tests, calibration and exact floors (p. 36); D.7 Robustness of rejection verdicts (p. 40); D.8 Which estimator and test each result uses (p. 41)
E Controlled perturbations: definitions and full results 42
E.1 Protocol (p. 42); E.2 Definitions and severity ladders (p. 43); E.3 Full results on Rayleigh–Bénard (p. 45); E.4 Other physical systems (p. 46); E.5 Monotonicity in severity (p. 49); E.6 Temporal suite (p. 50); E.7 Specificity: symmetries of the governing equations (p. 52); E.8 Rollout resilience to corrupted conditioning frames (p. 53); E.9 Raw scores (p. 53)
F IC-conditional laws: full results 60
F.1 The permutation test of Section 4.2 (p. 60); F.2 The single-trajectory exceedance probe (p. 61); F.3 Unseen physics: the turbulent radiative layer (p. 63); F.4 Structure of the RB reference (p. 63)
G Physical diagnostics and model ladders 64
G.1 Model roster (p. 64); G.2 Physical observables (p. 65); G.3 Agreement with physical diagnostics (p. 66); G.4 Per-functional correlations (p. 68); G.5 Scoring trained neural simulators: details (p. 70); G.6 Euler compression ladder (p. 72); G.7 Paired inversions and CRPS against climatology (p. 73)
H Probes of the feature map 75
I Out-of-distribution generalization 76
I.1 Reference-versus-reference axes and their triviality (p. 76); I.2 Controlled disturbances out of distribution (p. 78); I.3 Sim-to-real gap (p. 79); I.4 Ranking weather forecasts (p. 82); I.5 Frozen features as regression inputs (p. 83); I.6 A held-out-family feature map (p. 83)
J Qualitative examples 83

A FIGURE 1 ACROSS PHYSICS

A.1 PROTOCOL OF FIGURE 1

Each row of Figure 1 compares one candidate trajectory with one reference RB trajectory ($\mathrm{Ra} = 10^9$, $\mathrm{Pr} = 0.5$, $\Delta T_{\mathrm{init}} = 0.4$, perturbation seed 40; test split) over native frames 61–199 (139 frames). The null row changes only the perturbation seed (40 to 44). The coefficient row changes only Ra, to 10^8; it is a training-split trajectory with the same Pr, ΔT_{init} and seed. The symmetry row applies a legal RB symmetry to the reference's own frames: a reflection of the periodic horizontal axis plus a periodic shift of 184 of its 512 cells. This transformation is draw L2 of the augmentation sampler of our training pipeline, which the released feature map was *not* trained with; we fixed it before scoring as the lowest-indexed draw that exercises both operations. The blur row applies a Gaussian blur of $\sigma = 3$ native pixels (the L3 parameter, applied on the native grid) to the reference's own frames.

With one trajectory per arm, the IC-level estimator is undefined, so the PDE-KD column uses the frame-level U-statistic. Frames of one trajectory are autocorrelated, so the comparison is descriptive, not a test. The energy-distance column is the energy distance between the two frame populations: a frame-level U-statistic of the per-pixel RMS distance between native z-scored frames. As a finite-sample U-statistic, it can be slightly negative. The stored column is half this energy distance; the column scaling below cancels the factor. With autocorrelated frames, the frame-level statistic is not centered for these designs. It tends upward for independent trajectories (null and coefficient rows) and downward for rows built from the reference's own frames (symmetry and blur rows), so small values in either direction carry no verdict. The vRMSE column (the per-frame variance-normalized RMSE of Appendix C.6) and the spectral error pair equal-index frames. Each metric column is divided by its largest value over the four scored rows, so its most severe response reads 1.00 (two decimals). Cell shading is a within-column square-root ramp over the raw values and carries no statistical verdict.

A.2 THREE MORE PHYSICAL SYSTEMS

This subsection repeats the design of Figure 1 (Appendix A.1) on three more physical systems, each regenerated with a controllable seed and scored with the same functions. Shear flow uses the Well's Dedalus generator (Burns et al., 2020) with a seeded, divergence-free perturbation. Gray–Scott uses the Well's Fourier/ETDRK4 scheme (Kassam & Trefethen, 2005) from seeded Gaussian-cluster initial conditions. Compressible Kelvin–Helmholtz uses a GPU finite-volume solver that approximates Poseidon's CE-KH (Herde et al., 2024). We infer the time span of Poseidon's 21 frames by matching: with the released parameters, our dynamics come closest to Poseidon's when these frames span $t \in [0, 8]$ (KD 0–1.5 against released trajectories, -0.6 to 0.3 between our own seeds). The shear-flow initial condition departs from the Well's. Under the Well's own initial condition plus a 10^{-3} seeded perturbation (design A), seed partners do not decorrelate (Figure 6c). The design used here (design B) therefore drops the Well's deterministic blob seeding of the vertical velocity and triggers the Kelvin–Helmholtz instability with a 10^{-2} seeded perturbation. Design-B frames lie somewhat off the feature map's training distribution, equally for every row. Figures 4–8 report the energy distance of Figure 1 without its column scaling. Symmetry rows are exact symmetries of both the equations and the seeded initial-condition law.

Verdict questions. Every case is repeated over several seeds per law: 10 for RB, 6 for Gray–Scott and Kelvin–Helmholtz, and for shear flow 4 at the reference law and 2 at the other. The seed pairs at the reference law are the *same-law pairs*; their spread gives a range check, not a test. Each score is asked three questions against them: (1) does *every* cross-law pair score above *every* same-law pair; (2) does the blur score above every same-law pair; (3) does the symmetry copy score at or below the largest same-law pair? No permutation test was run. On shear flow none could reject at $\alpha = 0.05$: 4 and 2 trajectories admit only 15 labelings, so the smallest attainable p is $1/15$.

Results. PDE-KD answers all three questions correctly in all nine reported cases, and it is the only score that does. The energy distance and vRMSE never separate a coefficient change from a fresh seed or place the blur above every seed pair. vRMSE ranks the blur closest in all nine cases, and the energy distance does so in all but the two Kelvin–Helmholtz cases. The spectral error answers all

	PDE-KD	ImageNet	Poseidon	Random init	Energy distance	vRMSE	Spectral error
RB, Ra $10^9 \to 10^8$	✓✓✓	✓✓✓	✗✗✓	✗✓✓	✗✗✓	✗✗✓	✓✓✓
Shear flow, Re $10^5 \to 10^4$	✓✓✓	✓✓✓	✗✗✗	✓✗✓	✗✗✓	✗✗✓	✗✓✓
Shear flow, Re $10^5 \to 5 \times 10^4$	✓✓✓	✗✓✓	✗✗✗	✗✗✓	✗✗✓	✗✗✓	✗✓✓
Gray–Scott maze, F +10%	✓✓✓	✗✓✓	✓✓✓	✓✓✓†	✗✗✓	✗✗✓	✓✓✓
Gray–Scott spirals, F +10%	✓✓✓	✗✓✓	✓✗✓	✓✓✓	✗✗✓	✗✗✓	✗✓✓
Gray–Scott spots, F +10%	✓✓✓	✓✓✗	✓✓✓	✓✓✗†	✗✗✓	✗✗✓	✓✓✓
Gray–Scott gliders, F +10%	✓✓✓	✓✓✓	✓✓✓	✓✓✓	✗✗✓	✗✗✓	✗✓✓
Compressible KH, ΔU +20%	✓✓✓	✗✓✓	✓✗✓	✓✓✗†	✗✗✓	✗✗✓	✓✓✓
Compressible KH, ΔU +10%	✓✓✓	✗✓✓	✗✗✓	✓✓✗†	✗✗✓	✗✗✓	✗✓✓

Figure 4: **Verdicts across nine cases.** Each cell answers questions (1)–(3) of the text in order: coefficient change separated from every seed pair, blur above every seed pair, symmetry copy within the seed range; a check mark is a correct answer. Shading runs from green (all three correct) to red (none correct). Random init: the untrained PDE-KD network, mean over three initialization seeds of each pair's KD (Appendix C.2); †: verdict differs from the majority of the seeds. RB: Rayleigh–Bénard; KH: Kelvin–Helmholtz.

three questions on RB, the Gray–Scott maze and spots, and Kelvin–Helmholtz at $\Delta U + 20\%$; in the other five cases it misses the coefficient change. ImageNet and Poseidon features each answer all three questions in three cases (Figures 4 and 5).

Per-case figures. Figures 6–8 show every case, one pair per row; they are descriptive. On shear flow at $\mathrm{Re} = 10^5 \to 5 \times 10^4$, the coefficient change at the same seed barely moves the flow pointwise (vRMSE 0.059, against 0.81 for a fresh seed). vRMSE, the energy distance and the spectral error therefore all rank it closer to the reference than the fresh seed; only the blur (vRMSE, energy distance) or the symmetry copy (spectral error) ranks closer. In Kelvin–Helmholtz, frame 0 is the initial condition, which already carries the change of ΔU in its velocity, so the coefficient row differs from the reference from the first frame. The no-decorrelation control (Figure 6c) scores frames 61–105 of runs stopped at $t \approx 11$, with no seed-pair spread. Its seed partner stays close to the reference (tracer relative $L^2 \approx 0.02$), so every score, the pointwise ones included, places it below both the coefficient change and the blur. The failure of the pointwise metrics in Figure 1 therefore requires decorrelation. This control's PDE-KD and pointwise values were recomputed and reproduce the original run. Its ImageNet and Poseidon values come from that run alone, because the run's inputs are only partly preserved.

Case selection and pilot cases. No pre-registered rule fixed these nine cases. They were chosen after pilot scans that also generated six further cases. On Gray–Scott these are $F - 10\%$ in the gliders, spirals and spots regimes (three seeds each); the worms and bubbles regimes were scanned too and dropped, because our generator does not reproduce them. On Kelvin–Helmholtz they are $\Delta U - 10\%$, an adiabatic index $\gamma = 1.3$ and a 50% larger interface perturbation (one seed each). We scored the six pilot cases afterwards with the pipeline, questions and reference law of the reported cases; scores computed during the scans themselves were not recorded. With three coefficient seeds (Gray–Scott) or one (Kelvin–Helmholtz), the coefficient question compares 18 or 6 cross-law pairs rather than 36, which makes it easier to pass. We therefore compare each pilot case with its reported sibling case restricted to the same coefficient seeds. On the spirals, spots and $\Delta U - 10\%$ cases, every column returns the verdicts of the corresponding reported case. PDE-KD answers all three questions on these three cases and on $\gamma = 1.3$, and on all four its answers match those of the restricted reported case. On spots, two of the three coefficient seeds decay to a uniform state, and the verdicts are unchanged on the surviving seed. The $\gamma = 1.3$ change is large pointwise: it shifts the stored energy $p/(\gamma - 1)$ pointwise from the first frame, so vRMSE is at least 16 across the law change, against at most 1.08 between seeds. On gliders, all three coefficient seeds decay to

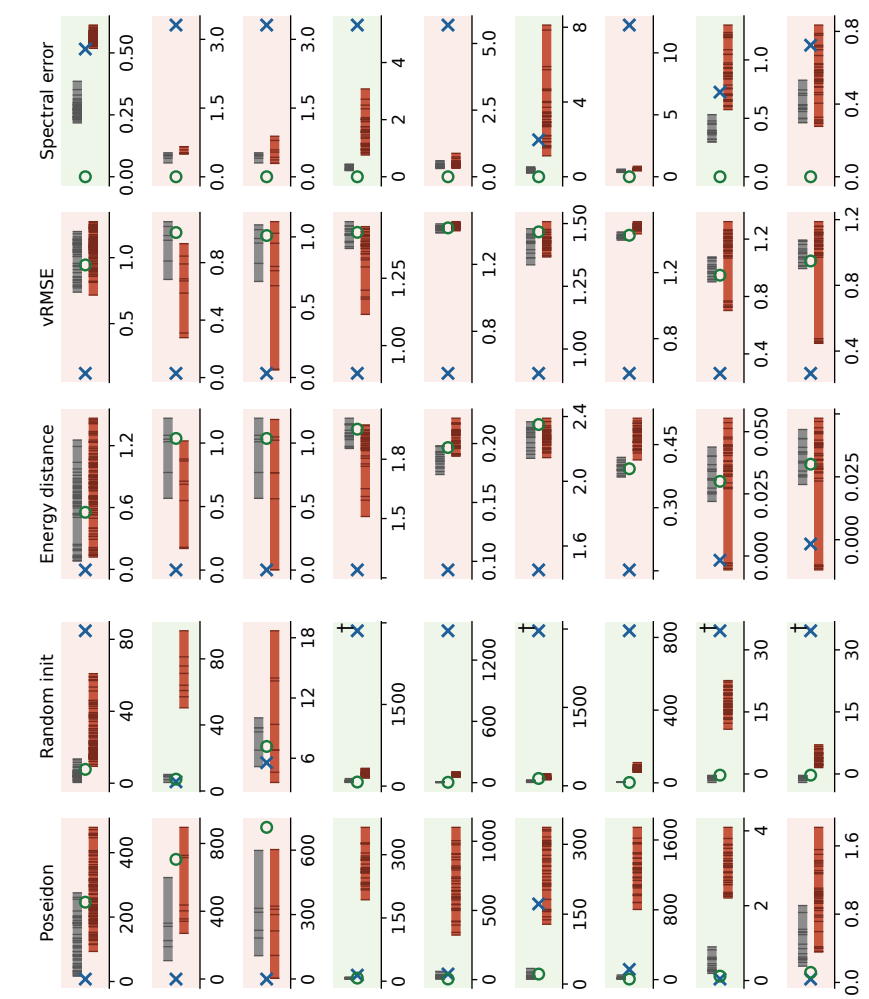


Figure 5: **The ranges behind Figure 4.** Each panel shows one score on its own horizontal axis; KD columns show KD $\times 10^3$. Gray: same-law seed pairs; red: cross-law pairs; $\times$: blur; $\circ$: symmetry copy. A green panel means every cross-law pair scores above every same-law pair. Random init and † (panel corner) as in Figure 4.

	Same system?	PDE-KD	ImageNet	Poseidon	Random init	Energy distance	vRMSE	Spectral error
Reference								
Symmetry-augmented	✓	0.77	106	706	7.17	1.036	1.010	0.001
Different noise seed	✓	1.09	62.17	311	9.03	0.768	0.811	0.494
Different coefficient	✗	21.64	1665	341	64.05	0.179	0.310	0.500
Blurred	✗	5.12	653	-0.33	5.56	0.001	0.030	3.312

(a) $\mathrm{Re} = 10^5 \to 10^4$ (one decade), frames 61–199.

	Same system?	PDE-KD	ImageNet	Poseidon	Random init	Energy distance	vRMSE	Spectral error
Reference								
Symmetry-augmented	✓	0.77	106	706	7.17	1.036	1.010	0.001
Different noise seed	✓	1.09	62.17	311	9.03	0.768	0.811	0.494
Different coefficient	✗	1.52	162	3.43	3.59	0.002	0.059	0.292
Blurred	✗	5.12	653	-0.33	5.56	0.001	0.030	3.312

(b) $\mathrm{Re} = 10^5 \to 5 \times 10^4$ (a factor of two, one step of the Re grid), frames 61–199.

	Same system?	PDE-KD	ImageNet	Poseidon	Random init	Energy distance	vRMSE	Spectral error
Reference								
Symmetry-augmented	✓	-0.25	-7.44	-0.61	-0.38	-0.009	0.019	0.000
Different noise seed	✓	-0.26	-7.45	-0.62	-0.40	-0.010	0.011	0.036
Different coefficient	✗	28.18	2280	87.51	58.24	0.108	0.179	0.481
Blurred	✗	8.84	792	-0.53	13.10	0.005	0.036	7.491

(c) No-decorrelation control (design A): the Well's initial condition plus a 10^{-3} perturbation, $\mathrm{Re} = 10^5 \to 10^4$, frames 61–105.

Figure 6: **Shear flow.** One pair per row. Left block: KD $\times 10^3$ in four feature spaces (Random init: mean over three initialization seeds); right block: the energy distance, vRMSE and spectral error of Figure 1. Values are not normalized per column, unlike those of Figure 1; shading is per column, from green (smallest) to red (largest).

	Same system?	PDE-KD	ImageNet	Poseidon	Random init	Energy distance	vRMSE	Spectral error
Reference								
Symmetry-augmented	✓	24.24	101	7.32	74.37	1.952	1.420	0.000
Different noise seed	✓	13.55	185	6.68	81.54	1.973	1.432	0.405
Different coefficient	✗	636	376	258	169	1.612	1.180	0.912
Blurred	✗	4966	3313	14.37	2854	1.240	0.897	5.312

(a) Maze.

	Same system?	PDE-KD	ImageNet	Poseidon	Random init	Energy distance	vRMSE	Spectral error
Reference								
Symmetry-augmented	✓	1.50	6.21	4.57	3.46	0.197	1.418	0.000
Different noise seed	✓	10.26	5.23	34.83	3.97	0.185	1.429	0.490
Different coefficient	✗	273	21.08	322	58.21	0.197	1.425	0.503
Blurred	✗	2983	1410	40.87	1487	0.093	0.549	5.691

(b) Spirals.

	Same system?	PDE-KD	ImageNet	Poseidon	Random init	Energy distance	vRMSE	Spectral error
Reference								
Symmetry-augmented	✓	26.07	62.28	21.41	140	2.351	1.467	0.000
Different noise seed	✓	14.70	16.45	14.83	82.99	2.292	1.430	0.302
Different coefficient	✗	479	324	244	221	2.268	1.418	3.423
Blurred	✗	2978	1706	172	2967	1.452	0.903	1.976

(c) Spots.

	Same system?	PDE-KD	ImageNet	Poseidon	Random init	Energy distance	vRMSE	Spectral error
Reference								
Symmetry-augmented	✓	0.74	1.54	6.49	2.05	0.393	1.426	0.000
Different noise seed	✓	3.23	3.13	10.23	2.85	0.376	1.414	0.590
Different coefficient	✗	164	31.29	1320	88.48	0.437	1.462	0.654
Blurred	✗	4176	1041	117	837	0.152	0.590	12.242

(d) Gliders.

Figure 7: **Gray–Scott**, F +10% at fixed k, frames 862–1000. Format as Figure 6.

Reference	Same system?	PDE-KD	ImageNet	Poseidon	Random init	Energy distance	vRMSE	Spectral error
Symmetry-augmented	✓	-0.48	-20.59	0.12	-0.29	0.030	0.948	0.000
Different noise seed	✓	-0.46	-6.88	0.62	-0.88	0.025	0.939	0.319
Different coefficient	✗	4.14	-11.53	2.32	14.58	-0.003	0.769	0.655
Blurred	✗	31.59	1125	0.04	34.61	-0.002	0.265	0.724

(a) ΔU +20%.

Reference	Same system?	PDE-KD	ImageNet	Poseidon	Random init	Energy distance	vRMSE	Spectral error
Symmetry-augmented	✓	-0.48	-20.59	0.12	-0.29	0.030	0.948	0.000
Different noise seed	✓	-0.46	-6.88	0.62	-0.88	0.025	0.939	0.319
Different coefficient	✗	0.42	-22.42	0.36	2.02	-0.011	0.475	0.306
Blurred	✗	31.59	1125	0.04	34.61	-0.002	0.265	0.724

(b) ΔU +10%.

Reference	Same system?	PDE-KD	ImageNet	Poseidon	Random init
Symmetry-augmented	✓	1.84	6.22	246	7.70
Different noise seed	✓	0.91	9.58	275	7.03
Different coefficient	✗	6.95	1211	405	15.85
Blurred	✗	11.57	7533	7.64	84.75

(c) Rayleigh–Bénard, $\mathrm{Ra} = 10^9 \to 10^8$: the feature-space companion of Figure 1.

Figure 8: **Compressible Kelvin–Helmholtz and Rayleigh–Bénard.** Kelvin–Helmholtz: shear velocity ΔU, all 21 frames ($t \in [0, 8]$). Format as Figure 6; panel (c) shows the feature spaces only.

a uniform state, so that case tests pattern extinction rather than a coefficient change; every metric except vRMSE separates it. Restricting the reported gliders case to the same seeds flips the energy-distance and vRMSE answers to the coefficient question, but not PDE-KD's. No metric separates the 50% larger interface perturbation, which changes a parameter of the initial-condition law rather than of the equations. There, PDE-KD's cross-law pairs (-0.44 to 0.49, KD $\times 10^3$) overlap its same-law pairs (up to -0.31), whereas the reported ΔU cases pass on the same seed. A 25-point (F, k) map around the maze regime (one seed each) was also generated; two of its points are the reported maze reference and $F + 10\%$ runs, and the other 23 were not scored.

Untrained network. The Random init column shows what training adds. Judged on the seed mean, Random init separates the coefficient change in 7 of 9 cases: all four Gray–Scott patterns, both Kelvin–Helmholtz cases and the one-decade shear-flow case. It misses the coefficient change on RB and on the factor-of-two shear-flow case, and it misses the blur on both shear-flow cases. It answers all three questions on the Gray–Scott maze, spirals and gliders. By the majority of its three seeds, it separates the coefficient change in only 5 of 9 cases, because two of the three seeds also miss it on maze and spots. On Gray–Scott F and Kelvin–Helmholtz ΔU, which no coefficient head regresses, the seed mean, like PDE-KD, separates the coefficient change and places the blur above every seed pair in all six cases. Across these six cases, the seed mean misses only the symmetry copies of spots and of both Kelvin–Helmholtz cases, which PDE-KD places within the seed range. On both Kelvin–Helmholtz cases, two of the three untrained seeds individually place the symmetry copy within the seed range as well. Random init's BatchNorm statistics come from RB and are used unchanged here (Appendix C.2). Its kernel scale was not measured on these runs, so off RB its verdicts also test that calibration.

Caveats. Frames of one trajectory are autocorrelated. The shear-flow spreads are small (6 same-law and 8 cross-law pairs). Of RB's 20 spread trajectories, 17 are in the feature map's training split; all other systems use fresh simulations. Gray–Scott's (F, k) enters training only through its six named patterns, six widely separated (F, k) values that serve as regime-classification targets (Table 6); $F + 10\%$ is not among them. The $\sigma = 3$ blur is severe on the 128^2 Gray–Scott and Kelvin–Helmholtz grids.

B Datasets and feature map

The feature map is a single checkpoint (step in Table 8), exported to PyTorch with parity checks on features and logits (Appendix B.2). Every per-frame (2D) PDE-KD number in the paper uses this checkpoint. PDE-KD (3D) uses the clip feature map of Appendix C.3, and the probe and training-amount studies use sibling checkpoints, as stated where they appear.

B.1 Training corpus

The training corpus has nine classes (Table 5). Six are datasets of the Well (Ohana et al., 2024): shear flow, Rayleigh–Bénard (RB) convection, compressible Euler (multi-quadrant Riemann problems), active matter, viscoelastic instability and Gray–Scott reaction–diffusion. The Euler dataset is split by boundary condition into two classes (open and periodic). The remaining two classes are the compressible-Euler and incompressible Navier–Stokes operator families of Poseidon's PDEgym (Herde et al., 2024). Each class contributes either a regression target (the PDE coefficients it sweeps) or a categorical target (named regimes), never both. Shear flow, RB, Euler and active matter also carry the datasets' regime identifiers, which no head reads. Gray–Scott carries feed and kill labels, but they are masked out of the coefficient loss. Gray–Scott is therefore supervised only through its six named patterns.

The trajectory counts in Table 5 are those present in our training copy of each dataset when the feature map was trained. They differ from the nominal split sizes in the dataset specifications. The active-matter copy holds 225 of the Well's 360 trajectories (175 / 24 / 26 train / validation / test). The Poseidon operator families were truncated when converted to our training format. Both Poseidon families' validation splits contained only a subset of their regimes at training time (CE: 2 of 4; NS: 1 of 2), so their regime accuracies in Table 12 measure a reduced task (for NS, with one class present, a trivial one).

Table 5: **Training corpus of the feature map.** C: number of native channels. Train / Valid / Test: trajectories present when the feature map was trained. Every family also carries the 9-way family label. [a]In the validation split, viscoelastic trajectories have $T = 60$ (CAR, EIT) or $T = 20$ (the other 5 regimes). [b]210 at training time, because the $\mathrm{Ra} = 10^9$, $\mathrm{Pr} = 2$ test directory duplicated its 40-trajectory training directory; the decontaminated test split scored in this paper has 175. [c]Enlarged after training: Euler open-BC test to 500 trajectories; Poseidon validation and test doubled, adding the missing IC families. [d]Only CE-CRP, CE-RP and NS-Sines were present in validation and test at training time, and in the validation split scored in Table 12.

Family	$H \times W$	T	Fields (decode order)	C	Train	Valid	Test	Supervised targets
Shear flow	256×512	200	tracer, pressure, velocity	4	896	112	112	Re (4), Sc (7)
Rayleigh–Bénard	512×128	200	buoyancy, velocity, pressure	4	1,400	175	210[b]	Ra (5), Pr (7)
Euler, open BC	512×512	101	energy, momentum, pressure, density	5	4,000	500	0[c]	γ (10)
Euler, periodic BC	512×512	101	energy, momentum, pressure, density	5	4,000	500	500	γ (10)
Active matter	256×256	81	concentration, velocity, $\mathbf{E}$, $\mathbf{D}$	11	175	24	26	ζ (9), α (5)
Viscoelastic	512×512	20/60[a]	pressure, c_{zz}, velocity, $\mathbf{C}$	8	213	22	22	7 flow regimes
Poseidon CE	128×128	21	energy, momentum, pressure, density	5	32,808	240[c]	480[c]	4 IC families[d]
Poseidon NS	128×128	21	velocity, tracer	3	26,318	120[c]	240[c]	2 IC families[d]
Gray–Scott	128×128	1001	A, B	2	960	120	120	6 (F, k) patterns
Total					70,770	1,813	1,710	

Table 6: **Supervised targets.** Grid: the distinct coefficient values of the validation and test splits; $[0, 1]$ map: the normalization applied before the coefficient loss. Regime classes are in head order, with Poseidon training trajectories per class in parentheses; valid coverage: classes present in the validation split.

Family	Target	Grid / classes	$[0, 1]$ map / valid coverage	Size
Shear flow	Re	$10^4, 5{\times}10^4, 10^5, 5{\times}10^5$	$\log_{10}\mathrm{Re}/8$	4
Shear flow	Sc	0.1, 0.2, 0.5, 1, 2, 5, 10	$(\log_{10}\mathrm{Sc} + 2)/4$	7
Rayleigh–Bénard	Ra	$10^6, 10^7, 10^8, 10^9, 10^{10}$	$\log_{10}\mathrm{Ra}/12$	5
Rayleigh–Bénard	Pr	0.1, 0.2, 0.5, 1, 2, 5, 10	$(\log_{10}\mathrm{Pr} + 2)/4$	7
Euler, open BC	γ	1.13, 1.22, 1.3, 1.33, 1.365, 1.4, 1.404, 1.453, 1.597, 1.76	$\gamma - 1$	10
Euler, periodic BC	γ	1.13, 1.22, 1.3, 1.33, 1.365, 1.4, 1.404, 1.453, 1.597, 1.76	$\gamma - 1$	10
Active matter	ζ	1, 3, 5, 7, 9, 11, 13, 15, 17	$\zeta/20$	9
Active matter	α	$-5, -4, -3, -2, -1$	$-\alpha/5$	5
Viscoelastic	regime	AH, CAR, EIT, transtochaos_EIT_L, transtochaos_EIT_SAR, transtononchaos_EIT_L, transtononchaos_EIT_SAR	7/7 classes in valid	7
Poseidon CE	regime	CE-CRP (9,640), CE-Gauss (6,992), CE-KH (6,536), CE-RP (9,640)	2/4 classes in valid	4
Poseidon NS	regime	NS-Sines (19,640), NS-Gauss (6,678)	1/2 classes in valid	2
Gray–Scott	regime	Bubbles, Gliders, Maze, Spirals, Spots, Worms	6/6 classes in valid	6

B.2 Channel dictionary and preprocessing

This subsection describes how every frame is mapped onto the feature map's fixed input and how the trained network is exported.

Channel dictionary. Every frame is written into 13 canonical slots: 0 pressure; 1 primary scalar (energy, buoyancy, tracer, c_{zz}, concentration, or Gray–Scott A); 2 secondary scalar (density or Gray–Scott B); 3–4 a vector (velocity or momentum); 5–8 a first rank-2 tensor; 9–12 a second rank-2 tensor. Tensors are flattened row-major as (xx, xy, yx, yy), and unoccupied slots are exactly zero. Slots are capacity roles, not shared physical variables. Poseidon's compressible-Euler "momentum" field is physically a velocity: the ideal-gas relation holds with $\gamma = 1.4$ only under that reading. It nevertheless occupies the vector slots used by Euler momentum. Table 7 lists the occupancy.

Normalization and resizing. Each native channel is z-scored with one mean and standard deviation per (family, channel), estimated on the training split and collapsed over time by the law of total variance, $\sigma_c^2 = \overline{\sigma_{tc}^2} + \mathrm{Var}_t(\mu_{tc})$, with σ_c clamped at $\geq 10^{-4}$. There is no per-coefficient or per-sample normalization, so the coefficient-dependent amplitude that the coefficient head needs is kept. For viscoelastic instability the statistics cover only frames 0–19, although two of its regimes run to 60 frames (Table 5, note a). For it and for active matter, at most 128 trajectories per coefficient setting entered the statistics. The z-score is applied at native resolution. The frame is then bilinearly

Table 7: **Occupancy of the 13 canonical input slots.** ·: slot held at exactly 0; component index 0 is the H axis. Symbols: p pressure, e energy, b buoyancy, s passive tracer, c concentration, c_{zz} conformation zz-component, ρ density, A, B Gray–Scott species, u velocity, m momentum, E/D active-matter strain-rate / orientation tensors, and C conformation tensor.

	scalars			vector		tensor 0				tensor 1				
Family	0	1	2	3	4	5	6	7	8	9	10	11	12	#real
Shear flow	p	s	·	u_0	u_1	·	·	·	·	·	·	·	·	4
Rayleigh–Bénard	p	b	·	u_0	u_1	·	·	·	·	·	·	·	·	4
Euler, open BC	p	e	ρ	m_0	m_1	·	·	·	·	·	·	·	·	5
Euler, periodic BC	p	e	ρ	m_0	m_1	·	·	·	·	·	·	·	·	5
Active matter	·	c	·	u_0	u_1	E_{00}	E_{01}	E_{10}	E_{11}	D_{00}	D_{01}	D_{10}	D_{11}	11
Viscoelastic	p	c_{zz}	·	u_0	u_1	C_{00}	C_{01}	C_{10}	C_{11}	·	·	·	·	8
Poseidon CE	p	e	ρ	m_0	m_1	·	·	·	·	·	·	·	·	5
Poseidon NS	·	s	·	u_0	u_1	·	·	·	·	·	·	·	·	3
Gray–Scott	·	A	B	·	·	·	·	·	·	·	·	·	·	2

resized to 512×512 without preserving aspect ratio and scattered into the slots. RB (512×128) is thus stretched $4\times$ along its 128-point wall-normal axis, whose Chebyshev–Gauss grid is treated as uniform. Every native grid is at most 512 per axis, so the resize only upsamples, and the antialiasing filter of the TensorFlow resize has no effect. The training pipeline and the PyTorch evaluation port agree bit-exactly on the z-score and to 4.8×10^{-7} on the resize, measured on a fixture of RB validation frames. No data augmentation is used.

Export and parity. Features are extracted in fp32 with TF32 off and BatchNorm in inference mode using the stored exponential-moving-average statistics. With batch statistics instead, family accuracy collapses from 0.999 to 0.217 (Appendix B.5), so the stored statistics are part of the feature space. The PyTorch port reproduces the Flax feature map to a maximum absolute difference of 6.6×10^{-7} on the 512-dimensional features and 7.6×10^{-6} on the logits, on a fixture of 16 RB validation frames (both on CPU). The CUDA forward pass of the evaluation pipeline agrees with the CPU port to 3.8×10^{-6} (features) and 1.3×10^{-5} (logits). The weights are pinned by SHA-256 and loaded strictly.

B.3 Objective and training budget

A single linear layer on the 512-dimensional pooled feature has 46 outputs: 9 family logits; 18 coefficient outputs (two per family, used only where a family sweeps coefficients); and $7 + 4 + 2 + 6$ regime logits for viscoelastic instability, Poseidon CE, Poseidon NS and Gray–Scott. The true family label selects which coefficient or regime outputs enter the loss. The loss is Eq. (9) with $\alpha = 2$, $\beta = 0.5$. $\mathcal{L}_{\text{coeff}}$ is the mean squared error over the swept coefficients of the batch, each normalized to $[0, 1]$ by a fixed analytic map (for example $\log_{10} \mathrm{Ra}/12$; Table 6), which reproduces every stored target to within 10^{-6}. $\mathcal{L}_{\text{regime}}$ is the mean cross-entropy over the frames of regime-classified families. Table 8 lists the optimization recipe and the compute budget, and Table 9 the per-family exposure. The run has two legs: a uniform-mix run on a v5e-16 pod, resumed from its pre-cooldown checkpoint on a v5e-64 pod with the same recipe. Its chip-hours include 214 replayed steps and exclude per-launch setup.

Sampling. Each batch element is a single frame, drawn from a uniformly chosen family (weight $1/9$), a random trajectory and a frame index uniform over the whole trajectory, including $t = 0$. Uniform family sampling equalizes the number of frames drawn per family. It does not equalize their loss contribution: the coefficient loss counts two swept coefficients per frame for active matter, RB and shear flow, one for Euler and none for the other families. Nor does it equalize frame exposure: small families (active matter, viscoelastic instability) are revisited tens to hundreds of times more often than Gray–Scott or Euler (Table 9). The only overfitting we observed from this revisiting occurred in continued training beyond the released step, on active matter (Training-length row of Table 13).

Table 8: **Training recipe and budget of the released feature map** (step 22,176). Chip-hours are logged-step wall time × chips, summed over launches. Per-family exposure is in Table 9.

Item	Value
Architecture	ResNet-18 v1.5: 7×7 stem, basic blocks [2, 2, 2, 2] with widths 64/128/256/512, input 512×512×13, global average pool → 512-d feature, one linear head → 46 logits/outputs
Parameters	11,231,470 trainable (head 23,598) + 9,600 BN running statistics; trained from scratch
Loss	$\mathcal{L}_{\text{fam}} + 2\,\mathcal{L}_{\text{coef}} + 0.5\,\mathcal{L}_{\text{reg}}$ (CE; masked MSE over swept coefficients of the true family; CE over regime families)
Optimizer	AdamW, $\beta = (0.9, 0.999)$, weight decay 10^{-4} on tensors with ndim>1, no gradient clipping
LR schedule	linear warmup to 10^{-4} over 5 nominal epochs (840 steps), constant, then linear to 0 over steps 21,168–22,176; not rescaled when the global batch grew
Batch / hardware	64 frames per host; TPU v5e, steps 0–15,120: 4×4 = 16 chips, global batch 256; 15,120–22,176: 16×4 = 64 chips, global batch 1,024
Sampling	families uniformly (1/9 each); one uniformly random frame per sampled trajectory; no augmentation
Steps / examples	22,176 steps (132 nominal epochs of 168 steps); 11,096,064 frames (3,870,720 + 7,225,344)
Passes	3.17–3.18 mean passes over 3.49–3.50 M realized unique train frames (the range comes from viscoelastic T); highly non-uniform by family
Precision	bf16 compute, fp32 parameters and optimizer state; features for the metric are extracted in fp32
Compute	41.86 h pod wall-clock (28.78 + 13.08); 1,297.77 v5e chip-hours (460.48 + 837.29)

Table 9: **Per-family exposure over 11,096,064 sampled frames.** Passes: expected draws per family (1,232,896 under the uniform family mix, spread evenly over trajectories and frames) divided by unique train frames. Viscoelastic: a range, because the per-regime train split was not read (289.4 passes on a $T = 20$ trajectory, 96.5 on a $T = 60$ one). Static spec: the nominal per-split size $N \times T$ of the dataset specification, not the trajectories present at training time (Table 5).

	realized (at training)				static spec	
Family	Train traj.	T	Unique frames	Passes	Frames	Passes
Shear flow	896	200	179,200	6.88	179,200	6.88
Rayleigh–Bénard	1,400	200	280,000	4.40	280,000	4.40
Euler, open BC	4,000	101	404,000	3.05	808,000	1.53
Euler, periodic BC	4,000	101	404,000	3.05	808,000	1.53
Active matter	175	81	14,175	86.98	23,328	52.85
Viscoelastic	213	20/60	4,260–12,780	96.5–289.4	4,160	296.37
Poseidon CE	32,808	21	688,968	1.79	809,760	1.52
Poseidon NS	26,318	21	552,678	2.23	824,880	1.49
Gray–Scott	960	1001	960,960	1.28	960,960	1.28
All (mean)	70,770		3,488,241–3,496,761	3.17–3.18	4,698,288	2.36

B.4 SUPERVISED PERFORMANCE

Tables 10–12 report held-out performance on the full validation split (288,704 frames, all nine classes) and on the test split where it was scored (RB, shear flow, active matter). These splits hold out initial conditions, not coefficient values: on RB and shear flow the coefficient grids are identical across train, validation and test. Per-frame family accuracy on validation is 0.9994. All 165 errors are swaps between the two Euler classes, which share every field and differ only in boundary condition. On validation, coefficient regression is strongest for Euler γ (R^2 0.99), the active-matter ζ and α (0.96–0.97) and the Rayleigh number (0.93). It is weakest for the shear-flow Reynolds number (R^2 0.66; macro grid accuracy 0.57 over all validation frames, including the early transient), which a single frame constrains only partly. Exact energy, enstrophy and palinstrophy of the velocity predict it with $R^2 = 0.688$, and with $R^2 = 1.000$ once time derivatives are added (Appendix H). The multi-family mix is not what limits the Reynolds regression. A shear-only retrain with the same input and loss, step-matched to a dedicated single-frame shear regressor, matches the released feature map's z-normalized error to two decimals (Table 13, multi- vs single-family row). Longer nine-class training (180 to 360 cumulative epochs) lowers the nine-class feature map's error to between 0.28 and 0.31.

Comparison with published coefficient-regression baselines. Against the four baselines in Table 1 of Qu et al. (2026), the released feature map is best on active matter, between JEPA and MPP on shear flow, and behind all four on RB. We quote errors in their z-normalized units: MSE averaged over a family's two coefficients. The released feature map reaches 0.029 / 0.441 / 0.199 on validation and 0.037 / 0.531 / 0.209 on test for active matter / shear flow / RB. The published values are 0.079 / 0.38 / 0.13 (JEPA, Qu et al., 2026), 0.160 / 0.67 / 0.18 (VideoMAE, Tong et al., 2022),

Table 10: **Family classification (9-way, per frame, EMA BatchNorm).** *traj. vote* is the valid accuracy of a per-trajectory majority vote over frames; tied votes (0 in valid) go to the higher mean probability. The test split was scored for 3 families only.

	valid		test		traj.	
Family	frames	recall	frames	recall	vote	valid errors → predicted
Shear flow	22,400	1.0000	22,400	1.0000	1.000	–
Rayleigh–Bénard	35,000	1.0000	35,000	1.0000	1.000	–
Euler, open BC	50,500	0.9971	–	–	0.998	144 → Euler, periodic BC
Euler, periodic BC	50,500	0.9996	–	–	1.000	21 → Euler, open BC
Active matter	1,944	1.0000	2,106	1.0000	1.000	–
Viscoelastic	680	1.0000	–	–	1.000	–
Poseidon CE	5,040	1.0000	–	–	1.000	–
Poseidon NS	2,520	1.0000	–	–	1.000	–
Gray–Scott	120,120	1.0000	–	–	1.000	–
All	288,704	0.9994	59,506	1.0000		165 errors

Table 11: **Coefficient regression** in normalized $[0, 1]$ units (Table 6). $R^2 = 1 - \mathrm{SSE}/\mathrm{SST}$ over the split's frames. Grid accuracy snaps each prediction to the nearest grid value: *micro* over frames, *macro* the mean per-grid-value recall (chance $= 1/\mathrm{grid}$). Trajectory accuracy: a majority *vote* of the snapped frame predictions, or the *mean* frame prediction, snapped; tied votes (3) go to the grid value nearest the mean prediction. Validation: the full split; test: scored for 3 families.

							frame grid acc.		trajectory acc.		
Family	Coef.	Split	frames	MSE$\times 10^3$	R^2	grid	micro	macro	n	vote	mean
Shear flow	Re	valid	22,400	1.993	0.657	4	0.571	0.571	112	0.670	0.670
Shear flow	Re	test	22,400	2.214	0.619	4	0.564	0.564	112	0.652	0.652
Shear flow	Sc	valid	22,400	7.383	0.738	7	0.442	0.442	112	0.562	0.509
Shear flow	Sc	test	22,400	8.652	0.693	7	0.422	0.422	112	0.509	0.491
Rayleigh–Bénard	Ra	valid	35,000	1.017	0.927	5	0.911	0.911	175	1.000	1.000
Rayleigh–Bénard	Ra	test	35,000	0.918	0.934	5	0.918	0.918	175	1.000	1.000
Rayleigh–Bénard	Pr	valid	35,000	5.316	0.811	7	0.758	0.758	175	1.000	0.749
Rayleigh–Bénard	Pr	test	35,000	5.565	0.803	7	0.765	0.765	175	1.000	0.743
Euler, open BC	γ	valid	50,500	0.232	0.992	10	0.880	0.880	500	0.890	0.894
Euler, periodic BC	γ	valid	50,500	0.123	0.996	10	0.903	0.903	500	0.946	0.936
Active matter	ζ	valid	1,944	2.172	0.963	9	0.821	0.820	24	1.000	1.000
Active matter	ζ	test	2,106	2.395	0.964	9	0.823	0.824	26	1.000	0.962
Active matter	α	valid	1,944	1.982	0.969	5	0.977	0.977	24	1.000	1.000
Active matter	α	test	2,106	2.945	0.953	5	0.962	0.954	26	1.000	1.000

Table 12: **Regime classification** on the full validation split (the scored test split contains no regime family). *classes*: classes present in validation / head width. The Poseidon rows therefore measure a reduced task. Macro recall averages over the classes present; tied trajectory votes (1) go to the higher mean probability.

Family	classes	frames	acc.	macro rec.	traj.	vote	per-class recall
Viscoelastic	7/7	680	0.8382	0.7804	22	0.773	AH 1.000, CAR 1.000, EIT 1.000, transtochaos_EIT_L 0.550, transtochaos_EIT_SAR 0.433, transtononchaos_EIT_L 0.917, transtononchaos_EIT_SAR 0.562
Poseidon CE	2/4	5,040	1.0000	1.0000	240	1.000	CE-CRP 1.000, CE-RP 1.000
Poseidon NS	1/2	2,520	1.0000	1.0000	120	1.000	NS-Sines 1.000
Gray–Scott	6/6	120,120	0.9895	0.9895	120	0.992	Bubbles 0.999, Gliders 0.997, Maze 0.996, Spirals 0.947, Spots 0.999, Worms 0.999

0.057 / 0.13 / 0.01 (DISCO, Morel et al., 2025) and 0.230 / 0.59 / 0.08 (MPP, McCabe et al., 2024). On active matter the z-map leaves the targets with variance other than 1 on both splits (0.875 for ζ and 0.810 for α on validation), so that column is approximate. R^2 is invariant to affine maps of the target, and the z-map is affine in $\log \mathrm{Re}$, $\log \mathrm{Ra}$, ζ and α, so for these coefficients the R^2 of Table 11 equals R^2 in the z-units. Sc and Pr are log-scaled in our map but linear in the z-map, so their R^2 differs. The comparison is indicative only, because the protocols differ. The baselines are frozen pretrained encoders read out by an attentive probe (MPP is fine-tuned end to end) on 16-frame clips. The PDE-KD feature map was trained with a supervised head on these coefficients and sees one frame. The baselines are scored on the validation split, as our validation row is, but

on a 9,600-window subset of it for shear flow and RB (MPP at its best epoch); our row uses every validation frame.

Out-of-range coefficients. Out of range, every RB frame is still classified as RB, and the Rayleigh head preserves the ordering of the coefficients. We generated RB trajectories at $\mathrm{Ra} = 10^5$ and 10^{11}, one decade outside the training grid, with our own solver setup. All 14,000 out-of-range frames, and all 35,000 in-range test frames, are classified as RB. Over all frames 0–199, the Rayleigh head extrapolates with a median bias of $+0.61$ decades at 10^5 and -0.46 at 10^{11}, against ≤ 0.07 in distribution. We report medians because early frames fall back towards the grid center (22.4% of frames 1–9). We have no in-distribution control from our solver, so solver differences remain confounded with the coefficient shift.

B.5 Feature-map design choices and ablations

Table 13 collects the comparisons made while building the feature map. All are single-seed. Most are continuations or siblings within one lineage; the family-sampling, normalization-layer, regime-weight and held-out-family runs were trained from scratch. The rows marked (FID) and the training-amount row report FID in each variant's feature space; the others report supervised metrics only. Only the training-amount checkpoints were also scored with KD, on an earlier version of the RB spectral suite. We did not train a feature map with family supervision alone, obtain a held-out result for one trained below 512^2 resolution, or evaluate a ResNet-50 backbone.

Released checkpoint and loss weights. The released checkpoint was the best by coefficient MSE when the disturbance suites were first run on it; no other selection criterion is recorded. The later big-batch continuations (Training-length row) lower the validation MSE of every swept coefficient but were not adopted for the 2D feature map; the clip feature map's parent is one of them (Appendix C.3). The loss weights $\alpha = 2$, $\beta = 0.5$ were not swept on this architecture. A sweep on an earlier velocity-only feature map moved family accuracy by at most 0.002 and viscoelastic regime accuracy by 0.12 across $\alpha = \beta \in \{0.1, 0.2, 0.3, 0.5\}$.

Feature-map generations. Generation 1 read the two raw velocity or momentum channels of eight classes (no Gray–Scott). Generation 2 read one z-scored channel at a time, with a channel-type head, on nine classes. The released feature map reads all 13 slots jointly.

B.6 Evaluation data

Table 14 lists the datasets scored in the paper and appendix and the pools drawn from them. Scored pools drawn from a dataset come from its held-out (validation or test) splits, with the following exceptions.

- The RB IC-conditional experiment uses all splits. Its training-split exposure is controlled for the exceedance probe of Appendix F but not for the permutation test of Section 4.2.
- The turbulent radiative layer, which is not in the feature map's training corpus, uses all splits.
- The HRRR hour and month axes, and their triviality and regional tests, use the HRRR training split (July 2024 to June 2025); HRRR is not in the feature map's training corpus.
- The coefficient row of Figure 1 is an RB training trajectory, and 17 of the 20 RB trajectories behind the seed-pair spread of Appendix A.2 come from the training split.

Clip pool. The PDE-KD (3D) clip suite scores the same 350-IC validation-plus-test pool as the per-frame perceptual suite (Appendix E.1), split 175 clean / 175 disturbed. Its clip features were first extracted from a copy of the RB test split in which the $\mathrm{Ra} = 10^9$, $\mathrm{Pr} = 2$ cell held the 40 trajectories of the corresponding training cell (Table 5, note b). These 40 of the copy's 385 reference ICs (10%) were therefore training data of the clip feature map and of its 2D parent. We dropped these 40 trajectories and extracted the cell's 5 test trajectories with the same checkpoint. The disturbed clips of these 5 trajectories, and of one IC moved to balance the arms, come from a later version of the disturbance sampler. The arms follow the original split of the clip extraction rather than split seed 0. The interleave and switching disturbances mix each clip with the next five candidate-arm ICs (Appendix E.2). The disturbed clips of the up to five ICs that preceded the dropped trajectories in

Table 13: **Feature-map ablations** (single seed; full validation split unless noted). Coeff.: pooled coefficient MSE in $[0, 1]$ units; z: per-family z-normalized MSE of Qu et al. (2026) for active matter / shear flow / RB; *: an earlier feature-map generation (see text); FID: computed in that variant's feature space.

Choice	Variants	Result
Input representation	raw velocity (generation 1)*; one channel at a time with a channel-type head (generation 2)*; 13-slot joint input (released)	Euler γ at chance (grid accuracy ≈ 0.10) in generation 1 (no thermodynamic channels) and generation 2 (one channel at a time), and for a momentum-only feature map trained on Euler-open alone with an eightfold frame budget (0.096 / 0.106; an identifiability limit, not a mixing effect); the released feature map reaches MSE 1.2–2.3 $\times$ 10^{-4}. z: generation 2 (best single-channel readout; different budget), 0.120 / 0.760 / 0.279 $\rightarrow$ released, 0.029 / 0.441 / 0.199
Family sampling	$\propto \sqrt{N_{\text{frames}}}$ vs uniform $1/9$ (step 19,152)	coeff. 0.00264 vs 0.00254; z active matter 0.055 vs 0.034; shear 0.579 vs 0.549; RB 0.221 vs 0.224
Batch / epochs / LR	$+36$ epochs at batch 1024 (released) / $+90$ at 256 / $+36$ at $4\times$ LR	coeff. 0.00208 / 0.00217 / 0.00218: larger batch $\approx$ longer training; higher LR does not help the coefficients
Training length	90, 180, 270, 360 cumulative flat epochs at batch 1024 (released: 36)	coeff. 0.00183, 0.00157, 0.00142, 0.00136; active-matter MSE $+60\%$ from 270 to 360 epochs (0.00107 to 0.00171), still below the released feature map's 0.00208 (mean over ζ and α, Table 11)
Regime weight β	0.5 (90 + 24 cooldown epochs) vs 0.1 (90 + 6), from scratch	coeff. 0.00254 vs 0.00271; viscoelastic regime 0.837 vs 0.825; confounded with cooldown length, no conclusion
Normalization layer	BatchNorm vs GroupNorm(16), twins differing only in the normalizer ($\beta = 0.1$, 90 + 6 epochs from scratch; not the released recipe)	coeff. 0.00271 vs 0.00338 (GroupNorm worse)
BatchNorm at inference	stored EMA / EMA recalibrated on 500 batches / PreciseBN on 500 batches / batch statistics	validation loss 0.0221 / 0.0260 / 0.0260 / 4.52; batch statistics collapse family accuracy to 0.217
BatchNorm inside the metric (FID)	stored EMA vs batch statistics, released checkpoint, RB perceptual suite	split-half FID floor 0.218 vs 0.156; the range of FID across the ladder shrinks from about 5.4 to 1.7 orders of magnitude (salt-and-pepper L5: 54,823 vs 7.76)
Checkpoint (FID)	step 19,152 vs the released 22,176, RB perceptual suite	floor 0.224 vs 0.218; every family strictly monotone in level in both; a modest shift in relative sensitivity
Training amount $\rightarrow$ FID floor	six checkpoints from step 19,152 to 76,608	RB split-half FID floor falls monotonically 0.224 $\rightarrow$ 0.125 (single draws; each checkpoint is its own feature space, so descriptive); relative sensitivity shifts toward phase and fine scales because band-energy sensitivity shrinks
Loss weights*	$\alpha = \beta \in \{0.1, 0.2, 0.3, 0.5\}$	family accuracy 0.9960 $\rightarrow$ 0.9943; viscoelastic regime 0.48 $\rightarrow$ 0.60
Augmentation*	none vs translate/rotate/reflect	family accuracy 0.9954 vs 0.9942; neutral to slightly negative; the released feature map uses none
Multi- vs single-family	13-slot input and loss on shear flow only: 96 epochs / step-matched to the dedicated regressor / the latter plus a 120-epoch LR cooldown	shear z 0.493 / 0.443 / 0.303 (BatchNorm recalibration alone reaches 0.349, about two-thirds of the cooldown gain), vs 0.441 for the nine-class feature map and 0.296 for a dedicated single-frame shear regressor
Held-out family	feature-map recipe with shear flow removed (uniform 1/8 mix, batch 1024 throughout, from scratch)	8-class validation vs the released feature map: pooled coeff. 0.00151 vs 0.00142, active-matter error doubles, regime cross-entropy 0.197 vs 0.031; not scored with KD (Appendix I.6)

that order therefore still mix in training trajectories; they were not re-extracted. The PDE-KD (3D) clip-suite results (Table 2 and Appendix E.6), including its clean-versus-clean rates in Table 20, use the corrected pool, and its spectral-suite results never used the uncorrected copy. The test ICs of every per-frame pool are the 175 of the test split, without the duplicated trajectories.

Table 14: **Evaluation datasets.** In corpus: the data come from one of the feature map's training classes (the out-of-range RB set shares the family but not the data source or coefficient values). Frames are native frame indices of the scoring window. [a]The training split is in the corpus. [b]Generated with our own solver setup.

Dataset	In corpus	Units scored	Used in
RB (Well), test	yes	175 ICs (35 (Ra, Pr) cells × 5); frames 61–199 (spectral suite); leads 1–183 = native frames 17–199 (model ladders)	Section 4.1, Appendices E, G.3 and G
RB, validation + test	yes	350 ICs: perceptual full-pool tests (175 vs 175; power sweeps draw from a fixed 175-IC subset); symmetry suite	Section 4.1, Appendix E
RB, validation + test (clips)	yes	PDE-KD (3D): 350 ICs (validation 175, test 175; corrected pool, see text), split 175 clean / 175 disturbed; JEPA: 350 (175 / 175); clips start at frames 65–180 (stride 5)	Section 4.1, Appendix E.6
RB, all splits	yes[a]	1,750 trajectories = 35 cells × 5 ΔT_{init} × 10 seeds; 20-frame windows	Section 4.2, Appendix F
RB, out-of-range Ra	no[b]	14 files at $\mathrm{Ra} \in \{10^5, 10^{11}\}$, 7 Pr; 14,000 frames (all frames 0–199)	Appendix B.4
Shear flow (Well), test	yes	112 ICs, frames 61–199	Table 3
Euler, periodic (Well), test	yes	175 of 500 ICs, frames 20–100 (disturbances); 500 ICs, frames 0–100 every 4th (symmetry audit); 500 ICs, leads 64–88 (trained neural simulators; Walrus vRMSE, TV and EOS values at leads 61–91)	Table 3, Section 4.3, Appendices E.7 and G.5
Euler, open (Well), test	yes	500 ICs; leads 64–88 (ladder); frames 0–100 every 4th (symmetry suite)	Appendices G.3, G and E.7
Euler, both BCs	yes	10 adiabatic-index cells × 50 ICs	Appendix G.6
Gray–Scott (Well), test	yes	120 ICs, frames 862–1000	Table 3
Turbulent radiative layer (Well)	no	9 cooling times × the same 10 ICs, frames 0–99 (IC memory); all 90 trajectories, frames 20–100 (controlled disturbances)	Appendices F and I.2
FSI (RealPDEBench)	no	17 (Re, m^*) cells × 3 replicates, PIV and CFD arms, 128^2, frames 61–199	Appendix I
HRRR surface	no	363 training-split day trajectories (hour and month axes); 120 test-split day trajectories (controlled disturbances)	Table 3, Appendix I
ERA5 surface	no	24 month trajectories	Appendix I
WeatherBench-2	no	8 forecast systems (of 10 scored; the two 12-hourly GenCast systems enter only an even-lead comparison, Appendix I.4), 117 initializations in 2020 (24 for the ENS-member system), 40 leads at 6 h; 2 m temperature, mean-sea-level pressure and 10 m wind on the 1.5° WeatherBench-2 grid	Appendix I.4

C Comparison feature spaces and metrics

Every space except Spectrum-native and JEPA reads the same canonical tensor as the PDE-KD feature map (Appendix B.2): z-scored at native resolution, upsampled to 512×512 and scattered into 13 slots. Poseidon then resizes its occupied slots to 128^2, and PDE-KD (3D) reads 20-frame clips of the tensor. Disturbances therefore corrupt identical inputs for every such space. Every KD uses Eq. (7) with d the space's own dimension. Table 15 summarizes the spaces.

C.1 Poseidon features and Fréchet Poseidon Distance

Cheng et al. (2025) define the Fréchet Poseidon Distance (FPD) for single-field problems. The field enters Poseidon's first input channel, the other channels are zero-filled, and a Fréchet distance is taken on the mean-pooled encoder embedding. Its direct multi-field extension, which we tested, runs one pass per field and concatenates the embeddings (768 per field, 3072 on RB). We instead run one forward pass with the physical fields routed to the input channels on which Poseidon was pretrained (768 dimensions), and we score it with KD like every other space.

Routing is the stronger construction. We compare the two constructions by their separation from each space's own null, a quantity comparable across spaces (Appendix C.7). On matched RB ICs, KD at the top level of the perceptual suite exceeds the clean split-half floor by 146, 62 and 183 of that floor's jackknife standard errors for occlusion, noise and salt-and-pepper in the routed space, against 29, 17 and 135 in the per-field space. Neither space resolves blur (one run per space; descriptive). The FID floors of the two spaces (0.81 against 3.21) are not comparable and mainly reflect the 768 versus 3072 dimensions. On an unseen physics, concatenation also dilutes information: in the per-field space, one field's block carries most of the variance while another carries most of the IC identity. The per-field construction was scored on the RB disturbance suites but is not in the power tables.

Routing requires that the physics provide the fields. On both Euler datasets (periodic, and open as used in Table 47), slot 1 holds energy. Poseidon's density channel therefore receives energy, its

Table 15: **Feature spaces.** d is the feature dimension on RB (4 channels; the spectra treat the two velocity components as one field, so $765 = 3 \times 255$ and $189 = 3 \times 63$). "Fit" is the data the space was fit on, if any.

Space	Construction	d	Fit
PDE-KD	ResNet-18, 13-slot input, global-average-pooled penultimate layer (Appendix B)	512	nine classes (train splits)
ImageNet	torchvision ResNet-18 (`IMAGENET1K_V1`); stem inflated from 3 to 13 input channels by averaging the RGB filters and scaling by $3/13$; same z-scored input, no ImageNet normalization; average-pool features	512	ImageNet (Deng et al., 2009)
Poseidon	Poseidon-B (ScOT) encoder; occupied slots antialias-resized to 128^2; one forward pass with (density, v_x, v_y, pressure) filled from slots (1, 3, 4, 0); mean over last-stage tokens	768	PDE pretraining
Random init	the PDE-KD network (same ResNet-18, 13-slot input, pooled penultimate layer) with random weights and BatchNorm statistics calibrated on RB training frames (Appendix C.2); three seeds, reported as their mean	512	BatchNorm statistics only (RB train frames)
Pixels	128 fixed grid sites (seed 0), all occupied channels read at the same sites	512	none
Spectrum	per field: remove the mean along the periodic axis, orthonormal 2D FFT, sum power over vector components, radial shells up to $\lvert k\rvert \le 0.5$ cycles/px, DC dropped, $\log_{10}(P + 10^{-6})$	765	none
Spectrum-native	as Spectrum on the native 512×128 grid (no perceptual suite)	189	none
PCA-white	top $k = 370$ principal components of the flattened canonical tensor, divided by $\sqrt{\lambda_i}$	370	RB train, 14,000 frames
PCA-raw	as PCA-white without the division	370	RB train (same basis)
PDE-KD (3D)	ResNet3D-18 inflated from step 46,368 of a separate continuation of the 2D feature map's training (not the released checkpoint; Appendix C.3); temporal convolutions replicate-pad and are valid in time; clips of $T = 20$ consecutive frames	512	nine classes (clips)
JEPA	convolutional JEPA encoder of Qu et al. (2026), retrained on RB with sample-axis VICReg (Bardes et al., 2022) statistics (checkpoint 29, after 30 epochs); RB-specific crop and 224^2 staging; spatial mean pool; clips of $T = 16$	128	RB (self-supervised)

velocity channels momentum and its pressure channel pressure, and density is dropped. On Gray–Scott it receives only the A field, and on the turbulent radiative layer it receives no density, because that field occupies a slot that is not routed. Staging also costs resolution. Poseidon's 128^2 input is a $4\times$ downsampling of the 512^2 canonical grid, after which only 17%, 49% and 67% of the root-mean-square change made by a $\sigma = 1$, 3 and 5 pixel blur (levels L1, L3 and L5) survive on RB, against 98% for an L5 occlusion. Poseidon's blur results, and plausibly its results on other fine-scale disturbances, carry this handicap.

C.2 Untrained-network control (Random init)

Randomly initialized networks have been proposed as embeddings for evaluating generative models (Naeem et al., 2020). To separate what training contributes from what the architecture, the input staging and the kernel contribute, we score the PDE-KD network *without training*, labeled Random init in tables and figures. The network is that of the released feature map: ResNet-18 with the 13-slot stem and the same padding, with the pooled 512-dimensional penultimate layer as the feature. Its weights follow the standard torchvision initialization: Kaiming-normal convolutions (fan-out), BatchNorm (Ioffe & Szegedy, 2015) scale 1 and offset 0, and no zero-initialized residual branches. We do not use the Flax initialization from which the feature map was trained, because it zero-initializes the last BatchNorm scale of every residual block, so an untrained copy would reduce to the stem and the three projection shortcuts. That zero-residual initialization is the random floor of the probes in Appendices H and I.5; the untrained network scored here instead keeps every residual branch active. BatchNorm running statistics are set to their cumulative average over 4,760 RB *training* frames (140 trajectories, every sixth frame), shuffled across trajectories. No weight is trained, and no label or evaluation trajectory is involved. Without this calibration, BatchNorm is the identity and the features drift in scale: $\gamma\langle x, x\rangle$ of Eq. (7), with the kernel scale $\gamma = 1/d$, is 3.1, against 0.72–0.78 over the three seeds after calibration and 0.18 for the trained feature map. The same calibrated networks are used on every physical system, as the trained feature map is; they are not recalibrated per system.

Protocol. Every experiment runs under the protocol of the other spaces: the same reference pools, frame windows, split seed 0, estimators, designs and drivers. Single IC-level permutation tests are reported at both $B = 2{,}000$ and $B = 20{,}000$ for every space. Power and n_{80} use the sweep convention ($B = 999$ per replicate, 200 IC replicates). The conditional permutation test enumerates

its relabelings, the conditional exceedance probe uses a per-pair threshold, and the model-ladder correlations use exact permutation floors, so none of them has a B.

Seeds and coverage. Three initialization seeds use the same recipe and calibration frames. The RB disturbance, conditional and model-ladder tables and figures report one *Random init* row or column: the mean over the three seeds. The multi-physics suite (Tables 3 and 31) includes it with the same RB-calibrated networks. On Euler, the kernel scale $\gamma\langle x, x\rangle$ of these networks is 2.2–4.3, against 0.72–0.78 on RB, 0.7–1.0 on shear flow and Gray–Scott, and 1.2–2.3 on out-of-distribution HRRR. One Euler seed exceeds 3.1, the upper end of the range that our feature-scale check accepts, so on Euler the RB-calibrated BatchNorm is off-scale. Appendix A also includes it, as the seed mean of each pair's KD, and so do the Euler-open per-functional entries (Table 49). Among the out-of-distribution datasets, only HRRR includes it (Table 3). The clip suites, the symmetry suite (Figure 9) and the physical-law panels (Figure 11) do not include it.

Averaging over seeds. Power is averaged over seeds at each rung and sample size, and n_{80}, censoring and the bold-face rule of the power tables are derived from that averaged power curve. With 200 replicates per seed, the averaged power is the rejection rate over 600 replicates of a test whose network is drawn at random from the three. Averaging the three n_{80} values instead is undefined whenever the seeds are censored differently. Single-test rejection counts, Spearman correlations and conditional power are averaged over seeds directly. Standard errors and bootstrap intervals are conditional on the three networks. The seed rows of Table 16 give the seeds separately.

Results. Random init responds strongly to large changes of band energy (low- and medium-k injection and removal, small-scale tilt, severe high-k injection from L3) and to high-amplitude local defects (occlusion at L4–L5, salt-and-pepper at every level). It responds weakly to phase, truncation, high-k removal, large-scale tilt, noise and blur, and to occlusion at L1–L3 (Tables 29 and 30). On these families PDE-KD reaches 80% power on noise at every level, on occlusion, blur and large-scale tilt from L2, on both phase families from L3, on truncation from L4, and on high-k removal at L5 of the canonical ladder and from L2 of the severe ladder (Tables 29, 30 and 27). Table 16 gives, for every space and suite, the rungs reaching 80% power at 175 ICs and the rungs whose single full-pool test rejects at $B = 2{,}000$ and at $B = 20{,}000$. Raising B from 2,000 to 20,000 changes at most two verdicts of any space. The Random init columns of Tables 29, 30, 26 and 27 compare it with PDE-KD rung by rung, and those of Tables 43, 47 and 32 give its conditional and model-ladder entries. The three seeds reject in 0 to 4 of 40 clean-versus-clean draws per cell (both estimators, all three designs; Table 20), and no cell is flagged by the rule of Appendix D.6. The severe and canonical removal ladders share their top rung, and the untrained features on that rung are bit-identical between the two ladders for every seed.

Table 16: **Random init: rung counts on the RB disturbance suites.** Each cell: *detected / rejected at* B=2,000 */ rejected at* B=20,000. Detected: disjoint KD power at the full 175-IC pool reaches 0.80; rejected: the single IC-level permutation test rejects at α=0.05. Random init: the untrained PDE-KD network, mean over three initialization seeds (Appendix C.2); it detects on the seed-averaged power curve and averages the rejection counts. Seed 0–2: rungs detected on each seed's own power curve; their rejection counts enter only the Random init mean (–). –: not applicable (Spectrum-native scores only the spectral suites).

space	perceptual (20)	spectral, energy $\leq$2 (55)	severe inject (15)	severe remove (15)
PDE-KD	18 / 18 / 18	30 / 41 / 42	14 / 15 / 15	14 / 15 / 15
ImageNet	18 / 20 / 20	24 / 26 / 26	11 / 12 / 12	9 / 10 / 10
Poseidon	12 / 14 / 14	15 / 29 / 29	13 / 13 / 13	10 / 13 / 13
Random init	7 / 16.7 / 16.3	12 / 21.0 / 21.3	13 / 15.0 / 15.0	10 / 9.3 / 9.3
seed 0	7 / – / –	11 / – / –	13 / – / –	10 / – / –
seed 1	8 / – / –	14 / – / –	13 / – / –	10 / – / –
seed 2	7 / – / –	12 / – / –	13 / – / –	10 / – / –
Pixels	9 / 12 / 11	16 / 18 / 18	12 / 13 / 13	9 / 10 / 10
Spectrum	19 / 19 / 19	21 / 22 / 22	8 / 12 / 12	9 / 9 / 9
Spectrum-native	–	17 / 18 / 18	6 / 8 / 8	10 / 9 / 9
PCA-white	13 / 13 / 13	39 / 39 / 39	14 / 14 / 14	13 / 14 / 14
PCA-raw	2 / 7 / 7	7 / 9 / 9	9 / 11 / 11	5 / 5 / 5

C.3 THE CLIP FEATURE MAP, PDE-KD (3D)

The clip feature map is the 3D inflation (Carreira & Zisserman, 2017) of Section 3. Every convolution except the $1 \times 1 \times 1$ shortcut projections gains a temporal width of 3, initialized so that the network computes its 2D parent frame by frame. Temporal padding replicates edge frames, and the temporal stride is 1, which gives a temporal receptive field of 35 frames (radius 17), wider than the 20-frame clip. Features are the global average over time and space (512 dimensions; 33.3M parameters).

The 2D parent is step 46,368 of a two-stage continuation that branched from the released run at step 21,168, before its cooldown; it is not the released checkpoint. The parent is stronger than the released checkpoint on the supervised tasks (z-errors 0.020 / 0.312 / 0.166 against 0.029 / 0.441 / 0.199 for active matter / shear flow / RB on validation). The clip feature map was fine-tuned on clips of $T = 20$ consecutive frames from the same nine-class training split with the same loss. The optimizer was AdamW (Loshchilov & Hutter, 2019) with a 5-epoch linear warmup to 10^{-4}, a flat phase, and a linear decay to 0 over the last 6 of 96 epochs. Training used 32 clips per step on a v5e-16 pod for 129,312 steps (4.1M clips, about 716 v5e chip-hours). An auxiliary temporal-shuffle loss was switched off. With that loss on, temporal order becomes perfectly decodable (probe AUC 1.000 against 0.966), with no consistent gain on the supervised tasks (z-errors 0.009 / 0.427 / 0.129 with it and 0.010 / 0.459 / 0.107 without, in the same order). The clip feature map improves on its parent for active matter and RB but not for shear flow.

At evaluation, clips start every 5 frames at frames 65–180 (24 starts) and are scored as unordered populations of clips. The clip-suite pool is the 350-IC validation-plus-test pool of the frame rows, with the correction of one test coefficient cell described in Appendix B.6. A difference between the PDE-KD (3D) and PDE-KD rows of Table 2 confounds the clip input with two other changes: a stronger parent (step 46,368 against the released 22,176) and 96 epochs of fine-tuning on clips. It is not a controlled comparison of clips with frames.

C.4 THE JEPA COMPARISON ENCODER

The spatiotemporal comparison encoder is the convolutional JEPA of Qu et al. (2026), which we retrain on RB. Pretraining uses the RB training split of the Well (all 35 coefficient pairs) as 16-frame windows starting every 4 frames, cropped to the central 128×128 region of the 512×128 field, resized bilinearly to 224^2 and kept in physical units. The objective is VICReg (Bardes et al., 2022) with the coefficients of Qu et al. (2026) ($\lambda, \mu, \nu = 2, 40, 2$): the invariance term acts on the full, unpooled embedding, and the variance and covariance terms act on the sample axis of the mean-pooled clip embedding, gathered across devices, as in the published objective. An earlier run computed these statistics over the flattened batch and spatial axes, and its features collapsed with training (Appendix E.6). We train with AdamW for 30 epochs, against 6 in Qu et al. (2026), with learning rate 10^{-4}, 6 warm-up epochs, cosine decay to 10^{-6}, weight decay 0.05 and an effective batch of 256 clips (about 122 GPU-hours on eight GPUs). KD reads the 128-dimensional spatial mean pool of the final checkpoint (checkpoint 29, after 30 epochs) on 16-frame clips.

C.5 NON-LEARNED SPACES

The non-learned spaces of Table 15 (Pixels, Spectrum, Spectrum-native, PCA-white and PCA-raw) involve no network training. This subsection gives the construction details that the table omits.

Pixels. The 128 sites are drawn once, without replacement, and shared across channels, frames, arms and disturbances, so each feature vector is a joint sample of all channels at common locations.

Spectrum. The height-dependent mean profile of RB is removed before the FFT; otherwise that profile would dominate the lowest shells. Summing power over vector components makes the feature invariant to rotating those components. Because the canonical tensor is an upsampled grid, power at vertical wavenumbers above the native Nyquist limit comes from interpolation. Resizing also does not commute with a phase change, so the canonical spectrum is *not* phase-blind. Spectral disturbances are therefore also scored on the native grid (Spectrum-native), which is phase-blind up to float32 round-off (maximum change 4.8×10^{-7}).

PCA-white and PCA-raw. PCA-white retains the smallest of three cutoffs: parallel analysis against a column-permuted null (370), the one-standard-error rule on held-out reconstruction (976), and 99% explained variance (not reached within 1,024 components). Whitening then up-weights the last retained direction about $42\times$ relative to the first. PCA-raw omits the whitening. The RB basis is fit on 14,000 RB training frames. Other systems use their own bases, fit in-sample (on disjoint trajectory subsets of the scored split in the cross-physics suite, Appendix E.4), which favors this control.

C.6 Pointwise, ensemble and spectral metrics

vRMSE and CRPS act on z-scored fields per IC, lead and channel, with one global standard deviation per channel; the spectral error is computed in physical units.

- **vRMSE**: $\sqrt{\mathrm{SE}/(\mathrm{Var}_{\mathrm{frame}}(u) + 10^{-7})}$ per (IC, frame, channel), averaged over channels, leads and ICs (Ohana et al., 2024). For an ensemble model the scored prediction is its ensemble mean (Lola: 16 members), which is smoother than any member and therefore favored by vRMSE; KD scores the dumped members themselves. On RB, whose velocity starts at rest, the channel-mean vRMSE is dominated by velocity before lead ≈ 40. Over the 45 pairs of independent reference trajectories in one (Ra, Pr) cell used in Appendix A ($\mathrm{Ra} = 10^9$, $\mathrm{Pr} = 0.5$, native frames 61–199), channel-mean vRMSE has median 1.05 (range 0.74–1.20; 29 of 45 pairs above 1.0), comparable to the value 1.0 of a constant field at the spatial mean. After decorrelation, pointwise error therefore ranks a blurred prediction above a sharp but out-of-phase one (cf. Figure 1).
- **Spectral error**: four log-spaced radial bands in cycles per pixel; $\mathrm{specp}_b = \mathrm{rms}_{k\in b}|1 - (P_{\mathrm{pred}}(k) + \epsilon)/(P_{\mathrm{ref}}(k) + \epsilon)|$ with $\epsilon = 10^{-6}$ in physical units. Values above 1 can only come from excess power, so the error saturates on the deficit side.
- **CRPS**: the fair (ensemble-size-unbiased) CRPS (Ferro, 2014) per grid point, averaged. It is undefined for single-member models.
- **Energy distance and energy score** (Gneiting & Raftery, 2007; Székely & Rizzo, 2013; Sejdinovic et al., 2013): the energy distance is the same U-statistic as KD with kernel $k(x, y) = -\|x - y\|_2/\sqrt{d}$, computed in pixel space ($d = 262{,}144$ native RB) or in a feature space. The Appendix A figures report this energy distance between frame populations. With the Euclidean distance, the energy score of Section 2 is the corresponding proper scoring rule: its expectation over reference outcomes $u \sim P$ exceeds that of the ideal forecast $Q = P$ by half the energy distance between Q and P (before the $1/\sqrt{d}$ normalization). Replacing $\|\cdot\|_2$ by $\|\cdot\|_1/d$ gives a two-sample form of CRPS, which scores only per-pixel marginals.

Physical observables are defined in Appendix G.2.

C.7 Comparing spaces

KD magnitudes are not comparable across feature spaces. Three defensible normalizations of an ImageNet-to-PDE-KD effect ratio give 5.72, 0.92 and 0.11: the raw ratio, the ratio of the two spaces' effects each measured against its own response to an L5 occlusion on RB, and the ratio of effects each measured against the median symmetry-suite control that the space flags on the same physics (Appendix E.7). They disagree even on which space responds more strongly. We therefore compare spaces only through quantities defined on a common scale: rejection counts, power at a matched IC count, n_{80}, rank correlations, and separation from each space's own null.

D Statistical methodology

D.1 The KD estimator

Let arm X contain frames x_i and arm Y frames y_j, and let $g(\cdot)$ return the initial condition (IC) from which a frame was generated. With the cubic kernel of Eq. (7) ($\gamma = 1/d$, offset 1), we estimate

$$\widehat{\mathrm{KD}} = \frac{1}{|\mathcal{A}|}\sum_{(i,i')\in\mathcal{A}} k(x_i, x_{i'}) + \frac{1}{|\mathcal{B}|}\sum_{(j,j')\in\mathcal{B}} k(y_j, y_{j'}) - \frac{2}{|\mathcal{C}|}\sum_{(i,j)\in\mathcal{C}} k(x_i, y_j), \tag{10}$$

where $\mathcal{A} = \{(i, i') : g(x_i) \neq g(x_{i'})\}$ and $\mathcal{B}$ analogously exclude all within-IC pairs. When the two arms share ICs (the paired design), $\mathcal{C} = \{(i, j) : g(x_i) \neq g(y_j)\}$ also excludes same-IC cross-arm pairs; otherwise $\mathcal{C}$ is all pairs. This is the exact U-statistic over all frames of the cell, computed from per-IC block sums of one Gram matrix in float64, rather than the subset average of Bińkowski et al. (2018). It is unbiased for MMD_k^2 between the arms' marginal frame distributions when ICs are independent, identically distributed draws (for stratified pools, see Appendix D.2). It can be negative. It is undefined for fewer than two ICs per arm, and we mark cells with fewer than ten. Standard errors are delete-one-IC clustered jackknife estimates (Appendix D.4).

Split-half floor. Each cell's noise scale is a split-half floor: the mean over five random 50/50 IC splits of the reference against itself (split seeds 0–4). Each half holds half the ICs, so for FID this floor is conservative. On RB, a full-size validation-versus-test FID null is only 0.32–0.52 of the split-half floor in the three lead windows. For Lola-c16 at leads 61–183, for example, FID lies below its split-half floor but at 1.31 times its validation-versus-test floor. KD's split-half null is instead zero in expectation, because a random IC split is exchangeable; a stratified validation-versus-test null carries a small negative offset (Appendix D.2). We report FID only as a secondary diagnostic. Every RB cell is rank-starved for FID, because its 512×512 covariance rests on about 175 independent ICs.

D.2 Unit of independence

Frames of one trajectory are strongly dependent, so the textbook frame-level U-statistic (which drops only $i = i'$) and V-statistic are biased. The bias depends on lead time. Table 17 compares the three estimators on the same features: a 16-member Lola-c16 ensemble on 175 RB test ICs, each against its own validation-versus-test null of ground-truth frames. Both frame-level nulls are positive in every window ($+0.26$ to $+0.44 \times 10^{-3}$), whereas the IC-level null is slightly negative. At leads 1–20 both frame-level scores fall below their own null (by $8.0\times$ for frame-U and $6.7\times$ for frame-V). Relative to the IC-level estimator, frame-U is biased negative early, where cross-arm same-IC pairs compare a rollout with its own conditioning trajectory, and positive late, where within-IC autocorrelation dominates. Frame-V's self-pairs keep it above the IC-level value in every window. The gap to the IC-level estimator shrinks as $1/n_{\mathrm{IC}}$ along the IC axis. Along the ensemble-member axis it converges to a nonzero constant, so no ensemble size removes the bias of a frame-level estimator.

Table 17: **Estimator variants on the same features** (KD $\times 10^3$; in parentheses, each estimator's own validation-versus-test null; RB, Lola-c16, 175 ICs $\times$ 16 members). The IC-level estimator is Eq. (10).

Leads	IC-level	frame-level U	frame-level V
1–20	0.0641 (-0.085)	0.0512 ($+0.411$)	0.0660 ($+0.439$)
21–60	3.7408 (-0.271)	3.7764 ($+0.264$)	3.7881 ($+0.286$)
61–183	0.6621 (-0.244)	0.6744 ($+0.270$)	0.6770 ($+0.275$)

Residual bias under stratified pools. The IC-level estimator is exactly unbiased only when the two arms are exchangeable draws of ICs. Our reference pools are stratified: every arm holds the same number of trajectories per (Ra, Pr) cell. A within-arm term then sees $t - 1$ of an IC's t same-cell partners out of $n - 1$, while the cross term sees all t of n. Same-cell pairs have higher kernel values, so the estimate is pushed slightly negative. In a pure ground-truth null (validation versus test, leads 61–183), the IC-level estimate is -1.01×10^{-3} with 35 ICs per arm and -0.24×10^{-3} with 175. The predicted bias, $(\mathrm{frac}_{\mathcal{A}} + \mathrm{frac}_{\mathcal{B}} - 2\,\mathrm{frac}_{\mathcal{C}})\,\Delta k$ with the measured same-cell kernel excess Δk, agrees with the measurement in sign and order of magnitude (-1.32 to -0.26 predicted versus -1.01 to -0.24 measured, $\times 10^{-3}$, across five arm sizes; the prediction overshoots by 6–40%). The frame-level estimators are instead positive and decay as $1/n$. IC-level exclusion therefore removes the dominant, lead-dependent bias of frame-level estimators. What remains is a small negative offset specific to stratified validation-versus-test nulls. The disturbance tests instead split one pool at random, without stratification, so the expected KD under their null is exactly zero.

Shared versus disjoint ICs. KD may compare arms that share ICs; FID may not. On the 11 RB models of Appendix G.1, we split the 175 test ICs into halves of 87 and 88 with 10 seeds and score

nine lead windows (99 points). FID is *lower* on shared than on disjoint IC halves in 84 of the 99 points, by up to 0.33, or 3.9 standard deviations of the disjoint estimate across splits. The gap is 0.18–0.33 for every Lola model in every window, and near zero for the FNO and TFNO models, whose FID is high (10–43). FID has no same-IC exclusion, so sharing ICs biases it. KD excludes same-IC pairs, so its expectation over random splits is the same for shared and disjoint halves. Its observed gaps are at most 0.52×10^{-3} (2.0 disjoint standard deviations). They share one sign in 85 of the 99 points, because every model uses the same 10 splits: the disjoint KD correlates at 0.98 across models at leads 1–20. The 99 gaps are therefore one draw of split noise, not 99 independent draws.

D.3 Sample-count response

Table 18 varies the three sample axes of a KD cell independently: ensemble members per IC, leads per IC, and number of ICs. KD is flat along the member and lead axes, as an unbiased statistic should be. Along the IC axis its expectation is flat by construction. Every IC contributes the same number of frames, so Eq. (10), here in the paired design, is a U-statistic over ICs, and its average over all n-IC subsets equals the full-pool value. The 40-subset mean at 25 ICs lies 13% (2.3 standard errors of that mean) below the 175-IC value at leads 61–183. Because the subset-average identity fixes the expectation, this deviation is Monte Carlo error of the subset draws (subset s.d. 0.246), not small-sample bias, although a deviation this large has two-sided $p \approx 0.02$. A sixteen-member ensemble adds at most 4.9% effective samples to KD, measured by jackknife standard errors. The unit of evidence is the IC, not the frame or the member.

FID is not flat: it changes by up to 13% along the member axis and by 57% at one lead per IC. Along the IC axis (leads 61–183), the RB split-half FID floor rises tenfold, from 0.371 at 175 ICs to 3.839 at 15 ICs. In the validation-versus-test null, 83–87% of the FID null is its covariance term (means over five IC draws), which carries the dependence on n (0.52 at $n = 25$ to 0.16 at $n = 175$).

Table 18: **Sample-count response** of KD $\times 10^3$ and FID on Lola-c16 RB rollouts in the PDE-KD space. Members: 50-IC $\times$ 16-member pool, seeded random subsets. Leads per IC: 175-IC $\times$ 2-member pool, leads 61–183, mean of 5 seeded draws. IC count: 175-IC $\times$ 16-member pool, 40 seeded subsets (mean $\pm$ standard error of that mean). The three axes use different pools, so values are comparable within a row block only.

Axis		value of the axis					
Members		1	2	4	8	16	$\Delta_{1\to16}$
	KD, leads 21–60	1.2484	1.2611	1.2718	1.2976	1.3051	$+4.5\%$
	KD, leads 61–183	0.3117	0.3091	0.3090	0.3121	0.3164	$+1.5\%$
	FID, leads 61–183	0.2517	0.2336	0.2236	0.2191	0.2184	-13.2%
Leads per IC		1	4	16	64	123	
	KD, leads 61–183	0.692	0.694	0.722	0.719	0.719	
	FID, leads 61–183	0.418	0.293	0.272	0.266	0.266	
ICs		25	50	100	150	175	
	KD, leads 61–183	$0.574 \pm .039$	$0.656 \pm .033$	$0.669 \pm .018$	$0.660 \pm .009$	0.662	

D.4 Error bars

We calibrated the clustered jackknife against an exact null: $2n$ ICs drawn from one 350-IC ground-truth pool and split at random, so the expected KD is exactly 0. We used 2,000 draws per cell, $n = 10$–70 per arm, three windows and three estimators. The jackknife standard error is 1.15–1.35 times the resampling standard deviation (median 1.26), independent of n. The null distribution is skewed (skewness 1.6–2.8), so ± 1.96 standard errors cover 0.90–0.97 despite being too wide. Calibration of the paired-design standard error at $n \approx 175$ is not established. We therefore never read ± 2 standard errors near the null as a test.

D.5 Pooling over lead time and regime

A score pools every frame of a lead window (1–20, 21–60, 61 onward) into one distribution. It is therefore not linear in lead, and scores from different strides or windows are not comparable.

Pooling across within-family regimes θ gives the *marginal* score of Section 3.2, which the disturbance, trained-simulator and physical-alignment scores use. The regime-averaged score $\mathbb{E}_\theta\big[D(\cdot)\big]$ of the model ladder (Table 45) is a different estimand. A per-regime score does not decompose a pooled one: for Walrus-FT on RB at leads 61 onward, the pooled score is 2.36×10^{-3}, while every per-Rayleigh group lies in 3.10–6.12 ($\times 10^{-3}$). On the 11 RB models of Appendix G.1, the two estimands rank models almost identically. Spearman correlation between paired marginal and regime-conditional scores is $+0.99$ over 33 (model, window) points, and none of the 61 resolved KD-versus-error inversions of Appendix G changes direction between them.

Whether conditioning on regime changes what a space resolves depends on the noise scale. On the nine-model RB rollout ladder (paired KD, test split, nine lead windows), we split each cell from 175 ICs to 35 by Rayleigh number and count the cells in which the top two models differ by more than twice the noise scale. Against their jackknife standard error, the fraction of such cells falls from 0.78 to 0.33 for PDE-KD and from 0.66 to 0.49 for ImageNet, and it rises from 0.11 to 0.67 and 0.64 for the two spectra (Spectrum and Spectrum-native). On this scale the spaces move in opposite directions, which a change in sample size alone would not produce. Against the seed-to-seed spread of the split-half floor, which is conservative at this step though not at every IC count, no space falls and two rise.

D.6 Hypothesis tests, calibration and exact floors

This subsection states the permutation tests behind every rejection, checks that the shared test code is exact, discloses per-cell false-positive rates and gives exact floors for rank correlations.

Tests. Every KD test is one-sided. The null H_0 states that the two arms' per-IC feature samples are exchangeable; the alternative is that KD is larger than under relabeling. With disjoint arms, we use a two-sample IC-label permutation test: pool the ICs of both arms, reassign them to arms of the original sizes B times, recompute Eq. (10), and set $p = (1 + \#\{\mathrm{KD}^* \geq \mathrm{KD}\})/(B+1)$ (Phipson & Smyth, 2010). When the units are independent ICs, the permutation null is exact under H_0. Some units share an IC or are consecutive in time: the 90 turbulent-radiative-layer trajectories are 10 ICs $\times$ 9 cooling times, and HRRR's units are consecutive days. There, exactness rests instead on each unit's clean and disturbed versions being exchangeable under H_0, the same assumption the paired design makes. The relabelings of these two-sample tests are sampled, not enumerated. With the "+1" in numerator and denominator, this Monte Carlo test has size at most $\lfloor \alpha(B+1) \rfloor/(B+1) \leq \alpha$ at any B; ties, counted as $\geq$, only make it more conservative. Permutation replicates are recomputed from persisted per-IC block sums, which reproduce the directly computed values to floating-point rounding (about 10^{-15} relative; ties are counted within 10^{-12} relative). We use three resolutions. The power sweeps of Section 4.1 use $B = 999$ per replicate. Single full-pool tests on RB (Table 28) and in every column of Table 3 use $B = 2{,}000$; Appendix D.7 gives the counts at $B = 20{,}000$. The clip-spectral suite, the severe spectral ladders and the re-runs of single tests at higher resolution (Appendices C.2, D.7 and I.2) use $B = 20{,}000$. The IC-conditional test of Section 4.2 instead enumerates all $\binom{10}{5} = 252$ relabelings of its ten trajectories, so its $p = \#\{\mathrm{KD}^* \geq \mathrm{KD}\}/252$ has no Monte Carlo error (floor $2/252$, since KD is symmetric in the two arms).

Paired and conditional designs. For paired arms (the same U ICs, clean and disturbed), the null flips each IC's two arm labels with an independent fair coin. The p-value is Monte Carlo over B relabelings: 2,000 for the disturbance tests and 20,000 in the symmetry suite. For a cell with $U \leq 12$ ICs, the relabelings are drawn from its 2^{U-1} enumerated distinct labelings; otherwise they are coin flips. The *paired design* applies this to the whole paired set; the *conditional design* applies it within each coefficient cell and tests the mean of the per-cell KD over cells. Both rest on the assumption that an IC's two arms are exchangeable under the null, which the design assumes and does not test. The clean-versus-clean draws below cannot test it either, because their two halves are exchangeable by construction.

Exactness. Every KD and energy-distance test behind a reported rejection count, power or n_{80} is a randomization test whose null relabels the units exactly as the design assigns them to arms. Given the exchangeability stated above, its size is therefore at most α. The single-trajectory comparisons of Appendix F relabel autocorrelated frames and are not exact.

Code check. We check that the shared test code (the disturbance suites and the symmetry suite's relabeling test) realizes this exactness. For each of the 11 (system, design) pairs whose tests the paper reports or discloses, 4,000 fresh clean-versus-clean draws of the design's own randomization, each a test with $B = 999$, should reject at rate 0.05 (Table 19). Ten pairs pass at the Bonferroni threshold $0.05/11$. Shear flow's disjoint design rejects in 241 of its first 4,000 draws ($p = 0.004$); with correct code, at least one of the 11 primary checks fails by chance with probability 0.049. A replication on fresh draws, which the rule did not pre-specify, rejects in 203 ($p = 0.83$), and the pooled count is 444 of 8,000 ($p = 0.026$). The pooled rate, 0.056 (0.051–0.061), allows an excess of up to about one percentage point; we accept the design and disclose this excess. The check runs KD; exactness does not depend on the kernel, so the energy-distance tests share the checked code. Two tests have their own permutation code and are outside the check. The IC-conditional test of Section 4.2 enumerates its 252 relabelings, and its clean-versus-clean rejection rate is $12/252$ in all 11 of its runs, as exact enumeration requires. The reference-versus-reference tests of Appendix I.1 are not checked; the within-day hour relabeling of the HRRR diurnal axis additionally assumes that hours within a day are exchangeable.

Per-cell false-positive rates. For every reported (space, estimator, design) cell we disclose the clean-versus-clean rejection rate over repeated null draws, with its Clopper–Pearson 95% interval and the number R of draws (Tables 20 and 21; Table 61 for the out-of-distribution suite; the Random init cells on HRRR below). Each draw is a test of the cell's design with $B = 999$ on a new IC split. R is 40, or 200 where the 40-draw run was extended: for the clip spaces, for Spectrum, Spectrum-native and PCA-white on RB, and on the other simulated systems for every space whose disjoint 40-draw rate reached 0.075. A 200-draw run contains its 40-draw run. The spectral identity rung (severity 0) is a second measurement per space, pooled over the swept pool sizes at which the exact test can reach α ($n \geq 12$ on RB). A cell would be flagged only if its rate were significant after Holm correction over all 441 measured cells (those not tabulated are in the source data) and the excess reproduced on at least 400 fresh, non-nested draws. No cell is flagged: the smallest Holm-corrected p is 0.31 (Poseidon on HRRR, 8 of 40). Individual intervals do exclude 0.05: 18 lie above it and 3 below. Those above include PDE-KD on HRRR (6 of 40, 0.057–0.298) and on the turbulent radiative layer (17 of 200, 0.050–0.133), Poseidon on HRRR (8 of 40) and on shear flow (19 of 200, 0.058–0.144), and two RB cells, PCA-white KD in the disjoint design and Spectrum-native KD in the conditional design (17 of 200 each, 0.050–0.133). Across the 441 cells, 36 have a one-sided $p \leq 0.05$, against about 22 expected; the cells are correlated, because the spaces of a system share their IC splits. The flag rule has little power. At this family size a flag needs at least 10 of 40 or 24 of 200 rejections, so a true rate of 0.10 would be flagged with probability 0.005 at $R = 40$ and about 0.2 at $R = 200$. The absence of a flag is therefore not evidence of calibration; the intervals are that evidence. The Random init cells on HRRR were added to Table 3 after this family was fixed, so they lie outside it and outside Tables 21 and 61. Their KD tests ($R = 40$, $B = 999$) reject in 2, 3 and 2 of 40 disjoint draws for the three seeds, in 4, 3 and 2 of 40 paired draws, and in 3, 1 and 0 of 40 conditional draws. All nine counts are below the flag threshold of 10 of 40, and every Clopper–Pearson interval contains 0.05 (the widest is 0.028–0.237, at 4 of 40). On RB, the identity rungs of the frame spaces reject at 0.046–0.058 (KD) and 0.041–0.058 (energy distance), and those of the clip spaces at up to 0.061. These identity-rung sweeps were run before a correction to tie handling that affects only the smallest pools; the pooled rates exclude $n \leq 8$, and the residual effect at $n = 12$ is about 0.001. Table 61 instead reports the identity rung at the full pool ($R = 200$). Every space, PCA-white included, enters every comparison; the rates are disclosed, not used to exclude cells.

Exact rank-correlation floors. Model ladders have few models, so Spearman p-values must be exact. SciPy's asymptotic p is exactly 0 for any perfect ordering, whereas at $n = 4$ the exact two-sided p is $1/12$. Table 22 lists the smallest $|\rho|$ that reaches two-sided $p \leq 0.05$ under exact permutation. With five models only a perfect ordering qualifies ($p = 2/120$). With four or fewer, nothing does. Ties move the floor.

Power and n_{80}. The replicate design, the n grid, the logistic fit and censoring are given in Appendix E.1. The full per-level power tables (Tables 29–30) show where each curve crosses 0.8. Three properties matter when reading these quantities.

Table 19: **Code check of exactness.** For each (system, design) pair on which a reported result rests: rejections k in S clean-versus-clean draws of the design's own label randomization (PDE-KD features, $B = 999$ per test) and the two-sided exact binomial p against 0.05. A pair passes if $p > 0.05/11$. TRL2D: turbulent radiative layer. –: no reported result uses that design on that system.

	disjoint	paired	conditional
system	k/S p	k/S p	k/S p
RB	210/4000 .468	204/4000 .772	206/4000 .663
Shear flow	241/4000 .004	–	–
fresh draws (not pre-specified)	203/4000 .828		
pooled	444/8000 .026		
Euler (periodic)	203/4000 .828	–	178/4000 .119
Gray–Scott	213/4000 .345	–	–
FSI	190/4000 .491	–	–
HRRR	226/4000 .064	–	–
ERA5	206/4000 .663	–	–
TRL2D	198/4000 .913	–	–

Table 20: **Clean-versus-clean false-positive rates, RB.** Cell: rejections k/R at $\alpha = 0.05$ and the Clopper–Pearson 95% interval. Design columns: repeated clean-versus-clean tests of that design. Identity rung: the spectral suite's severity-0 rung, disjoint test, pooled over $n = 12$–175. Last block (clip spaces): design columns on the clip suite's windows, identity rung from the clip-spectral suite; the other JEPA checkpoints and readouts (Appendix E.6) have only the identity rung. –: not measured.

	disjoint	paired	conditional	identity rung
KD (cubic kernel)				
PDE-KD	2/40 .006–.169	2/40 .006–.169	3/40 .016–.204	88/1800 .039–.060
ImageNet	4/40 .028–.237	1/40 .001–.132	4/40 .028–.237	97/1800 .044–.065
Poseidon	2/40 .006–.169	3/40 .016–.204	1/40 .001–.132	87/1800 .039–.059
Poseidon (per-field)	2/40 .006–.169	2/40 .006–.169	0/40 .000–.088	–
Random init (seed 0)	1/40 .001–.132	3/40 .016–.204	0/40 .000–.088	83/1800 .037–.057
Random init (seed 1)	1/40 .001–.132	4/40 .028–.237	0/40 .000–.088	86/1800 .038–.059
Random init (seed 2)	1/40 .001–.132	3/40 .016–.204	0/40 .000–.088	84/1800 .037–.057
Pixels	0/40 .000–.088	2/40 .006–.169	0/40 .000–.088	87/1800 .039–.059
Spectrum	12/200 .031–.102	9/200 .021–.084	12/200 .031–.102	105/1800 .048–.070
Spectrum-native	11/200 .028–.096	12/200 .031–.102	17/200 .050–.133	101/1800 .046–.068
PCA-white	17/200 .050–.133	5/200 .008–.057	8/200 .017–.077	101/1800 .046–.068
PCA-raw	2/40 .006–.169	2/40 .006–.169	2/40 .006–.169	84/1800 .037–.057
Energy distance				
PDE-KD	3/40 .016–.204	2/40 .006–.169	2/40 .006–.169	80/1800 .035–.055
ImageNet	4/40 .028–.237	1/40 .001–.132	4/40 .028–.237	99/1800 .045–.067
Poseidon	3/40 .016–.204	2/40 .006–.169	0/40 .000–.088	93/1800 .042–.063
Poseidon (per-field)	2/40 .006–.169	0/40 .000–.088	1/40 .001–.132	–
Random init (seed 0)	3/40 .016–.204	0/40 .000–.088	0/40 .000–.088	77/1800 .034–.053
Random init (seed 1)	2/40 .006–.169	3/40 .016–.204	0/40 .000–.088	81/1800 .036–.056
Random init (seed 2)	2/40 .006–.169	0/40 .000–.088	0/40 .000–.088	74/1800 .032–.051
Pixels	1/40 .001–.132	2/40 .006–.169	0/40 .000–.088	97/1800 .044–.065
Spectrum	14/200 .039–.115	11/200 .028–.096	7/200 .014–.071	89/1800 .040–.060
Spectrum-native	15/200 .043–.121	6/200 .011–.064	9/200 .021–.084	95/1800 .043–.064
PCA-white	11/200 .028–.096	5/200 .008–.057	7/200 .014–.071	89/1800 .040–.060
PCA-raw	1/40 .001–.132	2/40 .006–.169	0/40 .000–.088	104/1800 .047–.070
Clip spaces, KD (RB windows)				
PDE-KD (3D)	16/200 .046–.127	6/200 .011–.064	14/200 .039–.115	–
PDE-KD (3D, torch port)	–	–	–	85/1800 .038–.058
JEPA	9/200 .021–.084	9/200 .021–.084	11/200 .028–.096	109/1800 .050–.073
JEPA (ep. 3)	–	–	–	109/1800 .050–.073
JEPA (shipped)	–	–	–	86/1800 .038–.059
JEPA (post-MLP tap)	–	–	–	95/1800 .043–.064
JEPA (pre-MLP tap)	–	–	–	90/1800 .040–.061

PDE-KD (3D, torch port): the same checkpoint via our PyTorch port (the clip spectral suite's bank).

Table 21: **Clean-versus-clean false-positive rates, other systems.** Format as Table 20; the identity rung is pooled over the swept n at which the exact test can reach α. On Euler and HRRR, Spectrum-native and Spectrum are one measurement. Energy-distance rows: the energy-distance tests of the out-of-distribution suite. Last block: Euler's conditional design, which backs the symmetry suite's relabeling test. TRL2D: turbulent radiative layer.

	Shear flow	Euler	Gray–Scott	FSI	HRRR	ERA5	TRL2D
Clean-vs-clean draws, KD, disjoint design							
PDE-KD	2/40 .006–.169	0/40 .000–.088	13/200 .035–.109	2/40 .006–.169	6/40 .057–.298	1/40 .001–.132	17/200 .050–.133
ImageNet	12/200 .031–.102	2/40 .006–.169	8/200 .017–.077	2/40 .006–.169	3/40 .016–.204	1/40 .001–.132	18/200 .054–.139
Poseidon	19/200 .058–.144	2/40 .006–.169	1/40 .001–.132	4/40 .028–.237	8/40 .091–.356	2/40 .006–.169	18/200 .054–.139
Random init (seed 0)	9/200 .021–.084	2/40 .006–.169	13/200 .035–.109	–	–	–	–
Random init (seed 1)	10/200 .024–.090	2/40 .006–.169	12/200 .031–.102	–	–	–	–
Random init (seed 2)	10/200 .024–.090	12/200 .031–.102	11/200 .028–.096	–	–	–	–
Pixels	1/40 .001–.132	1/40 .001–.132	10/200 .024–.090	3/40 .016–.204	5/40 .042–.268	0/40 .000–.088	16/200 .046–.127
Spectrum	11/200 .028–.096	2/40 .006–.169	2/40 .006–.169	1/40 .001–.132	3/40 .016–.204	1/40 .001–.132	15/200 .043–.121
Spectrum-native	11/200 .028–.096	2/40 .006–.169	2/40 .006–.169	1/40 .001–.132	3/40 .016–.204	1/40 .001–.132	14/200 .039–.115
PCA-white	1/40 .001–.132	2/40 .006–.169	2/40 .006–.169	3/40 .016–.204	4/40 .028–.237	4/40 .028–.237	–
PCA-raw	–	–	–	2/40 .006–.169	2/40 .006–.169	1/40 .001–.132	–
Spectral identity rung, KD, disjoint test (sweep, pooled over n)							
PDE-KD	74/1600 .036–.058	86/1800 .038–.059	81/1600 .040–.063	66/1200 .043–.069	65/1400 .036–.059	39/800 .035–.066	72/1400 .040–.064
ImageNet	82/1600 .041–.063	93/1800 .042–.063	91/1600 .046–.069	63/1200 .041–.067	55/1400 .030–.051	41/800 .037–.069	68/1400 .038–.061
Poseidon	82/1600 .041–.063	83/1800 .037–.057	80/1600 .040–.062	70/1200 .046–.073	73/1400 .041–.065	42/800 .038–.070	67/1400 .037–.060
Random init (seed 0)	77/1600 .038–.060	69/1800 .030–.048	93/1600 .047–.071	–	–	–	–
Random init (seed 1)	80/1600 .040–.062	76/1800 .033–.053	82/1600 .041–.063	–	–	–	–
Random init (seed 2)	79/1600 .039–.061	76/1800 .033–.053	103/1600 .053–.078	–	–	–	–
Pixels	82/1600 .041–.063	85/1800 .038–.058	85/1600 .043–.065	70/1200 .046–.073	71/1400 .040–.064	40/800 .036–.067	75/1400 .042–.067
Spectrum	–	–	–	56/1200 .035–.060	66/1400 .037–.060	42/800 .038–.070	–
Spectrum-native	91/1600 .046–.069	97/1800 .044–.065	94/1600 .048–.071	62/1200 .040–.066	66/1400 .037–.060	41/800 .037–.069	69/1400 .039–.062
PCA-white	88/1600 .044–.067	88/1800 .039–.060	87/1600 .044–.067	65/1200 .042–.069	64/1400 .035–.058	41/800 .037–.069	–
PCA-raw	–	–	–	64/1200 .041–.068	85/1400 .049–.075	37/800 .033–.063	–
Clean-vs-clean draws, energy distance, disjoint design (out-of-distribution energy-distance arm)							
PDE-KD	–	–	–	1/40 .001–.132	6/40 .057–.298	1/40 .001–.132	–
ImageNet	–	–	–	2/40 .006–.169	5/40 .042–.268	1/40 .001–.132	–
Poseidon	–	–	–	3/40 .016–.204	8/40 .091–.356	1/40 .001–.132	–
Pixels	–	–	–	1/40 .001–.132	7/40 .073–.328	0/40 .000–.088	–
Spectrum	–	–	–	1/40 .001–.132	3/40 .016–.204	2/40 .006–.169	–
Clean-vs-clean draws, KD, conditional design (Euler: the symmetry suite's relabeling test)							
PDE-KD		4/40 .028–.237					
ImageNet		4/40 .028–.237					
Poseidon		3/40 .016–.204					
Random init (seed 0)		1/40 .001–.132					
Random init (seed 1)		3/40 .016–.204					
Random init (seed 2)		7/200 .014–.071					
Pixels		1/40 .001–.132					
Spectrum		1/40 .001–.132					
Spectrum-native		1/40 .001–.132					
PCA-white		3/40 .016–.204					

Table 22: **Smallest $|\rho|$ reaching exact two-sided $p \leq 0.05$** over n untied points (models or levels).

n	3	4	5	6	7	8	9
$\lvert\rho\rvert$	none	none	1.000	0.886	0.786	0.738	0.700

- In our RB suites, every L5 cell with a finite n_{80} also rejects in the full-pool test (an empirical regularity, not a law), whereas a censored "175+" implies nothing. The converse does not hold: on the severe spectral ladders of Appendix E.2, 15 of 120 injection and 11 of 120 removal (space, band, level) cells reject in their single test while needing more ICs than the pool for 80% power. Spectrum's low-band injection at L3, for example, rejects at $p = 0.0084$ but reaches power 0.415 at 175 ICs.
- At n equal to the pool, power is a property of this pool rather than of the population (Appendix E.3).
- Rung counts are descriptive tallies of single tests at $\alpha = 0.05$ with no family-wise correction; at a true null, about one rung in twenty would reject.

D.7 Robustness of rejection verdicts

This subsection checks four analysis choices behind the rejection verdicts: the counting rule, the arm design, the permutation resolution and the kernel.

Counting rule. A descriptive alternative counts a rung as cleared when its KD exceeds two jackknife standard errors. The rungs it clears are a strict subset of those the permutation test rejects, its denominators are space-dependent, and it reorders spaces. On the RB spectral suite, PDE-KD / ImageNet / Poseidon clear 30 / 22 / 15 rungs under it, against 41 / 26 / 29 rejections under the permutation test. All disturbance rung counts and n_{80} values therefore use the permutation test; the descriptive analyses of Table 23 keep the rules listed there.

Arm design. We score disturbances in the disjoint design, with a different half of the ICs in each arm, because the paired design rejects the null for any change and grades none. Pairing each disturbed IC with its own clean version makes every space reject essentially every rung. On the RB spectral suite, the disjoint design leaves 14 (PDE-KD), 29 (ImageNet) and 26 (Poseidon) rungs not rejected, against 0 in the paired design. At one rung (PDE-KD, high-band removal L1), the paired permutation threshold is $936\times$ smaller than the disjoint one (1.26×10^{-6} versus 1.18×10^{-3}), under a pairwise-exchangeability assumption that this suite cannot check. Pairing saturates pointwise metrics too. A paired sign-flip test on per-IC vRMSE and on the spectral error rejects all 30 (metric, family) configurations at L5 at its resolution limit ($p = 1/2001$), with $n_{80} < 8$ paired ICs everywhere (perceptual effects measured against an analytic zero). Every per-IC effect is positive, and with all effects of one sign, five paired ICs already reach $p = 1/32$. The paired design is therefore uninformative for ranking representations or metrics by disturbance power, and no reported disturbance rung count, power or n_{80} outside this paragraph uses it. It remains wherever two arms share their ICs by construction: model-versus-reference scores and the symmetry suite (Table 23), and the paired coefficient contrasts of Appendix F.

Permutation resolution. Re-running the 120 canonical spectral injection tests at $B = 20{,}000$ changes one verdict: PDE-KD high-k injection L3, $p = 0.0515 \to 0.0491$. The single-test spectral count for PDE-KD would move from 41 to 42 of 55 (Table 28 uses $B = 2{,}000$); its L5 coverage in Table 3 is unchanged. No other space changes. Re-running the tests of Table 3 at $B = 20{,}000$ (every space on all five systems; the Random init control on RB and HRRR) changes two RB perceptual entries. Pixels moves from 2 to 1 (Gaussian noise L2, $p = 0.046 \to 0.052$), and Random init from 3.0 to 2.3 (Gaussian blur L2 in two of the three seeds, $p = 0.048 \to 0.053$ and $0.049 \to 0.052$). Every other re-run entry is unchanged. The Random init control on shear flow, Euler and Gray–Scott was not re-run at $B = 20{,}000$.

Kernel. Replacing the cubic kernel by the energy-distance kernel changes 25 of 1,600 rejection decisions on the RB suites (23 energy-only, 2 KD-only; all but one in the disjoint design). On the

Table 23: **Estimator, arm design and inference per result.** IC-level KD is Eq. (10); frame-level KD drops only identical frames.

Result	Estimator	Arms	Inference
Controlled disturbances (Tables 2 and 3, Appendices E and I.2)	IC-level KD, pooled over regimes	disjoint clean vs disturbed ICs	IC-permutation test; power over 200 replicates; n_{80}
IC-conditional laws (Section 4.2, Figures 3 and 10)	IC-level KD, per coefficient pair and window	five vs five trajectories, disjoint noise seeds	exact permutation test (252 relabelings); power over the 252 seed splits
IC-conditional exceedance probe (Tables 43 and 44, Appendix F.2)	frame-level KD	one 20-frame trajectory window per arm	95th percentile of the group's same-cell null
Monotonicity in severity (Table 32, Appendix E.5)	IC-level KD	disjoint clean vs disturbed ICs	Spearman ρ over five levels per ladder (descriptive; mean over 15 ladders, 13 for Spectrum)
Physical alignment (Table 47, Appendices G.3 and G.4)	IC-level KD, paired, pooled over regimes	model vs reference on the same ICs	Spearman ρ over models (family entries: exact test at nine RB models, descriptive on Euler; per-functional cells against the exact cutoff of Table 22)
Trained neural simulators (Table 4, Appendix G.5)	IC-level KD, paired, pooled over regimes; one rollout per IC	model vs reference on the same ICs	count of within-family pairs whose point estimates follow the expected order (descriptive); parentheses: KD $-$ floor $\leq$ 2 s.d. of the floor over split seeds
Model ladder (Table 45)	IC-level KD, paired, per (Ra, Pr), averaged	same ICs	descriptive; "resolved" = mean $-$ 1 s.d. across the 35 regimes > 0
CRPS against climatology (Figure 12, Appendix G.7)	IC-level KD, paired, pooled	same ICs	descriptive; "resolved" = KD $-$ floor $>$ 2 s.d. of the floor over split seeds
Inversions (Appendix G.7)	paired contrasts of regime-conditional KD	same ICs	both contrasts beyond ± 2 SE (descriptive)
Symmetry suite (Figure 9)	IC-level KD (and pixel energy distance, CRPS), paired, regime-conditional	transformed vs untransformed	one-sided mean $>$ 2 SE across cells (descriptive)
Conditioning-frame corruption (Appendix E.8)	IC-level KD, paired, pooled	disturbed-conditioning rollout vs reference on the same ICs	descriptive ratios to the clean rollout (not comparable across models)
Euler compression ladder (Appendix G.6)	IC-level KD, paired, per (adiabatic-index cell, lead window)	rollout vs reference, 50 or 5 ICs per cell	count of designed orderings recovered (exact binomial against chance); sign test over paired draws
OOD axes (Table 54)	IC-level KD	reference cell vs reference cell	relabeling permutation (exhaustive on FSI and ERA5, sampled on HRRR; hour labels swapped within each day)
Weather ranking (Appendix I.4)	IC-level KD, paired, per lead	forecast vs ERA5 on the same initializations	Spearman vs published scores
Figure 1	frame-level KD	one trajectory per arm, 139 frames	descriptive

55 spectral rungs, Spearman correlation between the two statistics over the paired ladders is $+0.997$ for PDE-KD. The cubic kernel is not characteristic: its MMD compares moments only up to order three. Out of distribution, the two estimators disagree on FSI blur, where only the energy distance rejects in the PDE-KD space, in a single uncorrected test (Appendix I.2). One rejection does not establish that the cubic kernel misses a difference the features carry. A pre-registered rule tests this reading on the 24 out-of-distribution (dataset, space, family) cells with high-wavenumber removal. It counts 2 cells that support it (the energy distance rejects at L5 and cubic-kernel KD does not), 16 that complicate it (cubic-kernel KD rejects at L5), and 6 in which neither estimator rejects at L5.

D.8 Which estimator and test each result uses

Every power or rejection claim for controlled disturbances (Section 4.1) uses a permutation test (Appendix D.6), as does the IC-conditional test of Section 4.2. The IC-conditional exceedance probe instead thresholds at a percentile of its empirical same-cell null (Appendix F). Five analyses use descriptive rules instead: the symmetry suite, the model ladder, the CRPS comparison, the inversion counts and the trained-simulator tables.

Table 23 lists, for the principal results, the estimator variant, the arm design and the inference rule. Scores computed under different rows are different estimands and are not interchangeable. For

example, Walrus-FT's pooled paired KD on RB at leads 61 onward is 2.36×10^{-3}, its regime-conditional score in Table 45 is 12.41×10^{-3}, and its score within every Rayleigh group lies in 3.10–6.12 ($\times 10^{-3}$; Appendix D.5).

E CONTROLLED PERTURBATIONS: DEFINITIONS AND FULL RESULTS

This appendix defines the controlled perturbations of Section 4.1 and reports their full results. It covers the protocol and the operators, the complete RB results, three further physical systems, monotonicity in severity, the temporal suite, invariance to symmetries of the equations, corruption of conditioning frames, and the raw single-test scores.

E.1 PROTOCOL

This subsection fixes the design shared by the frame-level disturbance experiments on RB; Appendix E.4 adapts it to three further systems and Appendix E.6 to clips.

Pools. All frame disturbances are applied to RB reference frames 61 onward. Perceptual power sweeps draw from a fixed 175-IC subset of the 350 pooled validation and test ICs, and perceptual single tests use all 350 (175 per arm). Spectral experiments use all 175 ICs of the RB test split. Because the two suites use different pools, their results are not paired. Within each condition, every feature space sees the same source frames, disturbance and partition. Spectral disturbances are applied on the native field before representation-specific staging, so Spectrum-native retains the 512×128 RB grid. Perceptual disturbances act after canonical resizing and therefore have no Spectrum-native row.

Power sweeps and the identity rung. At each total IC count n of the grid 8, 12, 16, 24, 32, 48, 64, 96, 128 and 175, 200 replicates split the sampled ICs approximately equally into clean and disturbed arms (87 versus 88 at $n = 175$). An IC appears in only one arm, coefficients are pooled, and a one-sided IC-label permutation test with $B = 999$ permutations rejects at $\alpha = 0.05$. Empirical power is the rejection fraction, with maximum Monte Carlo standard error $0.5/\sqrt{200} = 0.0354$. Spectral severity zero is an identity control. At the full 175-IC pool it rejects at 0.035–0.070 across spaces (200 replicates per space, 600 pooled for the Random init seed mean; Table 30), and pooled over the swept pool sizes $n \geq 12$ at 0.046–0.058 for the frame spaces' KD (Table 20). The perceptual suite has no identity rung; the clean-versus-clean false-positive rates of its disjoint design are measured separately and disclosed in Table 20 (Appendix D.6).

Families and levels. The perceptual suite comprises occlusion, Gaussian noise, salt-and-pepper and Gaussian blur. The spectral suite comprises low-, mid- and high-k band injection and removal, small- and large-scale tilt, truncation, and independent and shared phase scrambling. The temporal suite comprises seven ladders that act on clips (Appendices E.2 and E.6). Each family has five ordinal severity levels (Table 25); levels are not equal physical doses across families.

Sample efficiency. At L5 we fit

$$\operatorname{logit} \pi(n) = a + b \log n, \qquad b \geq 0, \tag{11}$$

and solve for the IC count n_{80} at which the fitted power reaches 0.80. Crossings outside the sampled range are reported as censored bounds, 8− or 175+. A fitted n_{80} has a parametric-bootstrap 95% interval (Appendix E.3).

Reading Table 2. Its cells use the format and colors defined in the caption of Table 26. Its perceptual columns omit salt-and-pepper, which is reported in Table 29. Its spectral columns are independent phase scrambling and band injection and removal at low and high k, one column per band. Injection and removal there use the severe band ladder (amplitude gain g; Table 25), whose injection rungs lie at or above 1σ of the IC-to-IC spread of band energy. The injection rungs of the canonical ladder lie below that spread; its results are in Appendix E.3. The medium band is omitted from Table 2 and reported in Tables 26 and 27. The Spectrum row uses canonical-resize features for the perceptual columns and native-grid (Spectrum-native) features for every spectral column, phase scrambling included. Its two temporal columns are described in Appendix E.6.

E.2 DEFINITIONS AND SEVERITY LADDERS

This subsection defines every operator (Table 24) and its five severity levels (Table 25), and relates the spectral severities to the natural variability of band energy.

Table 24: **Disturbance operators.** x is a field, ξ an independent standard-normal field, $\hat{x}(k)$ its Fourier coefficients, and b a radial band. For the temporal operators, s is the clip's start frame and t the frame index within the clip. L1–L5 values are in Table 25.

Suite	Operator	Definition (emulated rollout defect)
Perceptual	Occlusion	one rectangle of sides $pH \times pW$ set to 0 (the channel mean after z-scoring), shared by all slots (missing region)
	Gaussian noise	$(1-\alpha)x + \alpha\xi$ (uncorrelated noise)
	Salt and pepper	each pixel with probability p set to the slot minimum or maximum (impulsive artifacts)
	Gaussian blur	separable Gaussian, σ pixels, reflect padding, truncated at 2.5σ (over-smoothing)
Spectral	Phase (indep./shared)	$\hat{x}(k) \mapsto \hat{x}(k)\, e^{i\alpha\eta_k}$, $\eta_k \sim U(-\pi, \pi)$ drawn per field or shared across fields; $\lvert\hat{x}\rvert$ unchanged (wrong spatial arrangement at correct power)
	Truncation	zero all $\lvert k\rvert > \rho\, k_{\max}$ (lost resolution)
	Tilt (small/large)	$\hat{x}(k) \mapsto \lambda\, (\lvert k\rvert / k_p)^{\pm\beta}\, \hat{x}(k)$ for $k \neq 0$, λ restoring the frame's energy (wrong spectral slope)
	Band removal	$\hat{x}(k) \mapsto \sqrt{E}\, \hat{x}(k)$ in band $b \in \{\text{low, mid, high}\}$, $E < 1$ (missing scales)
	Band injection	as removal with $E > 1$ (excess energy at some scales)
Temporal	Local swap	m swaps of adjacent frames (local order errors)
	Global swap	m swaps of arbitrary frame pairs (global order errors)
	Interpolation dropout	in n disjoint triplets, the middle frame replaced by the mean of its neighbors (synthesized frames)
	Interleave	frame t taken from trajectory $t \bmod k$ of the base and its first $k-1$ partners, at frame $s + \lfloor t/k \rfloor$, so each plays at $1/k$ speed (incoherent dynamics)
	Switching	every n frames the clip moves to the next trajectory of a six-member group (base and 5 partners), timestamps aligned (a smaller n is harsher)

Table 25: **Severity levels L1–L5 of each disturbance family.** The severe band ladders are those of Table 2.

Disturbance	Parameter	L1	L2	L3	L4	L5
Gaussian noise	α (mix)	0.05	0.1	0.15	0.2	0.25
Gaussian blur	σ (px)	1	2	3	4	5
Occlusion	p (side fraction)	0.15	0.3	0.45	0.6	0.75
Salt and pepper	p (pixel prob.)	0.02	0.04	0.06	0.08	0.1
Phase (indep./shared)	α	0.1	0.25	0.5	0.75	1
Truncation	ρ	0.5	0.35	0.25	0.18	0.125
Tilt (small / large scales)	$\pm\beta$	0.125	0.25	0.5	0.75	1
Band removal (low/mid/high k)	E	0.8	0.6	0.4	0.2	0
Band injection (low/mid/high k)	E	1.2	1.4	1.6	1.8	2
Band removal, severe (Table 2)	g ($E = g^2$)	0.5	0.25	0.1	0.03	0
Band injection, severe (Table 2)	g ($E = g^2$)	1.5	2	4	8	16
Local swap	m adjacent pairs	4	8	12	16	20
Low-severity local / global swap	m pairs	1	2	3	4	5
Global swap	m pairs	4	8	12	16	20
Interpolation dropout	n triplets	1	2	3	4	5
Interleave	k	2	3	4	5	6
Switching	n frames	1	2	3	4	5

Perceptual and spectral operators. Perceptual operators act on the z-scored canonical tensor (Appendix B.2), occupied slots only. Spectral operators act on the native RB field in physical units, after subtracting buoyancy's horizontal-mean (conductive) profile. They are applied in an orthogonal, unnormalized 2D FFT basis: a real FFT along the periodic x and an FFT along the wall-bounded z, read as uniform and periodic. This basis is exactly invertible, so severity 0 reproduces the clean features up to floating-point batching differences. On RB the maximum absolute feature difference at severity 0 is 8.3×10^{-7} for PDE-KD, 1.3×10^{-5} for ImageNet and 2.7×10^{-5} for Poseidon, against a 10^{-4} tolerance; the multiphysics checks of Appendix E.4 use a 10^{-3} tolerance on the maximum relative feature difference instead. The plain FFT basis matters for the phase operators. In a mixed Fourier–Chebyshev basis, a phase change that preserves amplitude in that basis still changes the plain-FFT band power that the spectral metrics read, by up to 1.27 in relative power error on one velocity band. In the plain FFT basis the same phase change leaves every Fourier amplitude unchanged to 2×10^{-16}. The two phase families reuse one random field η across their five levels,

and the other spectral operators are deterministic. The perceptual and temporal operators draw a new random field at each level.

Band edges. Band edges come from fixed per-field spectral landmarks, the energy-peak wavenumber k_p and the 95%-cumulative-energy wavenumber k_h, measured on $\mathrm{Ra} = 10^{10}$ validation trajectories. They are therefore energy quantiles of one regime, not physical scales of every regime. At $\mathrm{Pr} = 1$, buoyancy's own k_h ranges from 27 at $\mathrm{Ra} = 10^6$ to 328 at 10^{10} (in radians per unit length; 20–404 across all 35 regimes).

Temporal operators. Temporal operators act on clips of consecutive frames. The seven temporal ladders are local swap, global swap, their two low-severity variants, interpolation dropout, interleave and switching (Table 25). Swaps and interpolation dropout act within one clip. The two mixing operators, interleave and switching, take partners from the next five candidate-arm ICs in a fixed cyclic order, at the same window start. That order runs through the arm file by file, with one $(\mathrm{Ra}, \mathrm{Pr})$ regime per file, so partners often come from other regimes. These operators therefore test regime mixing as well as incoherent dynamics. On clips, each perceptual operator uses one realization per clip, shared by all its frames.

Magnitude of the canonical spectral severities. Measured against the IC-to-IC spread of band energy in the data, the strongest injection of the canonical ladder anywhere is 0.99σ (low band 0.23–0.89σ, mid 0.25–0.93, high 0.26–0.99). Not rejecting a high-k injection at these levels is therefore consistent with an effect-size limit rather than blindness.

Severe injection ladder. A more severe ladder in the same basis scales the band's amplitude by g rather than its energy by E (injection $g = 1.5$–16, i.e. 1.0–7.9σ; removal of 75% to 100% of band energy). Table 2 reports this ladder at low and high k, and Tables 26 and 27 give every band and level. At $B = 20{,}000$, its single tests reject the following numbers of its 15 injection and 15 removal levels: PDE-KD 15 and 15, ImageNet 12 and 10, Poseidon 13 and 13, Random init 15.0 and 9.3 (mean over its three seeds), Pixels 13 and 10, Spectrum 12 and 9, Spectrum-native 8 and 9, PCA-white 14 and 14, and PCA-raw 11 and 5. Under the pre-registered injection rule, 12 of the 24 (space, band) cells of the eight pre-registered frame spaces (all but Random init) confirm the effect-size account, and 11 already reject on the canonical ladder. The remaining cell returns the rule's refuting outcome: Spectrum-native's low band never rejects on the severe ladder. Its p-values fall monotonically with severity, so its rejection threshold lies above 7.13σ; this bounds the effect size and is not evidence of blindness. Among the confirming cells, Poseidon rejects the severe high-band injection at its L3, a severity of 3.9σ ($p = 0.00085$).

Severe removal ladder. The severe removal ladder was pre-registered for a different question. The canonical ladder has no rung between 80% of band energy removed and annihilation, and three severe rungs (94%, 99% and 99.9% removed) fall inside that interval. For a cell that does not reject at 80% but rejects at annihilation, the lowest rejecting severe level, reported with its severity in σ, localizes the threshold. For the five non-learned spaces the pre-registration placed six cells in this interval: four first reject at 94% removed (-3.7σ to -3.9σ), and two, the low band of Spectrum and of Spectrum-native, only at annihilation. Three cells had a canonical threshold between 60% and 80% removed. Of these, Pixels' mid band and PCA-raw's low band reject at 75%, and Spectrum-native's high band first rejects at 94%, which places its threshold between 75% and 80%. The three cells that do not reject even at annihilation on the canonical ladder (Pixels' high band, PCA-raw's mid and high bands) reject at no severe level, as predicted. The severe ladder's L5 ($g = 0$) removes the band entirely and is by construction the canonical ladder's L5 ($E = 0$); its n_{80} agrees with the n_{80} rows of Table 30 to within one IC (different replicate draws). Table 2 shows the severe ladder for removal as well as injection by a choice made after both ladders' results were known.

Canonical ladder and clip spaces. The canonical ladder is fully reported in Table 30. On it, no frame space reaches .80 L2 power on all four band columns of Table 2; PDE-KD's L2 power is .74 and .07 for low- and high-k injection and 1.00 and .10 for low- and high-k removal. The two clip spaces of Table 2 are scored on the severe ladder with the protocol of Table 34 (87 versus 88 test ICs). Each rejects all 15 injection and all 15 removal levels at $B = 20{,}000$, and neither rejects the

identity rung ($p = 0.119$ for PDE-KD (3D), 0.635 for JEPA). Their per-level values are the two clip columns of Tables 26 and 27.

Table 26: **Severe band injection ladder** (amplitude gain $g = 1.5, 2, 4, 8, 16$; band energy g^2), RB: n_{80} at each level | power at $n = 175$, and the number of the 15 levels whose single test rejects at $B = 20{,}000$. Clip spaces: protocol of Table 34. Random init: the untrained PDE-KD network, mean over three initialization seeds (Appendix C.2). Cell format and colors of every n_{80} | power table in the appendix and of Table 2: dark green marks $n_{80} < 8$ or power $\geq .80$ (power also bold), dark red n_{80} above the sampled range; power is never red; censored values print as 8− and 175+.

		learned / untrained control								non-learned										clip spaces			
band	level	PDE-KD		ImageNet		Poseidon		Random init		Pixels		Spectrum		Spectrum-native		PCA-white		PCA-raw		PDE-KD (3D)		JEPA	
low-k	L1	26	**1.00**	175+	.06	32	**1.00**	55	**1.00**	78	**1.00**	175+	.07	175+	.07	8−	**1.00**	175+	**.83**	32	**1.00**	65	**1.00**
	L2	10	**1.00**	175+	.23	11	**1.00**	19	**1.00**	27	**1.00**	175+	.09	175+	.07	8−	**1.00**	64	**1.00**	13	**1.00**	21	**1.00**
	L3	8−	**1.00**	44	**1.00**	8−	**1.00**	8−	**1.00**	11	**1.00**	175+	.42	175+	.09	8−	**1.00**	21	**1.00**	8−	**1.00**	8	**1.00**
	L4	8−	**1.00**	13	**1.00**	8−	**1.00**	8−	**1.00**	10	**1.00**	110	**1.00**	175+	.15	8−	**1.00**	13	**1.00**	8−	**1.00**	8−	**1.00**
	L5	8−	**1.00**	8−	**1.00**	8−	**1.00**	8−	**1.00**	10	**1.00**	44	**1.00**	175+	.37	8−	**1.00**	11	**1.00**	8−	**1.00**	8−	**1.00**
mid-k	L1	37	**1.00**	175+	.15	102	**1.00**	74	**.99**	175+	.55	175+	.13	175+	.11	8−	**1.00**	175+	.12	24	**1.00**	62	**1.00**
	L2	12	**1.00**	175+	**1.00**	22	**1.00**	19	**1.00**	62	**1.00**	175+	.24	175+	.17	8−	**1.00**	175+	.45	10	**1.00**	18	**1.00**
	L3	8−	**1.00**	33	**1.00**	8−	**1.00**	8−	**1.00**	11	**1.00**	107	**.98**	175+	**.84**	8−	**1.00**	29	**1.00**	8−	**1.00**	8−	**1.00**
	L4	8−	**1.00**	15	**1.00**	8−	**1.00**	8−	**1.00**	8−	**1.00**	33	**1.00**	73	**1.00**	8−	**1.00**	9	**1.00**	8−	**1.00**	8−	**1.00**
	L5	8−	**1.00**	9	**1.00**	8−	**1.00**	8−	**1.00**	8−	**1.00**	18	**1.00**	38	**1.00**	8−	**1.00**	8−	**1.00**	8−	**1.00**	8−	**1.00**
high-k	L1	175+	.36	175+	.34	175+	.05	175+	.19	175+	.09	175+	.22	175+	.20	175+	.28	175+	.06	175+	.66	124	**1.00**
	L2	90	**1.00**	175+	**.94**	175+	.08	175+	.68	175+	.28	175+	.75	175+	.60	110	**1.00**	175+	.06	56	**1.00**	31	**1.00**
	L3	14	**1.00**	38	**1.00**	130	**1.00**	35	**1.00**	35	**1.00**	48	**1.00**	62	**1.00**	12	**1.00**	175+	.07	9	**1.00**	10	**1.00**
	L4	8−	**1.00**	21	**1.00**	17	**1.00**	11	**1.00**	8	**1.00**	22	**1.00**	28	**1.00**	8−	**1.00**	175+	.29	8−	**1.00**	8−	**1.00**
	L5	8−	**1.00**	13	**1.00**	8−	**1.00**	8−	**1.00**	8−	**1.00**	14	**1.00**	18	**1.00**	8−	**1.00**	56	**1.00**	8−	**1.00**	8−	**1.00**
rungs rejected /15		15		12		13		15.0		13		12		8		14		11		15		15	

Table 27: **Severe band removal ladder** (amplitude gain $g = 0.5, 0.25, 0.1, 0.03, 0$; band energy retained g^2), as Table 26.

		learned / untrained control								non-learned										clip spaces			
band	level	PDE-KD		ImageNet		Poseidon		Random init		Pixels		Spectrum		Spectrum-native		PCA-white		PCA-raw		PDE-KD (3D)		JEPA	
low-k	L1	16	**1.00**	175+	.07	21	**1.00**	61	**1.00**	43	**1.00**	175+	.07	175+	.07	8−	**1.00**	163	**.97**	10	**1.00**	26	**1.00**
	L2	10	**1.00**	175+	.08	10	**1.00**	38	**1.00**	18	**1.00**	175+	.09	175+	.07	8−	**1.00**	75	**1.00**	8−	**1.00**	13	**1.00**
	L3	9	**1.00**	175+	.10	9	**1.00**	33	**1.00**	12	**1.00**	175+	.11	175+	.10	8−	**1.00**	65	**1.00**	8−	**1.00**	10	**1.00**
	L4	9	**1.00**	175+	.10	8	**1.00**	33	**1.00**	11	**1.00**	175+	.18	175+	.21	8−	**1.00**	61	**1.00**	8−	**1.00**	9	**1.00**
	L5	9	**1.00**	175+	.10	8	**1.00**	33	**1.00**	10	**1.00**	87	**1.00**	63	**1.00**	8−	**1.00**	59	**1.00**	8−	**1.00**	9	**1.00**
mid-k	L1	34	**1.00**	175+	.66	175+	**.83**	134	**.93**	175+	.64	175+	.17	175+	.17	8−	**1.00**	175+	.14	19	**1.00**	62	**1.00**
	L2	16	**1.00**	69	**1.00**	85	**1.00**	77	**1.00**	93	**1.00**	175+	**.90**	175+	**.96**	8−	**1.00**	175+	.27	10	**1.00**	31	**1.00**
	L3	14	**1.00**	42	**1.00**	56	**1.00**	72	**1.00**	62	**1.00**	59	**1.00**	48	**1.00**	8−	**1.00**	175+	.41	9	**1.00**	25	**1.00**
	L4	13	**1.00**	38	**1.00**	49	**1.00**	72	**1.00**	52	**1.00**	28	**1.00**	17	**1.00**	8−	**1.00**	175+	.51	9	**1.00**	23	**1.00**
	L5	12	**1.00**	37	**1.00**	47	**1.00**	73	**1.00**	49	**1.00**	11	**1.00**	8−	**1.00**	8−	**1.00**	175+	.54	9	**1.00**	21	**1.00**
high-k	L1	175+	.45	175+	**.86**	175+	.07	175+	.10	175+	.09	175+	.70	175+	**.81**	175+	.27	175+	.05	175+	**.90**	118	**.99**
	L2	98	**1.00**	43	**1.00**	175+	.11	175+	.13	175+	.17	54	**1.00**	41	**1.00**	175+	.77	175+	.05	68	**1.00**	52	**1.00**
	L3	54	**1.00**	26	**1.00**	175+	.17	175+	.13	175+	.22	20	**1.00**	14	**1.00**	145	**.97**	175+	.06	43	**1.00**	39	**1.00**
	L4	44	**1.00**	22	**1.00**	175+	.19	175+	.13	175+	.27	12	**1.00**	8	**1.00**	114	**1.00**	175+	.06	36	**1.00**	35	**1.00**
	L5	41	**1.00**	21	**1.00**	175+	.21	175+	.13	175+	.27	9	**1.00**	8−	**1.00**	105	**1.00**	175+	.06	33	**1.00**	34	**1.00**
rungs rejected /15		15		10		13		9.3		10		9		9		14		5		15		15	

E.3 Full results on Rayleigh–Bénard

This subsection reports every family, level and frame space of the perceptual suite and of the canonical spectral ladder on RB. Table 2 in the body shows a subset of families, with band injection and removal on the severe ladder (Tables 26–27). Table 28 counts the rungs rejected by one full-pool test. The same tests give the RB columns of Table 3: a perceptual family counts as rejected when its L2 test rejects, and a spectral family when its L5 test rejects; the Spectrum entries of the spectral columns use the Spectrum-native column. Tables 29 and 30 give power at the full pool for every level, including the severity-0 identity row of the spectral suite (Appendix E.1), and for each family n_{80} at L5. Raw KD per level, with each level's permutation null, is in Appendix E.9; the power tables summarize each space's response relative to its own null.

Power at the full pool. The power sweeps use the 175-IC pools of Appendix E.1. At $n = 175$ each pool is exhausted and replicates differ only in the 50/50 partition, so full-pool power is a rejection rate over partitions of one fixed sample, not over datasets.

PCA-white on the canonical spectral ladder. At the full pool, PCA-white reaches 80% power on 39 of the 55 canonical spectral rungs, more than any other frame space (PDE-KD 30, ImageNet 24, Poseidon 15; Table 30). At L5 it does so on 9 of the 11 families, against 10 for PDE-KD. In single full-pool tests, PDE-KD rejects 41 rungs and PCA-white 39 (Table 28). On the severe band ladder, single tests ($B = 20{,}000$) reject all 30 levels for PDE-KD and 28 for PCA-white (Tables 26–27). In single L5 tests, PCA-white rejects fewer spectral families than PDE-KD on every system of Table 3, RB included.

Intervals for n_{80}. For a fitted n_{80}, a 95% interval is a parametric bootstrap of the same monotone logistic over the 200 replicate rejections at each n (2,000 draws). At L5 the median interval width is 2.4 ICs and the largest 21, excluding one extrapolated fit. That fit is Poseidon's mid-k injection: its measured L5 power at the full pool is .90, yet its fitted n_{80} exceeds the pool (printed 175+, interval 186–243). The two readings of sample efficiency therefore differ in that cell. Counted by measured L5 power at the full pool (Table 30), PDE-KD reaches 80% power on 10 of the 11 spectral families, ImageNet on 7 and Poseidon on 6; counted by fitted n_{80} (the n_{80} rows of Table 30), Poseidon's count is 5. The interval reflects replicate noise only, conditional on the 175-IC pool. It does not reflect misfit of the logistic or the sampling of that pool, and apart from this extrapolated fit, censored values carry no interval.

Table 28: **Rungs rejected by a single full-pool permutation test** ($\alpha = 0.05$, $B = 2{,}000$; perceptual of 20, spectral of 55; arms: perceptual 175 versus 175 ICs, spectral 87 versus 88). Identity control p: the single test of the spectral severity-0 rung. Random init: mean over three initialization seeds of each seed's count; its identity entry is the number of seeds whose identity test rejects. Sensitivity to B: Appendix D.7.

	learned / untrained control				non-learned				
disturbance	PDE-KD	ImageNet	Poseidon	Random init	Pixels	Spectrum	Spectrum-native	PCA-white	PCA-raw
Perceptual: rungs rejected, of 5 per family									
Occlusion	4	5	4	3.0	3	5	–	4	3
Gaussian noise	5	5	5	4.7	4	5	–	4	2
Salt-and-pepper	5	5	5	5.0	5	5	–	5	2
Gaussian blur	4	5	0	4.0	0	4	–	0	0
total (of 20)	18	20	14	16.7	12	19	–	13	7
Spectral (isotropic): rungs rejected, of 5 per family									
Low-k injection	5	0	4	5.0	3	0	0	5	3
Mid-k injection	5	0	2	5.0	0	0	0	5	0
High-k injection	2	1	0	2.0	0	1	0	0	0
Low-k removal	4	0	5	3.0	3	1	1	5	2
Mid-k removal	3	2	3	1.7	2	1	1	5	0
High-k removal	2	2	1	0.0	0	1	2	1	0
Independent phase	4	4	3	0.0	2	2	0	4	2
Shared phase	4	4	0	0.0	2	2	0	4	1
Small-scale tilt	5	5	4	4.3	3	5	5	5	1
Large-scale tilt	4	4	3	0.0	3	4	4	5	0
Truncation	3	4	4	0.0	0	5	5	0	0
total (of 55)	41	26	29	21.0	18	22	18	39	9
identity control p	0.157	0.402	0.212	(0/3)	0.512	0.402	0.434	0.581	0.321

E.4 OTHER PHYSICAL SYSTEMS

Table 3 and Table 31 repeat the frame-level perceptual and spectral ladders on three further Well systems, to test whether the RB results carry over to other physics. All three are among the feature map's nine training families and are scored on their held-out test ICs:

- shear flow: 112 trajectories, frames 61–199 (RB's window), native 256×512;

Table 29: **Perceptual suite, RB: power at the full pool per level, and n_{80} at L5.** Power rows: bold marks power $\geq .80$, dark green $\geq .90$ and dark red $\leq .10$. n_{80} (L5) rows: total ICs at which a monotone logistic fit of the L5 power curve reaches 0.80; dark green marks $n_{80} < 8$ (printed 8–), dark red n_{80} above the 175-IC pool (175+, censored). Monte Carlo standard error at most $0.5/\sqrt{200} = 0.0354$. Random init: mean over three initialization seeds.

		learned / untrained control				non-learned				
disturbance	level	PDE-KD	ImageNet	Poseidon	Random init	Pixels	Spectrum	Spectrum-native	PCA-white	PCA-raw
Occlusion	L1	.14	.55	.07	.06	.06	**1.00**	–	.10	.06
	L2	**1.00**	**1.00**	.26	.10	.12	**1.00**	–	**1.00**	.10
	L3	**1.00**	**1.00**	**.99**	.33	.71	**1.00**	–	**1.00**	.34
	L4	**1.00**	**1.00**	**1.00**	**.98**	**1.00**	**1.00**	–	**1.00**	**.99**
	L5	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	–	**1.00**	**1.00**
	n_{80} (L5)	8–	9	9	22	23	11	–	8–	74
Gaussian noise	L1	**1.00**	**1.00**	.21	.13	.08	**1.00**	–	.26	.07
	L2	**1.00**	**1.00**	**.82**	.20	.16	**1.00**	–	**.96**	.12
	L3	**1.00**	**1.00**	**1.00**	.22	.44	**1.00**	–	**1.00**	.18
	L4	**1.00**	**1.00**	**1.00**	.27	**.91**	**1.00**	–	**1.00**	.32
	L5	**1.00**	**1.00**	**1.00**	.69	**1.00**	**1.00**	–	**1.00**	.58
	n_{80} (L5)	8–	8–	22	175+	97	8–	–	16	175+
Salt-and-pepper	L1	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	–	**.90**	.07
	L2	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	–	**1.00**	.09
	L3	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	–	**1.00**	.14
	L4	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	–	**1.00**	.23
	L5	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	–	**1.00**	.35
	n_{80} (L5)	8–	8–	8	8–	8–	8–	–	8–	175+
Gaussian blur	L1	.17	.67	.08	.09	.07	.26	–	.05	.06
	L2	**.89**	**1.00**	.07	.12	.07	**.98**	–	.06	.06
	L3	**1.00**	**1.00**	.08	.21	.07	**1.00**	–	.06	.06
	L4	**1.00**	**1.00**	.10	.31	.07	**1.00**	–	.08	.06
	L5	**1.00**	**1.00**	.10	.37	.07	**1.00**	–	.15	.06
	n_{80} (L5)	42	8–	175+	175+	175+	51	–	175+	175+

- compressible Euler with periodic boundaries: a fixed 175-trajectory subset of the 500 (the first 18 or 17 of each of the ten gas files), frames 20–100 (after the Riemann-quadrant start-up), native 512×512;
- Gray–Scott reaction–diffusion: 120 trajectories, frames 862–1000 (the last 139 frames, in the developed-pattern regime), native 128×128.

Disturbances, severities, the IC-level permutation test, the 200-replicate power curves and the n_{80} fit are those of Appendix E.1. The single full-pool tests of Table 3 use 2,000 permutations, and the power-curve replicates 999. We use only the test split of each system, so the perceptual candidate half is 56, 88 and 60 ICs and the spectral pool is 112, 175 and 120. Each system's IC grid is RB's, truncated at that maximum (with $n = 40$ added for the shear-flow and Gray–Scott perceptual curves); n_{80} is censored, and power read, at that maximum. A band component without a fitted n_{80} enters the band average at the system's spectral maximum rather than at 175, so band averages depend on the pool and are not comparable across systems.

Reading the two tables. Table 31 uses the canonical band ladder and Table 2 the severe one, so their injection and removal columns are not comparable. On its Random init row, Table 31 fits n_{80} to the seed-averaged power curve (Appendix C.2). The pooled binomial standard error behind that row's ‡ omits between-seed variation. Table 3 counts the families rejected (perceptual at L2, spectral at L5) by single tests on a disjoint split of the full pool, without multiplicity correction. These are descriptive rejection counts, not power estimates. Its HRRR column follows the out-of-distribution protocol of Appendix I.2. Its Spectrum entries use canonical-resize features for the perceptual columns and native-grid (Spectrum-native) features for the spectral columns, as in Table 2. Its Random init entries are the mean of the three seeds' counts (Appendix C.2).

System-specific inputs. Three inputs are built per system as for RB. First, the spectral band edges (the energy peak k_p and the 95%-energy wavenumber k_h of each channel) are measured on the scoring window of the raw fields. Unlike the RB band edges (Appendix E.2), they exclude the $k = 0$ cell and use unit pixel spacing. The spectrum peaks in the lowest radial bin for all five Euler-

Table 30: **Spectral suite, RB: power at the full pool per level**, including the severity-0 identity row (a check of the test's size against $\alpha = 0.05$), and n_{80} at L5. Row formats and colors as in Table 29; the identity row (three decimals) is not colored. Random init: mean over three initialization seeds.

		learned / untrained control				non-learned				
disturbance	level	PDE-KD	ImageNet	Poseidon	Random init	Pixels	Spectrum	Spectrum-native	PCA-white	PCA-raw
Low-k injection	L1	.20	.06	.14	.17	.11	.07	.07	**1.00**	.06
	L2	.74	.06	.68	.51	.24	.06	.07	**1.00**	.12
	L3	**1.00**	.05	**.99**	.79	.59	.06	.07	**1.00**	.21
	L4	**1.00**	.06	**1.00**	**.95**	**.92**	.06	.07	**1.00**	.34
	L5	**1.00**	.06	**1.00**	**.99**	**1.00**	.07	.07	**1.00**	.56
	n_{80} (L5)	37	175+	44	76	113	175+	175+	8	175+
Mid-k injection	L1	.14	.06	.06	.13	.08	.06	.06	**.98**	.06
	L2	.49	.07	.10	.33	.09	.06	.06	**1.00**	.05
	L3	**.87**	.07	.26	.60	.14	.08	.07	**1.00**	.06
	L4	**1.00**	.08	.58	**.81**	.20	.08	.08	**1.00**	.07
	L5	**1.00**	.12	**.90**	**.95**	.30	.11	.09	**1.00**	.10
	n_{80} (L5)	55	175+	175+	112	175+	175+	175+	9	175+
High-k injection	L1	.05	.09	.05	.04	.08	.07	.06	.04	.05
	L2	.07	.11	.06	.04	.08	.09	.08	.05	.05
	L3	.09	.13	.05	.07	.08	.11	.12	.09	.05
	L4	.14	.20	.05	.11	.08	.15	.15	.13	.05
	L5	.21	.22	.05	.14	.09	.17	.17	.18	.05
	n_{80} (L5)	175+	175+	175+	175+	175+	175+	175+	175+	175+
Low-k removal	L1	.22	.06	.22	.24	.13	.07	.07	**1.00**	.07
	L2	**1.00**	.06	**.96**	.71	.52	.07	.07	**1.00**	.20
	L3	**1.00**	.06	**1.00**	**.99**	**1.00**	.06	.07	**1.00**	.64
	L4	**1.00**	.07	**1.00**	**1.00**	**1.00**	.07	.07	**1.00**	**1.00**
	L5	**1.00**	.10	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	n_{80} (L5)	9	175+	8	33	10	87	63	8–	59
Mid-k removal	L1	.15	.06	.07	.16	.08	.06	.08	**1.00**	.04
	L2	.78	.07	.14	.41	.13	.08	.08	**1.00**	.06
	L3	**1.00**	.12	.35	.76	.25	.10	.11	**1.00**	.09
	L4	**1.00**	**.90**	**.95**	**.96**	**.85**	.24	.26	**1.00**	.17
	L5	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	.54
	n_{80} (L5)	12	37	47	73	49	11	8–	8–	175+
High-k removal	L1	.07	.06	.05	.05	.07	.07	.07	.05	.05
	L2	.10	.15	.06	.06	.07	.13	.13	.07	.05
	L3	.20	.47	.06	.08	.09	.24	.30	.14	.05
	L4	.69	**.98**	.08	.11	.11	**.88**	**.94**	.36	.05
	L5	**1.00**	**1.00**	.21	.13	.27	**1.00**	**1.00**	**1.00**	.06
	n_{80} (L5)	40	21	175+	175+	175+	9	8–	105	175+
Independent phase	L1	.06	.06	.05	.04	.07	.06	.07	.05	.04
	L2	.21	**1.00**	.08	.04	.09	.06	.07	**.81**	.06
	L3	**1.00**	**1.00**	.21	.05	.49	.13	.07	**1.00**	.33
	L4	**1.00**	**1.00**	.65	.05	**1.00**	**.93**	.07	**1.00**	**.98**
	L5	**1.00**	**1.00**	.76	.05	**1.00**	**1.00**	.07	**1.00**	**1.00**
	n_{80} (L5)	26	12	175+	175+	61	65	175+	11	116
Shared phase	L1	.05	.07	.05	.04	.07	.06	.07	.05	.04
	L2	.18	**1.00**	.05	.04	.09	.06	.07	**.93**	.06
	L3	**1.00**	**1.00**	.06	.05	.41	.13	.07	**1.00**	.32
	L4	**1.00**	**1.00**	.07	.05	**1.00**	**.93**	.07	**1.00**	**.94**
	L5	**1.00**	**1.00**	.07	.05	**1.00**	**1.00**	.07	**1.00**	**1.00**
	n_{80} (L5)	36	15	175+	175+	68	65	175+	10	140
Small-scale tilt	L1	.42	.31	.07	.12	.10	.24	.22	**1.00**	.04
	L2	**1.00**	**1.00**	.32	.39	.35	**.87**	**.85**	**1.00**	.05
	L3	**1.00**	**1.00**	**1.00**	**.99**	**1.00**	**1.00**	**1.00**	**1.00**	.17
	L4	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	.60
	L5	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
	n_{80} (L5)	8–	19	9	20	10	18	20	8–	123
Large-scale tilt	L1	.33	.49	.06	.08	.08	.32	.34	**1.00**	.06
	L2	**1.00**	**1.00**	.10	.13	.14	**.97**	**.97**	**1.00**	.07
	L3	**1.00**	**1.00**	.46	.22	.46	**1.00**	**1.00**	**1.00**	.08
	L4	**1.00**	**1.00**	**1.00**	.26	**.95**	**1.00**	**1.00**	**1.00**	.10
	L5	**1.00**	**1.00**	**1.00**	.27	**1.00**	**1.00**	**1.00**	**1.00**	.13
	n_{80} (L5)	11	8	83	175+	127	8–	8–	8–	175+
Truncation	L1	.05	.11	.06	.04	.08	**1.00**	**1.00**	.04	.05
	L2	.08	**1.00**	.17	.04	.08	**1.00**	**1.00**	.06	.05
	L3	.29	**1.00**	.17	.04	.09	**1.00**	**1.00**	.10	.05
	L4	**1.00**	**1.00**	.24	.05	.11	**1.00**	**1.00**	.21	.05
	L5	**1.00**	**1.00**	.22	.07	.17	**1.00**	**1.00**	.54	.05
	n_{80} (L5)	68	8	175+	175+	175+	8–	8–	175+	175+
identity (size check)	L0	.040	.060	.050	.040	.070	.065	.065	.035	.050

Table 31: **Per-system discriminative power and sample efficiency.** Each cell: n_{80} at L5 | power at L2 at the largest swept IC count (perceptual 56/88/60, spectral 112/175/120 for shear flow / Euler-periodic / Gray–Scott); 8– and n+ are censored. Injection and removal: canonical band ladder, low-, mid- and high-band values averaged before display censoring. Format and colors as in Table 26. ‡The L5 power curve falls as n grows by more than three standard errors, so the logistic n_{80} is not a meaningful summary (for a band average, in at least one component). Random init: mean over three initialization seeds (Appendix C.2); on its row, ‡ uses the pooled binomial standard error of the seed-averaged power curve, conditional on the three networks.

	perceptual			spectral		
space	Occlusion	Gaussian noise	Gaussian blur	Phase scrambling	Injection (band avg.)	Removal (band avg.)
Shear flow						
PDE-KD	8– | .39	8– | **1.00**	56+ | .12	112+ | .09	81 | .22	40 | .27
ImageNet	9 | **1.00**	8– | **1.00**	20 | .19	31 | .11	112+ | .03	79 | .05
Poseidon	22 | .08	45 | .16	56+ | .07	112+ | .05	112+ | .15	55 | .30
Random init	56+ | .04	56+ | .05	56+‡ | .04	112+ | .02	112+ | .08	112+‡ | .08
Pixels	22 | .07	56+ | .07	56+ | .04	112+ | .05	112+ | .08	62 | .08
Spectrum	8– | **1.00**	8– | **1.00**	27 | .57	112+ | .03	112+ | .03	48 | .04
PCA-white	8– | .11	37 | .16	56+ | .04	67 | .06	112+ | .11	49 | .14
Euler-periodic						
PDE-KD	8– | **.81**	8– | **1.00**	12 | **.81**	14 | **1.00**	140 | .45	22 | .62
ImageNet	8 | **1.00**	8– | **1.00**	8– | **1.00**	8– | **1.00**	134 | .14	64 | .52
Poseidon	33 | .06	88+ | .07	88+ | .05	175+ | .05	175+ | .08	128 | .10
Random init	88+ | .06	88+‡ | .06	88+ | .06	175+‡ | .03	175+ | .04	175+ | .05
Pixels	88+ | .05	88+ | .05	88+ | .05	175+ | .05	175+ | .05	175+ | .05
Spectrum	9 | **1.00**	8– | **1.00**	9 | **1.00**	175+ | .04	139 | .23	11 | .33
PCA-white	88+ | .08	88+ | .05	88+ | .06	175+ | .06	175+ | .07	168 | .08
Gray–Scott						
PDE-KD	12 | .09	8 | .46	22 | .21	10 | **1.00**	56 | .69	63 | .57
ImageNet	13 | .13	8– | **1.00**	13 | .36	9 | **1.00**	120+ | .11	59 | .26
Poseidon	54 | .10	60+ | .13	60+ | .08	120+ | .04	120+ | .04	120+ | .05
Random init	8– | .11	27 | .45	20 | .14	120+ | .44	62 | .48	84 | .47
Pixels	60+ | .08	60+ | .07	60+ | .07	120+ | .04	120+ | .04	120+ | .04
Spectrum	8– | **1.00**	8– | **1.00**	60+‡ | .06	120+ | .09	120+ | .09	48‡ | .10
PCA-white	60+ | .07	60+ | .06	60+ | .06	120+ | .04	120+ | .04	120+ | .05

periodic channels and for two of the four shear-flow channels (tracer and v_x), so k_p is that bin there. Second, the PCA-white bases ($k = 25$ for shear flow, 298 for Euler-periodic and 148 for Gray–Scott) are fitted on most of the scored test trajectories, with the held-out elbow on the rest: three of four per shear-flow file, 14 of 20 per Gray–Scott file, and seven of the ten Euler gas files. Because only the test splits are used, the basis is in-sample, which favors that control. Third, Poseidon's input routing follows Appendix C.1: on Euler-periodic its density channel receives energy and density is dropped, and on Gray–Scott only A enters.

Calibration. The clean-versus-clean false-positive rate of each system and space was measured on 40 draws, extended to 200 when the disjoint design's 40-draw rate reached 0.075 (Table 21). Poseidon on shear flow rejects in 19 of 200 draws (interval 0.058–0.144); this rate is not significant after Holm correction over all disclosed cells (Appendix D.6), and the cell enters every comparison. Every spectral identity check passes: at severity 0 the maximum relative feature difference from the clean features is at most 2.2×10^{-5}, against a tolerance of 10^{-3}.

E.5 Monotonicity in severity

FID was validated by showing that it increases with the level of each disturbance (Heusel et al., 2017). We apply the same test to KD: a useful score should grow with the size of a difference. Table 32 correlates KD with the severity level of the off-manifold disturbances of Section 4.1 on 15 RB ladders, and Appendix E.9 lists their per-level KD scores.

Protocol. The mean row of Table 32 averages the signed Spearman correlations between KD and severity over four perceptual and eleven spectral families, each with five levels (13 for Spectrum,

Table 32: **Monotonicity of KD in severity, per RB ladder.** Spearman ρ between KD and the five severity levels of each ladder; 1.00 means KD increases at every level, and red marks the ladders where it does not. Last row: column mean. –: undefined, KD constant across the five levels to four significant figures; Spectrum's mean is over its 13 defined ladders. Random init: mean over three initialization seeds, then each seed (s0–s2). Spectrum uses canonical-resize features for the perceptual ladders and native-grid features for the spectral ladders.

	learned / untrained control							non-learned		
ladder	PDE-KD	ImageNet	Poseidon	Random init	s0	s1	s2	Pixels	Spectrum	PCA-white
Perceptual										
Occlusion	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Gaussian noise	1.00	1.00	1.00	0.43	0.30	0.30	0.70	1.00	1.00	1.00
Salt-and-pepper	1.00	–0.60	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Gaussian blur	1.00	1.00	0.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Spectral (isotropic)										
Low-k injection	1.00	0.90	1.00	1.00	1.00	1.00	1.00	1.00	–0.10	1.00
Mid-k injection	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
High-k injection	1.00	1.00	0.70	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Low-k removal	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Mid-k removal	1.00	1.00	1.00	0.97	1.00	1.00	0.90	1.00	0.90	1.00
High-k removal	1.00	1.00	1.00	–1.00	–1.00	–1.00	–1.00	1.00	0.90	1.00
Independent phase	1.00	1.00	1.00	0.30	1.00	–1.00	0.90	0.90	–	1.00
Shared phase	1.00	1.00	1.00	0.33	1.00	–0.90	0.90	0.90	–	1.00
Small-scale tilt	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Large-scale tilt	1.00	1.00	1.00	–0.43	–0.60	0.00	–0.70	1.00	1.00	1.00
Truncation	1.00	1.00	0.90	–1.00	–1.00	–1.00	–1.00	1.00	1.00	1.00
mean	1.00	0.89	0.97	0.51	0.58	0.36	0.58	0.99	0.90	1.00

whose two phase-ladder correlations are undefined; see below). The perceptual and spectral suites use different IC pools, so their 15 effect directions are averaged descriptively rather than pooled into one test; a mean rank across ladders would mix unrelated severity axes. The severity correlations use each level's raw KD even where it lies near that feature space's split-half floor, so monotonicity describes the direction of the response and does not by itself establish a resolved disturbance response. Random init is the mean over three seeds of each seed's 15-ladder mean. With five levels, only a perfect ordering reaches the exact 5% threshold (Table 22), so these correlations are descriptive.

Severity. KD in PDE-KD and in PCA-white increases with severity on all 15 ladders. Pixels orders every ladder except the two phase-scrambling ladders ($\rho = 0.9$ on each), and Poseidon every ladder except blur, high-k injection and truncation ($\rho = 0.9$, 0.7 and 0.9). ImageNet orders every ladder except low-k injection ($\rho = 0.9$) and salt-and-pepper noise ($\rho = -0.6$), on which its KD is largest at L1 and smallest at L3, while every level is still rejected (Table 35). Spectrum's spectral ladders use native-grid features, which are phase-blind (Appendix C.5). Its KD on the two phase-scrambling ladders is therefore the same at every level to four significant figures (Spectrum-native in Table 40), so a rank correlation there would order round-off differences. Table 32 leaves these two entries undefined, and Spectrum's mean, 0.90, is over its 13 other ladders. Spectrum is also not monotone on low-k injection ($\rho = -0.1$) or on mid- and high-k removal (0.9 each). Random init has the lowest mean correlation, 0.51: its KD *falls* as truncation and high-k removal grow ($\rho = -1$ in every seed), and it is not monotone on noise and large-scale tilt in any seed, nor on phase scrambling in two of the three. The mean correlations therefore mainly separate Random init from the other six spaces, whose means lie between 0.89 and 1.00.

E.6 Temporal suite

This subsection compares the two clip spaces, PDE-KD (3D) and JEPA, on disturbed RB clips.

Design. The clip suite applies the four perceptual operators and the seven temporal ladders of Appendix E.2 to clips of $T = 20$ frames for PDE-KD (3D) and $T = 16$ for JEPA (Table 33). The canonical spectral ladder is scored on clips of the 175 RB test ICs, 87 versus 88 per arm (Table 34). For the clip suite, both clip spaces use 175/175 arms of the same 350-IC validation-plus-test pool. For PDE-KD (3D), one test coefficient cell of that pool was corrected (Appendix B.6); its clip-suite

counts and its perceptual and temporal cells in Table 2 use the corrected pool, and on the uncorrected pool its clip suite rejects 43 rungs. The comparison is not a controlled architectural ablation: the arms of the two clip spaces are drawn by different splits, and the two spaces use different clip preprocessing. Unlike the spectral suite, the clip suite has no severity-0 identity rung; the clean-versus-clean rates of both clip spaces on its windows are in Table 20. Table 2 shows two of the seven temporal ladders, local swap (4–20 adjacent swapped pairs) and low-severity global swap (1–5 arbitrary swapped pairs), selected for display rather than by a stated rule.

JEPA checkpoint and readout. The JEPA row of Table 2 is the sample-axis retrain of Appendix C.4, read from its 128-dimensional spatial mean pool. Checkpoints are indexed from zero, so epoch 29 below is the checkpoint after 30 epochs and epoch 3 the one after four. Our first training run of this encoder computed the VICReg variance and covariance statistics over the flattened batch–space axis instead; its epoch-3 and epoch-29 checkpoints are the columns "ep. 3" and "shipped" of Tables 33 and 34. With further training, that run collapsed. Its sample-varying share of the feature norm fell from 0.91 to 0.02, and its effective rank to about 1.5. Its coefficient-probe MSE rose from 0.0508 at epoch 3 to 0.2784 at epoch 29, and the temporal rungs it rejects fell from 21 to 4 of 35. The sample-axis retrain avoids this collapse: at epoch 29 its sample-varying share is 0.64, its effective rank 13.9 and its coefficient-probe MSE 0.1082, and it rejects 17 of the 35 temporal rungs. Single full-pool tests reject the following numbers of rungs, each out of 55, in the clip suite (perceptual and temporal) and in the spectral suite ($B = 20{,}000$):

Clip space and readout (column label in Tables 33–34)	Clip	Spectral
PDE-KD (3D)	41	50
JEPA, first run, epoch 3 ("ep. 3")	35	35
JEPA, first run, epoch 29, collapsed ("shipped")	20	42
JEPA, sample-axis retrain, epoch 29 ("JEPA")	32	38
its attention pooler after the MLP ("post-MLP tap")	28	35
its attention pooler before the MLP ("pre-MLP tap")	33	46

No JEPA checkpoint or readout dominates the others across the two suites. The two attention-pooler readouts have n_{80} values but no full-pool L2 power; their "–" power cells are missing measurements, not zero results. On the two temporal ladders of Table 2, the epoch-3 checkpoint of the first run reaches 80% power at L5 on low-severity global swap ($n_{80} = 93$), which the sample-axis retrain does not within the sampled range; on local swap no JEPA checkpoint or readout does. The JEPA verdict of Section 4.1 on these two ladders therefore depends on the checkpoint.

Other temporal ladders. Of the five temporal ladders not shown in Table 2, neither clip space reaches 80% power within the sampled range on low-severity local swap or interpolation dropout: n_{80} at L5 lies above that range (175+), and L2 power is .04 for PDE-KD (3D) and .04 and .05 for JEPA. Global swap, interleave and switching have n_{80} within the sampled range for both clip spaces, except JEPA on global swap (175+; L2 power .62).

Table 33: **Perceptual and temporal suite on RB clips: n_{80} at L5 | power at L2**, 175 versus 175 ICs ($T = 20$ frames for PDE-KD (3D), $T = 16$ for JEPA). JEPA: the sample-axis retrain used in the body; the other JEPA columns are checkpoint and readout variants (Appendix E.6). –: full-pool L2 power not measured. Format and colors as in Table 26.

disturbance	PDE-KD (3D)		JEPA		JEPA (ep. 3)		JEPA (shipped)		JEPA (post-MLP tap)		JEPA (pre-MLP tap)	
Occlusion	8–	**1.00**	8–	.60	8–	**1.00**	8–	**.97**	46	–	23	–
Gaussian noise	8	**1.00**	14	**1.00**	14	**.97**	8–	**1.00**	14	–	12	–
Salt-and-pepper	8–	**1.00**	8–	**1.00**	8–	**1.00**	8–	**1.00**	9	–	8–	–
Gaussian blur	69	.28	175+	.07	175+	.06	118	.09	101	–	23	–
Local swap	97	.22	175+	.06	175+	.08	175+	.07	175+	–	175+	–
Global swap	9	**1.00**	175+	.62	59	**1.00**	175+	.07	175+	–	140	–
Low-severity local swap	175+	.04	175+	.04	175+	.05	175+	.07	175+	–	175+	–
Low-severity global swap	12	**1.00**	175+	.10	93	.69	175+	.07	175+	–	175+	–
Interpolation dropout	175+	.04	175+	.05	175+	.04	175+	.07	175+	–	175+	–
Interleave	8–	**1.00**	31	**1.00**	21	**1.00**	175+	.01	12	–	8–	–
Switching	8–	**1.00**	22	**1.00**	24	**1.00**	175+	.10	26	–	11	–

Table 34: **Spectral suite on RB clips: n_{80} at L5 | power at L2** for the clip spaces, 87 versus 88 test ICs. Identity row: the severity-0 control. Format and colors as in Table 26.

disturbance	PDE-KD (3D)		JEPA		JEPA (ep. 3)		JEPA (shipped)		JEPA (post-MLP tap)		JEPA (pre-MLP tap)	
Low-k injection	43	.79	93	.43	51	.67	15	**1.00**	156	–	51	–
Mid-k injection	34	**.83**	96	.34	175+	.14	175+	.09	175+	–	62	–
High-k injection	175+	.09	175+	.13	175+	.07	48	.44	175+	–	61	–
Low-k removal	8–	**1.00**	9	.79	12	**.89**	8–	**1.00**	34	–	14	–
Mid-k removal	8	**.99**	21	.46	42	.20	20	.12	33	–	15	–
High-k removal	33	.12	34	.23	86	.08	8–	**1.00**	38	–	11	–
Independent phase	8–	**1.00**	10	**1.00**	8–	**1.00**	43	.09	13	–	8–	–
Shared phase	8–	**1.00**	9	**1.00**	8–	**1.00**	48	.09	12	–	8–	–
Small-scale tilt	8–	**1.00**	8–	**1.00**	13	**.80**	8–	**1.00**	8–	–	8–	–
Large-scale tilt	8–	**1.00**	54	**.97**	31	.62	8	**1.00**	41	–	12	–
Truncation	70	.19	24	**1.00**	89	.41	8–	**1.00**	33	–	12	–
identity (size check)	.055		.075		.070		.060		–		–	

E.7 SPECIFICITY: SYMMETRIES OF THE GOVERNING EQUATIONS

A useful score should *not* respond to transformations that are exact symmetries of the equations and boundary conditions. We call such transformations *legal* and apply them to every frame of one arm: periodic translations, reflections, and draws of the training pipeline's augmentation sampler, which combine the legal operations. The released feature map was trained without augmentation, so this family tests invariances outside its training. Each control configuration breaks a boundary condition: translations along a wall-bounded (RB) or non-periodic (Euler-open) axis, reflections across a wall, half-domain shears, and reflections that omit the sign change of the vector component normal to the mirror. The systems are RB (350 ICs; 12 legal and 12 control configurations) and Euler with open boundaries (500 ICs; 8 legal, 16 control). A third system, Euler with periodic boundaries, is excluded because periodic translations of its IC ensemble are not symmetries of that ensemble (*Euler with periodic boundaries* below). The score is a per-cell paired self-KD (transformed versus untransformed frames of the same ICs), averaged over coefficient cells. A configuration is flagged when that mean exceeds two standard errors across cells, an uncalibrated descriptive rule (Appendix D.4). The pixel-space energy distance and two-sample CRPS (the pointwise comparators) are scored in the same design. Figure 9 gives each score's false-positive rate (legal configurations flagged) and false-negative rate (controls not flagged), with the counts. No feature space flags a legal configuration on either system, and PDE-KD and ImageNet flag every control. Each pointwise comparator flags one legal RB configuration, the same realization of the augmentation sampler (one of three). On RB the CRPS misses all five half-shears and the unflipped-vector control.

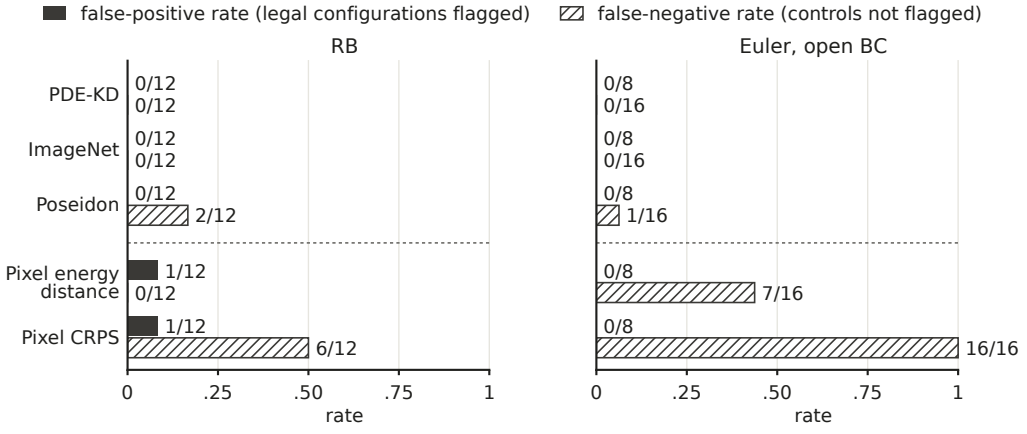


Figure 9: **Symmetry-invariance suite: false-positive and false-negative rates.** Per score, the fraction of legal configurations flagged (solid) and of controls not flagged (hatched), with counts beside the bars. Above the dashed line: KD in three feature spaces; below: pixel-space energy distance and two-sample CRPS.

Reading the flag rule. The across-cell standard error is optimistic: the estimand is nonnegative, yet some scores fall more than 2 SE below zero. These are 8 of the 144 KD scores in the figure (48 configurations in three feature spaces), 10 of 225 over all three systems, and 2 of the 96 pixel-space scores. The rule therefore flags too readily. The false-negative rates are optimistic, the zero false-positive rates of the feature spaces err on the conservative side, and the single legal flag of each pointwise comparator on RB may partly reflect the same optimism. "Not flagged" bounds the effect but does not establish invariance at a stated confidence.

Sources of the RB specificity. On RB the specificity is partly architectural: 11 of the 12 legal configurations contain a periodic x-roll (8 are pure rolls), and all three feature spaces end in a spatial mean pool, which makes them largely insensitive to rolls by construction. The Boussinesq mid-plane reflection, an exact symmetry of the RB equations and wall conditions, is not tested; it is a vertical flip with the vertical velocity and the buoyancy deviation from the conductive profile negated and the pressure adjusted. The wall-reflection control negates only the vertical velocity.

Marginal design. The counts also depend on the design and the rule. The conditional design followed a marginal design, which pools all coefficient cells and flags a configuration when its median paired change over ten split-half seeds exceeds the across-seed scatter of the clean-versus-clean floor. Under the marginal design, the energy distance and the CRPS flag 7 and 8 of the 12 legal RB configurations, because a legal roll shifts the pooled per-pixel distribution, while no feature space flags a legal RB or Euler-open configuration. Its KD control recall is 12/12 on RB and 15/16 on Euler-open for PDE-KD, 12/12 and 16/16 for ImageNet, and 7/12 and 3/16 for Poseidon. With a band of two across-seed standard deviations of the paired change instead, each pointwise comparator flags 1 of the 12 legal RB configurations, and KD recall is 11/12 and 15/16 for PDE-KD, 12/12 and 16/16 for ImageNet, and 7/12 and 0/16 for Poseidon. ImageNet thus flags one control more than PDE-KD under each marginal rule on Euler-open (the unflipped-vector control) and under the second rule on RB (the smallest half-shear). On Euler-open, vRMSE separates legal from control configurations with inverted polarity (legal 1.90–3.40, controls 0.36–1.39), because it measures how far pixels moved rather than whether the move is a symmetry.

Euler with periodic boundaries. An audit of this system's initial conditions showed that periodic translations of its finite ensemble are not distribution-preserving: about 25% of the boundaries between the initial condition's constant regions (24.9% of row and 25.3% of column boundaries) sit on the periodic seam. A calibrated relabeling test exposed the mislabel: it rejects 12 (PDE-KD), 15 (ImageNet) and 1 (Poseidon) of the 21 configurations labeled legal (about 1 expected by chance), all in the four families that roll the domain. In no space does it reject any of the five configurations that survive the audit. On the same system the conditional score flags 8, 14 and 1 of the 21.

E.8 Rollout resilience to corrupted conditioning frames

As a check that KD responds to errors that models produce, we corrupted only the conditioning frames of RB rollouts with each perceptual family at L5 on the native 512×128 grid (blur σ in native pixels; one realization shared by all conditioning frames; 175 test ICs; leads 61–183; paired design). Occlusion raises Lola-c16's score $8.0\times$ over its clean rollout, whereas blur leaves it unchanged ($0.98\times$). Noise and salt-and-pepper raise Walrus-PT's score $20.7\times$ and $35.7\times$. Walrus-PT's horizon vRMSE nevertheless falls under three of the four families, noise included, and rises under salt-and-pepper. The models condition on different numbers of frames (Lola 1, Walrus-FT 6, Walrus-PT 10), so these ratios are not comparable across models.

E.9 Raw scores

Tables 35–42 give the raw single-test results behind the rung counts of Table 28. For each space, family and level they list the KD score, the 95th percentile of that level's permutation null, and the permutation p. The null percentile differs from level to level, so a larger score does not by itself mean a stronger rejection. KD magnitudes are comparable only within one space, never across spaces (Appendix C.7); spaces are compared through power and n_{80}.

Table 35: **Raw response to the perceptual disturbances, RB (learned spaces).** Disjoint KD $\times 10^3$ at each severity level, the 95th percentile q_{95} of that level's own IC-permutation null, and the one-sided permutation p ($B = 2{,}000$; 175 versus 175 ICs per arm); bold: $p \leq 0.05$. Floor row: clean-versus-clean KD on the same partition (no test).

		PDE-KD			ImageNet			Poseidon		
disturbance	level	KD	q_{95}	p	KD	q_{95}	p	KD	q_{95}	p
clean vs clean (floor)	–	−0.290	–	–	−14.5	–	–	−4.61	–	–
Occlusion	L1	0.0792	0.703	.296	**67.45**	30.2	.005	1.66	14.1	.283
	L2	**5.14**	0.662	<.001	**236**	33.85	<.001	**23.6**	14.9	.015
	L3	**25.0**	0.629	<.001	**528**	24.8	<.001	**99.2**	14.8	<.001
	L4	**77.7**	0.796	<.001	**862**	21.8	<.001	**322.5**	14.3	<.001
	L5	**194**	1.695	<.001	**1209**	24.3	<.001	**1086**	18.85	<.001
Gaussian noise	L1	**8.19**	0.707	<.001	**2323**	32.9	<.001	**17.6**	14.4	.042
	L2	**72.3**	1.01	<.001	**4863**	51.6	<.001	**69.6**	15.8	<.001
	L3	**239**	2.46	<.001	**5896**	60.05	<.001	**156**	16.1	<.001
	L4	**518**	5.48	<.001	**6714**	66.05	<.001	**283**	16.6	<.001
	L5	**903**	7.87	<.001	**7431**	65.2	<.001	**459**	16.0	<.001
Salt-and-pepper	L1	**1.07×10^5**	901	<.001	**1.95×10^4**	163.5	<.001	**148**	14.4	<.001
	L2	**7.29×10^5**	6374	<.001	**8857**	78.5	<.001	**276**	15.8	<.001
	L3	**2.30×10^6**	2.04×10^4	<.001	**6589**	59.9	<.001	**516**	15.0	<.001
	L4	**5.26×10^6**	4.58×10^4	<.001	**6889**	61.5	<.001	**879**	16.9	<.001
	L5	**1.01×10^7**	8.91×10^4	<.001	**7402**	66.6	<.001	**1360**	18.6	<.001
Gaussian blur	L1	0.3225	0.718	.145	**84.6**	31.7	.005	−4.705	16.3	.681
	L2	**2.26**	0.633	.001	**694**	26.05	<.001	−4.80	14.3	.685
	L3	**4.56**	0.592	<.001	**1687**	28.0	<.001	−4.69	14.7	.685
	L4	**6.545**	0.642	<.001	**2491**	34.7	<.001	−4.41	15.8	.677
	L5	**8.40**	0.667	<.001	**2746**	38.9	<.001	−4.00	15.0	.613

Table 36: **Raw perceptual response, RB (Pixels, Spectrum).** As Table 35.

		Pixels			Spectrum		
disturbance	level	KD	q_{95}	p	KD	q_{95}	p
Occlusion	L1	−7.50	45.0	.566	**1.09×10^6**	3.25×10^4	<.001
	L2	20.3	38.2	.141	**1.42×10^6**	3.85×10^4	<.001
	L3	**106**	37.25	<.001	**1.61×10^6**	3.59×10^4	<.001
	L4	**279**	29.9	<.001	**1.73×10^6**	3.34×10^4	<.001
	L5	**547**	28.9	<.001	**1.79×10^6**	3.99×10^4	<.001
Gaussian noise	L1	6.21	45.0	.333	**2.12×10^6**	3.48×10^4	<.001
	L2	**41.1**	38.8	.046	**2.655×10^6**	4.31×10^4	<.001
	L3	**90.1**	40.4	.003	**2.91×10^6**	4.50×10^4	<.001
	L4	**151**	37.8	<.001	**3.05×10^6**	4.77×10^4	<.001
	L5	**222**	32.8	<.001	**3.14×10^6**	4.82×10^4	<.001
Salt-and-pepper	L1	**277**	52.7	<.001	**3.355×10^6**	4.90×10^4	<.001
	L2	**1123**	42.4	<.001	**3.36×10^6**	4.57×10^4	<.001
	L3	**2601**	46.45	<.001	**3.37×10^6**	4.43×10^4	<.001
	L4	**4683**	59.9	<.001	**3.38×10^6**	5.06×10^4	<.001
	L5	**7570**	73.4	<.001	**3.40×10^6**	4.87×10^4	<.001
Gaussian blur	L1	−11.45	48.2	.632	2.89×10^4	3.22×10^4	.061
	L2	−11.4	46.4	.624	**9.24×10^4**	2.62×10^4	.002
	L3	−11.2	44.2	.597	**1.455×10^5**	2.345×10^4	<.001
	L4	−10.9	46.6	.624	**1.98×10^5**	2.57×10^4	<.001
	L5	−10.45	50.7	.608	**2.44×10^5**	2.80×10^4	<.001

Table 37: **Raw perceptual response, RB (PCA spaces).** As Table 35.

		PCA-white			PCA-raw		
disturbance	level	KD	q_{95}	p	KD	q_{95}	p
Occlusion	L1	−0.613	1.78	.712	-3.89×10^{10}	1.31×10^{11}	.695
	L2	**16.6**	1.625	<.001	2.33×10^{10}	1.07×10^{11}	.293
	L3	**104**	2.48	<.001	$\mathbf{1.59\times10^{11}}$	9.22×10^{10}	.010
	L4	**327**	4.83	<.001	$\mathbf{3.28\times10^{11}}$	7.77×10^{10}	<.001
	L5	**653.5**	8.55	<.001	$\mathbf{4.54\times10^{11}}$	8.06×10^{10}	<.001
Gaussian noise	L1	1.395	1.645	.074	-1.67×10^{10}	1.26×10^{11}	.551
	L2	**7.41**	1.51	<.001	3.52×10^{10}	1.08×10^{11}	.229
	L3	**16.7**	1.45	<.001	9.525×10^{10}	1.02×10^{11}	.059
	L4	**29.3**	1.405	<.001	$\mathbf{1.58\times10^{11}}$	9.905×10^{10}	.010
	L5	**45.3**	1.38	<.001	$\mathbf{2.19\times10^{11}}$	8.79×10^{10}	.001
Salt-and-pepper	L1	**5.37**	1.76	<.001	-2.39×10^{10}	1.44×10^{11}	.584
	L2	**24.8**	1.72	<.001	1.95×10^{10}	1.16×10^{11}	.323
	L3	**57.4**	1.625	<.001	7.38×10^{10}	1.15×10^{11}	.115
	L4	**104**	1.815	<.001	$\mathbf{1.36\times10^{11}}$	1.18×10^{11}	.034
	L5	**166**	2.12	<.001	$\mathbf{2.04\times10^{11}}$	1.02×10^{11}	.006
Gaussian blur	L1	−1.28	1.73	.908	-5.28×10^{10}	1.37×10^{11}	.738
	L2	−1.19	1.68	.885	-5.23×10^{10}	1.45×10^{11}	.733
	L3	−0.956	1.59	.840	-5.15×10^{10}	1.38×10^{11}	.726
	L4	−0.537	1.72	.695	-5.03×10^{10}	1.37×10^{11}	.736
	L5	0.152	1.755	.407	-4.85×10^{10}	1.44×10^{11}	.717

Table 38: **Raw perceptual response, RB (Random init).** As Table 35, for Random init: KD and q_{95} are means over three initialization seeds; rej.: number of seeds whose test rejects at $\alpha = 0.05$; bold: at least two seeds reject. No p is printed, because p-values are not averaged.

		Random init		
disturbance	level	KD	q_{95}	rej.
Occlusion	L1	33.1	210	0/3
	L2	150	184	0/3
	L3	**494**	154	3/3
	L4	**1206**	154	3/3
	L5	**2289**	126	3/3
Gaussian noise	L1	**245**	173.5	3/3
	L2	**361**	163	3/3
	L3	**383**	162	3/3
	L4	**360**	166	2/3
	L5	**461**	155	3/3
Salt-and-pepper	L1	$\mathbf{6.17\times10^{10}}$	5.175×10^{8}	3/3
	L2	$\mathbf{4.75\times10^{11}}$	4.09×10^{9}	3/3
	L3	$\mathbf{1.64\times10^{12}}$	1.42×10^{10}	3/3
	L4	$\mathbf{3.99\times10^{12}}$	3.28×10^{10}	3/3
	L5	$\mathbf{8.00\times10^{12}}$	6.96×10^{10}	3/3
Gaussian blur	L1	104	198	0/3
	L2	**217**	185	3/3
	L3	**337**	185	3/3
	L4	**451**	179	3/3
	L5	**571**	204	3/3

Table 39: **Raw response to the spectral disturbances, RB (learned spaces).** As Table 35, with 87 versus 88 ICs per arm; L0: the identity rung (clean-versus-clean floor).

		PDE-KD			ImageNet			Poseidon		
disturbance	level	KD	q_{95}	p	KD	q_{95}	p	KD	q_{95}	p
identity	L0	0.587	1.29	.157	−6.26	69.6	.402	8.89	27.9	.212
Low-k injection	L1	**2.00**	1.31	.022	−7.20	74.1	.395	9.43	30.0	.198
	L2	**4.47**	1.38	<.001	−7.31	73.35	.417	**32.2**	29.6	.041
	L3	**7.81**	1.56	<.001	−6.72	66.5	.412	**70.4**	31.7	.004
	L4	**11.9**	1.53	<.001	−5.525	61.1	.375	**120**	29.1	<.001
	L5	**16.6**	1.66	<.001	−3.80	70.5	.374	**177**	29.4	<.001
Mid-k injection	L1	**2.16**	1.15	.006	−2.85	70.5	.356	10.3	28.3	.177
	L2	**4.45**	1.20	<.001	3.24	67.7	.289	17.2	27.1	.099
	L3	**7.29**	1.52	<.001	11.2	73.0	.241	29.0	30.3	.057
	L4	**10.6**	1.43	<.001	20.5	68.4	.180	**45.4**	28.0	.014
	L5	**14.4**	1.435	<.001	30.9	75.3	.149	**66.0**	28.2	.004
High-k injection	L1	0.750	1.29	.128	10.75	66.3	.240	8.22	32.5	.212
	L2	1.00	1.27	.080	30.6	73.2	.147	8.00	28.9	.214
	L3	1.33	1.37	.051	51.8	73.8	.088	8.14	32.0	.213
	L4	**1.71**	1.40	.033	73.6	74.3	.052	8.57	30.7	.209
	L5	**2.15**	1.25	.010	**95.4**	69.3	.024	9.22	30.5	.179
Low-k removal	L1	0.606	1.21	.149	−4.37	79.3	.391	**41.7**	29.5	.026
	L2	**2.89**	1.23	.004	−1.41	68.5	.346	**128**	30.6	.001
	L3	**9.67**	1.32	<.001	2.88	68.4	.305	**307**	33.9	<.001
	L4	**28.0**	1.55	<.001	9.09	78.4	.250	**672.5**	38.0	<.001
	L5	**70.1**	2.27	<.001	20.2	69.1	.177	**1669**	46.1	<.001
Mid-k removal	L1	−0.0493	1.19	.395	−5.38	70.5	.387	13.7	30.8	.140
	L2	0.623	1.20	.148	3.03	68.5	.304	25.8	29.4	.066
	L3	**3.285**	1.22	.001	26.2	66.5	.155	**47.9**	32.4	.013
	L4	**9.30**	1.14	<.001	**85.4**	62.3	.026	**87.0**	31.7	.001
	L5	**24.85**	1.36	<.001	**336.5**	50.7	<.001	**227**	33.2	<.001
High-k removal	L1	0.549	1.175	.142	−17.6	65.5	.630	10.15	29.3	.172
	L2	0.690	1.27	.128	−17.3	64.3	.639	12.2	32.8	.161
	L3	1.13	1.21	.058	10.3	59.9	.243	15.6	32.8	.133
	L4	**2.25**	1.21	.009	**117**	50.7	.010	21.3	29.2	.079
	L5	**8.82**	1.12	<.001	**785**	61.4	<.001	**40.6**	28.4	.024
Independent phase	L1	0.675	1.21	.123	6.27	74.8	.280	9.95	30.8	.186
	L2	**1.97**	1.27	.020	**185**	71.4	.004	17.0	30.2	.117
	L3	**8.12**	1.335	<.001	**986**	84.35	<.001	**42.2**	27.5	.019
	L4	**14.0**	1.29	<.001	**1587**	72.5	<.001	**67.4**	28.4	.004
	L5	**15.9**	1.32	<.001	**1651**	84.3	<.001	**75.8**	29.8	.003
Shared phase	L1	0.6285	1.17	.132	12.7	73.0	.221	9.21	29.3	.183
	L2	**1.475**	1.36	.044	**257**	83.8	.001	10.45	30.6	.176
	L3	**5.29**	1.35	<.001	**1105**	91.9	<.001	13.7	29.1	.133
	L4	**8.565**	1.31	<.001	**1543**	91.75	<.001	16.3	30.5	.117
	L5	**9.49**	1.36	<.001	**1579**	82.6	<.001	17.2	30.55	.113
Small-scale tilt	L1	**2.79**	1.16	.003	**122**	69.2	.012	17.5	28.9	.112
	L2	**9.48**	1.38	<.001	**330**	71.6	.001	**51.5**	30.3	.011
	L3	**46.3**	2.10	<.001	**721.5**	83.5	<.001	**231**	34.6	<.001
	L4	**126**	3.87	<.001	**1050**	81.6	<.001	**754**	35.6	<.001
	L5	**246**	6.32	<.001	**1315**	83.1	<.001	**1701**	55.75	<.001
Large-scale tilt	L1	1.10	1.32	.068	8.06	60.6	.237	12.9	31.7	.150
	L2	**3.58**	1.17	.002	**201**	50.9	<.001	21.2	26.6	.073
	L3	**12.1**	1.21	<.001	**1030**	48.0	<.001	**42.3**	29.5	.023
	L4	**21.8**	1.17	<.001	**1635**	52.4	<.001	**66.0**	28.9	.007
	L5	**30.65**	1.28	<.001	**1648**	51.3	<.001	**91.6**	29.4	.001
Truncation	L1	0.704	1.24	.125	0.770	65.5	.327	15.2	28.25	.131
	L2	1.11	1.19	.056	**100**	58.6	.014	**38.1**	26.0	.022
	L3	**2.19**	1.22	.013	**367**	54.3	<.001	**39.1**	28.2	.025
	L4	**4.51**	1.34	.001	**920**	53.9	<.001	**50.0**	30.7	.013
	L5	**5.485**	1.36	<.001	**1896**	59.8	<.001	**44.55**	32.4	.022

Table 40: **Raw spectral response, RB (Pixels, Spectrum, Spectrum-native).** As Table 39.

		Pixels			Spectrum			Spectrum-native		
disturbance	level	KD	q_{95}	p	KD	q_{95}	p	KD	q_{95}	p
identity	L0	−10.25	101.5	.512	−6716	8.73×10^{4}	.402	−8374	7.29×10^{4}	.434
Low-k injection	L1	30.7	119	.244	−5285	8.515×10^{4}	.366	−8573	7.17×10^{4}	.419
	L2	115	133	.068	−2965	1.00×10^{5}	.356	−8665	8.61×10^{4}	.428
	L3	**241.5**	154	.011	−149	8.665×10^{4}	.322	−8683	7.15×10^{4}	.429
	L4	**412**	201	.003	2935	8.245×10^{4}	.304	−8650	6.81×10^{4}	.423
	L5	**627**	219	.001	6157	9.32×10^{4}	.289	−8580	7.77×10^{4}	.420
Mid-k injection	L1	−2.22	102	.456	−131	9.38×10^{4}	.323	−4325	7.67×10^{4}	.367
	L2	13.9	95.8	.318	7460	8.395×10^{4}	.253	36.9	6.87×10^{4}	.309
	L3	37.0	101.5	.224	1.57×10^{4}	9.44×10^{4}	.229	4498	7.89×10^{4}	.286
	L4	66.4	107	.129	2.425×10^{4}	8.275×10^{4}	.189	8952	6.93×10^{4}	.258
	L5	102	116.5	.066	3.305×10^{4}	8.425×10^{4}	.157	1.33×10^{4}	6.895×10^{4}	.227
High-k injection	L1	−9.80	98.3	.508	1.12×10^{4}	8.32×10^{4}	.242	5491	7.05×10^{4}	.271
	L2	−8.28	103	.480	3.10×10^{4}	8.90×10^{4}	.167	2.02×10^{4}	7.34×10^{4}	.194
	L3	−5.895	98.35	.460	5.16×10^{4}	9.13×10^{4}	.115	3.51×10^{4}	7.40×10^{4}	.137
	L4	−2.80	99.0	.437	7.26×10^{4}	9.19×10^{4}	.068	4.975×10^{4}	7.62×10^{4}	.085
	L5	0.907	103	.412	$\mathbf{9.35\times10^{4}}$	8.54×10^{4}	.045	6.40×10^{4}	6.93×10^{4}	.059
Low-k removal	L1	−6.83	84.8	.506	−6666	9.57×10^{4}	.395	−8004	7.78×10^{4}	.429
	L2	44.6	65.5	.103	−4430	9.24×10^{4}	.362	−7340	7.66×10^{4}	.403
	L3	**153**	56.8	.001	769	8.92×10^{4}	.319	−6087	7.50×10^{4}	.381
	L4	**347**	48.0	<.001	1.08×10^{4}	9.29×10^{4}	.246	−3222	7.98×10^{4}	.358
	L5	**938**	46.2	<.001	$\mathbf{2.81\times10^{5}}$	8.085×10^{4}	.001	$\mathbf{2.86\times10^{5}}$	7.36×10^{4}	.001
Mid-k removal	L1	−8.605	89.9	.467	-1.16×10^{4}	9.58×10^{4}	.444	-1.165×10^{4}	7.70×10^{4}	.482
	L2	5.52	89.2	.369	-1.35×10^{4}	9.61×10^{4}	.468	-1.31×10^{4}	7.96×10^{4}	.497
	L3	37.6	86.2	.179	−8627	9.25×10^{4}	.418	−9514	8.00×10^{4}	.454
	L4	**102**	79.45	.022	1.75×10^{4}	9.34×10^{4}	.231	1.27×10^{4}	7.91×10^{4}	.253
	L5	**351.5**	69.4	<.001	$\mathbf{2.60\times10^{6}}$	1.07×10^{5}	<.001	$\mathbf{4.97\times10^{6}}$	1.08×10^{5}	<.001
High-k removal	L1	−9.31	102.5	.502	-2.10×10^{4}	8.95×10^{4}	.563	-1.97×10^{4}	7.49×10^{4}	.615
	L2	−6.415	96.05	.489	-2.75×10^{4}	9.32×10^{4}	.704	-2.41×10^{4}	8.00×10^{4}	.742
	L3	−0.515	100	.424	-1.40×10^{4}	9.59×10^{4}	.482	−8775	8.01×10^{4}	.432
	L4	11.1	89.7	.327	7.09×10^{4}	8.68×10^{4}	.065	$\mathbf{8.56\times10^{4}}$	7.43×10^{4}	.042
	L5	57.5	87.9	.115	$\mathbf{3.79\times10^{6}}$	1.28×10^{5}	<.001	$\mathbf{9.69\times10^{6}}$	2.04×10^{5}	<.001
Independent phase	L1	−14.55	93.1	.548	−2941	9.74×10^{4}	.353	−8374	7.79×10^{4}	.421
	L2	−25.0	86.2	.685	1.04×10^{4}	8.90×10^{4}	.248	−8374	7.61×10^{4}	.417
	L3	10.4	59.1	.293	3.69×10^{4}	8.74×10^{4}	.142	−8374	7.81×10^{4}	.392
	L4	**99.2**	45.3	.003	$\mathbf{1.11\times10^{5}}$	7.55×10^{4}	.019	−8374	6.45×10^{4}	.400
	L5	**145**	41.5	<.001	$\mathbf{2.81\times10^{5}}$	8.57×10^{4}	.002	−8374	7.09×10^{4}	.398
Shared phase	L1	−14.2	83.5	.529	−2889	9.23×10^{4}	.346	−8374	7.53×10^{4}	.413
	L2	−24.6	88.3	.687	1.06×10^{4}	8.68×10^{4}	.244	−8374	7.22×10^{4}	.421
	L3	3.025	60.4	.367	3.73×10^{4}	8.33×10^{4}	.130	−8374	7.63×10^{4}	.412
	L4	**80.5**	50.4	.009	$\mathbf{1.12\times10^{5}}$	8.72×10^{4}	.032	−8374	7.52×10^{4}	.415
	L5	**126**	40.3	.001	$\mathbf{2.81\times10^{5}}$	8.46×10^{4}	.002	−8374	7.30×10^{4}	.421
Small-scale tilt	L1	−2.56	91.8	.443	$\mathbf{1.35\times10^{5}}$	8.09×10^{4}	.015	$\mathbf{9.60\times10^{4}}$	6.61×10^{4}	.020
	L2	48.4	83.0	.136	$\mathbf{3.75\times10^{5}}$	7.62×10^{4}	.001	$\mathbf{2.565\times10^{5}}$	6.08×10^{4}	.001
	L3	**360**	83.9	<.001	$\mathbf{9.44\times10^{5}}$	8.82×10^{4}	<.001	$\mathbf{6.24\times10^{5}}$	7.12×10^{4}	<.001
	L4	**908**	73.9	<.001	$\mathbf{1.44\times10^{6}}$	8.83×10^{4}	<.001	$\mathbf{9.49\times10^{5}}$	6.96×10^{4}	<.001
	L5	**1393**	62.7	<.001	$\mathbf{1.81\times10^{6}}$	8.02×10^{4}	<.001	$\mathbf{1.21\times10^{6}}$	5.89×10^{4}	<.001
Large-scale tilt	L1	7.025	91.4	.347	-1.80×10^{4}	9.11×10^{4}	.507	-1.19×10^{4}	7.68×10^{4}	.451
	L2	37.2	104	.197	$\mathbf{1.33\times10^{5}}$	8.61×10^{4}	.017	$\mathbf{1.44\times10^{5}}$	7.45×10^{4}	.011
	L3	**111**	102	.042	$\mathbf{1.08\times10^{6}}$	9.49×10^{4}	<.001	$\mathbf{1.21\times10^{6}}$	9.06×10^{4}	<.001
	L4	**182**	100	.009	$\mathbf{2.98\times10^{6}}$	1.18×10^{5}	<.001	$\mathbf{3.59\times10^{6}}$	1.19×10^{5}	<.001
	L5	**244**	101	.002	$\mathbf{5.655\times10^{6}}$	1.47×10^{5}	<.001	$\mathbf{7.10\times10^{6}}$	1.58×10^{5}	<.001
Truncation	L1	−5.935	97.4	.464	$\mathbf{1.82\times10^{5}}$	8.46×10^{4}	.006	$\mathbf{7.58\times10^{5}}$	7.235×10^{4}	<.001
	L2	1.70	94.7	.401	$\mathbf{8.14\times10^{5}}$	7.62×10^{4}	<.001	$\mathbf{3.07\times10^{6}}$	8.29×10^{4}	<.001
	L3	8.495	92.4	.352	$\mathbf{1.895\times10^{6}}$	8.74×10^{4}	<.001	$\mathbf{6.64\times10^{6}}$	1.30×10^{5}	<.001
	L4	18.1	95.45	.297	$\mathbf{3.62\times10^{6}}$	1.21×10^{5}	<.001	$\mathbf{1.08\times10^{7}}$	2.09×10^{5}	<.001
	L5	33.4	95.5	.219	$\mathbf{6.00\times10^{6}}$	1.44×10^{5}	<.001	$\mathbf{1.60\times10^{7}}$	2.835×10^{5}	<.001

Table 41: **Raw spectral response, RB (PCA spaces).** As Table 39.

		PCA-white			PCA-raw		
disturbance	level	KD	q_{95}	p	KD	q_{95}	p
identity	L0	−0.515	3.71	.581	6.78×10^{10}	3.14×10^{11}	.321
Low-k injection	L1	**10.5**	3.80	<.001	2.47×10^{11}	4.49×10^{11}	.143
	L2	**37.1**	4.26	<.001	5.53×10^{11}	5.63×10^{11}	.052
	L3	**77.0**	4.87	<.001	**1.01×10^{12}**	7.96×10^{11}	.024
	L4	**129**	5.93	<.001	**1.64×10^{12}**	1.09×10^{12}	.012
	L5	**192.5**	7.10	<.001	**2.47×10^{12}**	1.38×10^{12}	.007
Mid-k injection	L1	**7.35**	4.15	.001	1.03×10^{11}	3.62×10^{11}	.258
	L2	**29.1**	4.19	<.001	1.47×10^{11}	3.66×10^{11}	.192
	L3	**62.85**	5.24	<.001	1.99×10^{11}	4.01×10^{11}	.165
	L4	**107.5**	5.63	<.001	2.59×10^{11}	4.40×10^{11}	.140
	L5	**163**	7.04	<.001	3.27×10^{11}	4.65×10^{11}	.107
High-k injection	L1	−0.475	3.76	.563	6.84×10^{10}	3.31×10^{11}	.297
	L2	−0.299	3.71	.525	6.90×10^{10}	3.39×10^{11}	.306
	L3	−0.0148	3.47	.484	6.97×10^{10}	3.13×10^{11}	.286
	L4	0.358	3.49	.399	7.04×10^{10}	3.33×10^{11}	.280
	L5	0.804	3.67	.323	7.11×10^{10}	3.29×10^{11}	.306
Low-k removal	L1	**8.33**	3.45	<.001	−7.84×10^{9}	2.38×10^{11}	.473
	L2	**45.7**	3.40	<.001	−3.48×10^{9}	1.52×10^{11}	.460
	L3	**131**	4.01	<.001	5.71×10^{10}	1.18×10^{11}	.190
	L4	**323**	6.57	<.001	**1.50×10^{11}**	8.82×10^{10}	.011
	L5	**1382**	33.5	<.001	**2.85×10^{11}**	8.53×10^{10}	<.001
Mid-k removal	L1	**9.07**	3.07	<.001	4.11×10^{10}	2.83×10^{11}	.331
	L2	**43.1**	3.25	<.001	2.41×10^{10}	2.61×10^{11}	.353
	L3	**117**	3.75	<.001	1.85×10^{10}	2.23×10^{11}	.392
	L4	**276**	5.93	<.001	2.87×10^{10}	1.86×10^{11}	.329
	L5	**1094**	26.0	<.001	9.78×10^{10}	1.26×10^{11}	.087
High-k removal	L1	−0.376	3.66	.532	6.72×10^{10}	3.23×10^{11}	.314
	L2	0.0155	3.58	.465	6.67×10^{10}	3.06×10^{11}	.280
	L3	0.802	3.44	.315	6.63×10^{10}	2.96×10^{11}	.310
	L4	2.35	3.48	.127	6.60×10^{10}	3.09×10^{11}	.297
	L5	**8.56**	3.55	<.001	6.66×10^{10}	3.13×10^{11}	.296
Independent phase	L1	−0.426	3.43	.569	4.68×10^{10}	3.02×10^{11}	.321
	L2	**4.81**	3.37	.013	−2.06×10^{10}	2.32×10^{11}	.502
	L3	**52.6**	4.12	<.001	−9.64×10^{9}	1.17×10^{11}	.505
	L4	**124**	7.29	<.001	**9.24×10^{10}**	8.99×10^{10}	.046
	L5	**150**	8.36	<.001	**1.28×10^{11}**	8.31×10^{10}	.010
Shared phase	L1	−0.282	3.48	.517	4.65×10^{10}	2.82×10^{11}	.316
	L2	**6.57**	3.45	.002	−2.065×10^{10}	2.315×10^{11}	.504
	L3	**67.55**	4.68	<.001	−1.64×10^{10}	1.28×10^{11}	.529
	L4	**159**	8.45	<.001	6.88×10^{10}	1.01×10^{11}	.099
	L5	**194**	10.5	<.001	**1.03×10^{11}**	8.17×10^{10}	.028
Small-scale tilt	L1	**14.6**	3.71	<.001	4.79×10^{10}	2.915×10^{11}	.333
	L2	**66.0**	4.15	<.001	3.20×10^{10}	2.70×10^{11}	.361
	L3	**266**	8.095	<.001	3.32×10^{10}	2.25×10^{11}	.354
	L4	**424**	14.1	<.001	1.10×10^{11}	1.56×10^{11}	.103
	L5	**424**	21.8	<.001	**2.015×10^{11}**	1.03×10^{11}	.005
Large-scale tilt	L1	**13.6**	3.37	<.001	8.88×10^{10}	3.07×10^{11}	.245
	L2	**49.8**	3.70	<.001	1.09×10^{11}	3.57×10^{11}	.230
	L3	**167**	5.74	<.001	1.44×10^{11}	3.40×10^{11}	.212
	L4	**322**	8.80	<.001	1.67×10^{11}	3.28×10^{11}	.161
	L5	**499**	13.4	<.001	1.81×10^{11}	3.09×10^{11}	.135
Truncation	L1	−0.409	3.54	.553	6.74×10^{10}	3.08×10^{11}	.296
	L2	−0.128	3.85	.489	6.70×10^{10}	3.26×10^{11}	.294
	L3	0.386	3.63	.402	6.67×10^{10}	3.13×10^{11}	.280
	L4	1.27	3.58	.244	6.64×10^{10}	3.07×10^{11}	.283
	L5	3.29	3.45	.061	6.63×10^{10}	3.24×10^{11}	.303

Table 42: **Raw spectral response, RB (Random init).** As Table 38, for the spectral suite (87 versus 88 ICs per arm).

		Random init		
disturbance	level	KD	q_{95}	rej.
identity	L0	109	424.5	0/3
Low-k injection	L1	**964.5**	636	3/3
	L2	**2889**	885	3/3
	L3	**6467**	1610	3/3
	L4	**1.245×10^4**	2381	3/3
	L5	**2.18×10^4**	3638	3/3
Mid-k injection	L1	**624**	515.5	3/3
	L2	**1567**	691	3/3
	L3	**3078**	796	3/3
	L4	**5331**	1118	3/3
	L5	**8527**	1462	3/3
High-k injection	L1	187	413	0/3
	L2	280.5	432	0/3
	L3	393	478.5	1/3
	L4	**526**	491	2/3
	L5	**681**	502	3/3
Low-k removal	L1	−105	290	0/3
	L2	24.7	234	0/3
	L3	**304**	215	3/3
	L4	**620**	199	3/3
	L5	**927**	209	3/3
Mid-k removal	L1	−95.15	318	0/3
	L2	−85.0	275	0/3
	L3	62.55	264	0/3
	L4	**283**	217	2/3
	L5	**475**	220.5	3/3
High-k removal	L1	46.8	356	0/3
	L2	−2.70	383	0/3
	L3	−40.8	413.5	0/3
	L4	−68.6	354.5	0/3
	L5	−84.3	363	0/3
Independent phase	L1	112	426	0/3
	L2	123	396	0/3
	L3	152	432	0/3
	L4	170	410	0/3
	L5	170	435	0/3
Shared phase	L1	109	399	0/3
	L2	111	398	0/3
	L3	128	397	0/3
	L4	150	398	0/3
	L5	154	442	0/3
Small-scale tilt	L1	529	499	1/3
	L2	**1834**	675.5	3/3
	L3	**1.91×10^4**	3479	3/3
	L4	**1.41×10^5**	2.17×10^4	3/3
	L5	**6.15×10^5**	7.20×10^4	3/3
Large-scale tilt	L1	−34.1	366	0/3
	L2	−84.3	358	0/3
	L3	−104	342	0/3
	L4	−99.3	340	0/3
	L5	−92.8	351	0/3
Truncation	L1	103	382	0/3
	L2	88.6	395	0/3
	L3	65.6	410	0/3
	L4	23.35	411	0/3
	L5	−26.25	393	0/3

F IC-CONDITIONAL LAWS: FULL RESULTS

Section 4.2 reports the power of an IC-level permutation test for each (Ra, Pr) coefficient pair, called a *group* below and in the tables. Appendix F.1 gives that test's design, its size and a physical reference. Appendix F.2 reports a second, pre-registered probe of the same 1,750 trajectories, a single-trajectory exceedance rate, averaged over groups and over coefficient adjacencies (Tables 43 and 44). Appendix F.3 reports a related analysis on the turbulent radiative layer, and Appendix F.4 describes the structure of the RB reference itself.

F.1 THE PERMUTATION TEST OF SECTION 4.2

IC-memory index. This index ranks the 35 groups by how well a single frame determines ΔT_{init}. A network trained from scratch on RB alone regresses ΔT_{init} from one frame; it is trained on five of the ten noise seeds and applied to every frame (0–199) of the other five. A group's error is the mean absolute error over those frames and trajectories, averaged over two independently trained networks. The index ranks the groups by this error (1: lowest error, longest IC memory; 35: highest error, shortest IC memory). It uses the continuous error rather than the accuracy (a prediction within 0.1 of the true ΔT_{init}), which is 1.0 for more than half of the groups. The index should not be read as a measure of chaos: over the 35 groups its Spearman correlation with $\log_{10}(\mathrm{Ra}\,\mathrm{Pr})$ is -0.965. Its ordering closely follows that of the solver's thermal diffusivity $(\mathrm{Ra}\,\mathrm{Pr})^{-1/2}$, consistent with diffusion relaxing the heat content set by ΔT_{init}; we did not test this mechanism. The group with the shortest memory ($\mathrm{Ra} = 10^6$, $\mathrm{Pr} = 0.1$) has the grid's smallest $\mathrm{Ra}\,\mathrm{Pr}$, i.e. its largest thermal diffusivity.

Design. For each group, pair of initial gradients and 20-frame window (frames 0–19, …, 180–199), one arm holds five trajectories at one gradient and the other arm five trajectories at the second gradient. The arms use disjoint noise seeds, because the RB grid reuses seed indices across gradients and a shared seed would pair the arms. The statistic is the IC-level KD of Section 3.2; the test enumerates all $\binom{10}{5} = 252$ relabelings and rejects at $p \le 0.05$. A comparison's power is the fraction of the 252 seed splits (every choice of five of the ten seeds for the first arm) that reject, and a group's power in a window is the mean over its ten gradient pairs. The max distinguishable frame is the last frame of the last window before power first falls below 0.8 after first reaching it, or frame 199 if it never falls. All 252 splits reuse the same 20 trajectories, so a group's power, and hence its 0.8 crossing, describes this pool of trajectories and carries no sampling interval. As for the exceedance probe (Appendix F.2), most trajectories come from the RB training split (12 to 20 of each comparison's 20; 80% of all 1,750), and no held-out control was run for this design. We also did not run a perturbed-twin test, which would restart the solver from a perturbed copy of one IC and score its footprint against seed-to-seed variability.

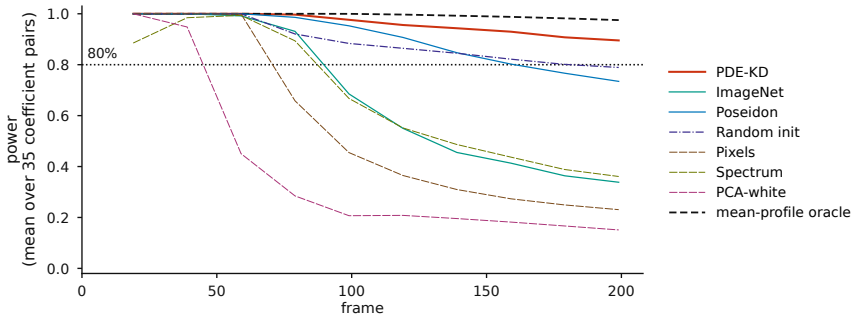


Figure 10: **Permutation power of Section 4.2, averaged over the 35 groups**, per 20-frame window (x: the window's last frame), for six spaces and Random init. Random init: the untrained PDE-KD network, mean over three initialization seeds (Appendix C.2). Dashed black line: the same test on the horizontal-mean buoyancy profile (mean-profile oracle); dotted black line: 80% power.

Size and censoring. KD is symmetric in the two arms, so the 252 relabelings take 126 distinct values and the smallest attainable p is $2/252$. The test rejects when at most 12 relabelings (the observed one and its mirror included) reach the observed statistic, so its size is $12/252 = 0.048$ by construction, in every space. The max distinguishable frame is right-censored at frame 199. Section 4.2 gives how many groups reach the end of the record for PDE-KD, Poseidon and Random init. Of the other spaces, only Pixels does, at 2 of the 35 groups. PDE-KD's max frame is strictly later than that of every other space at 11 groups and ties at least one space at the other 24.

Power at the extremes of IC memory. At the group with the longest IC memory, PDE-KD, Poseidon and Random init keep power 1.00 through the final window, while ImageNet, Spectrum, Pixels and PCA-white end at 0.42–0.64. At the group with the shortest, no representation keeps 80% power past frame 79, and all end at or below 0.034. At $\mathrm{Ra} = 10^7$, $\mathrm{Pr} = 0.1$, PDE-KD's final power is 0.38 and still falling, so the 200-frame record does not resolve the asymptotic behavior.

A physical reference. The same test applied to the horizontal-mean buoyancy profile $b(z)$ of each frame (128 values, standardized per height with one mean and standard deviation pooled over all 1,750 trajectories) keeps at least 80% power to frame 199 at 34 of the 35 groups, with mean final-window power 0.975. At the shortest-memory group it keeps 80% power until frame 119 and ends at 0.343, where PDE-KD keeps it until frame 79 and ends at 0.034. The ΔT_{init} difference therefore persists in the mean temperature profile throughout the record at 34 of the 35 groups. At those groups, a representation whose power falls no longer resolves a difference that these trajectories still carry, so its low power does not indicate converged populations. Averaged over groups (Figure 10), the final-window power is 0.895 for PDE-KD, 0.789 for Random init, 0.734 for Poseidon, 0.361 for Spectrum, 0.338 for ImageNet, 0.231 for Pixels and 0.151 for PCA-white. This average mixes groups that lose power within the record with groups that keep it (Figure 3).

F.2 The single-trajectory exceedance probe

Design counts. RB has 35 (Ra, Pr) groups $\times$ 5 initial temperature gradients $\Delta T_{\mathrm{init}} \in \{0.2, \ldots, 1.0\} \times$ 10 seeds, which gives 1,750 trajectories across all splits. Below, a *cell* is one $(\mathrm{Ra}, \mathrm{Pr}, \Delta T_{\mathrm{init}})$ combination, so each group holds five cells. The Well does not record the pair (ΔT_{init}, seed). We recovered it for all 1,750 trajectories by matching each trajectory's first frame against regenerated initial conditions (the generator's seeded noise plus a fitted ΔT_{init}; worst match correlation 0.999994, worst runner-up 0.0094). Per group and 20-frame window, the null pools $5 \times 45 = 225$ same-cell seed pairs. The cross-IC alternative uses 1,000 cross-ΔT_{init} pairs per group (35,000 per window). The coefficient controls compare cells one grid step apart at equal ΔT_{init}: 28 adjacencies in Ra (one decade) and 30 in Pr (0.301 or 0.398 decades).

Estimator and threshold. Each arm of this probe contains one 20-frame trajectory window, so the same-IC-excluded trajectory-level estimator is degenerate. We instead use the frame-level cubic-kernel U-statistic. Frames within a window are temporally dependent, so the unbiasedness that this statistic has for independent frames does not carry over, and the threshold comes from an empirical null instead. A cross-IC pair is judged against its group's 225-pair null pool; a coefficient-control pair, whose two cells lie in two groups, against the union of both groups' pools (450 values). The threshold is the pool's 95th percentile, so it is exceeded by 12 of the 225 null pairs themselves: the rule's structural false-positive rate is $12/225 = 0.053$, not an estimated one. The probe is thus a per-pair threshold rule, not a permutation test, and its power is an exceedance rate rather than the replicate power of Section 4.1. The arm and window geometry is identical in every space, so the spaces can still be compared with each other. The null pairs are not independent: each seed appears in 9 of its cell's 45 pairs, and a group pool reuses 50 ICs across 5 cells. The threshold's sampling error is therefore larger than that of 225 independent values; we report this dependence and do not correct for it. The rate $12/225$ is a property of the rule, not a clean-versus-clean check, and this probe has no such check. Power is reported per group and averaged over groups or over adjacencies, and the Ra and Pr controls are kept separate because their grid steps differ in size.

Group mean. Table 43 gives, per window, the *mean* power over the 35 groups, the summary that the pre-registered analysis reported. Because every group carries 1,000 pairs, this mean equals the power pooled over all alternative pairs. By the mean, PDE-KD holds power 0.8 longest: its mean

Table 43: **Cross-IC exceedance power per window, mean over groups** (pooled over all alternative pairs), all eight frame spaces and Random init, with the horizon: the first window whose power falls below .80 after reaching it. Bold: $\geq .80$. The per-axis coefficient controls are in Table 44. Random init: mean over three initialization seeds.

	learned / untrained control				non-learned				
frames	PDE-KD	ImageNet	Poseidon	Random init	Pixels	Spectrum	Spectrum-native	PCA-white	PCA-raw
0–19	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	.681	.672	**1.000**	**1.000**
20–39	**1.000**	**.987**	**1.000**	**1.000**	**1.000**	**.912**	**.916**	.399	**.966**
40–59	**1.000**	**.915**	**.996**	**.990**	.419	**.911**	**.884**	.006	.099
60–79	**.946**	.713	**.831**	.735	.048	.675	.736	.017	.004
80–99	**.848**	.386	.779	.651	.076	.387	.401	.034	.014
100–119	**.846**	.294	.747	.663	.086	.334	.334	.041	.018
120–139	**.811**	.241	.687	.641	.088	.270	.265	.045	.023
140–159	.782	.203	.631	.599	.082	.220	.219	.045	.026
160–179	.756	.155	.603	.563	.078	.187	.187	.050	.025
180–199	.729	.133	.568	.560	.068	.169	.178	.051	.017
horizon	140–159	60–79	80–99	60–79	40–59	60–79	60–79	20–39	40–59

first falls below 0.8 at frames 140–159 and is 0.729 at 180–199. Its group-bootstrap 95% intervals include 0.8 from 140–159 through 180–199 (0.618–0.834 at 180–199). Counted from the window where each curve first reaches 0.8, the next-longest horizon is Poseidon's (80–99); the horizons of the other spaces are in Table 43. At the final window, Random init's mean lies just below Poseidon's (0.56 against 0.57). For PDE-KD, the median over groups exceeds this mean (0.870 against 0.729 at 180–199), because its power is bimodal across groups: it never drops below 0.8 within the record at 19 of the 35 groups.

Per group. Over frames 0–59, PDE-KD's power is 1.00 at all 35 groups. At $\mathrm{Ra} \leq 10^7$ and $\mathrm{Pr} \leq 0.2$ (four groups), its mean falls to 0.08 by frames 180–199, and every space ends within 0.03 of the null rate. There, the mean first drops below 0.10 at frames 180–199 for PDE-KD, 140–159 for Poseidon, 120–139 for Random init, 80–99 for ImageNet and Spectrum, 60–79 for Pixels and 40–59 for PCA-white. At $\mathrm{Pr} \geq 5$ (ten groups), PDE-KD's mean is at least 0.96 in every window and each group ends at 0.98 or above. Poseidon, ImageNet, Spectrum, Pixels and PCA-white end there at 0.82, 0.19, 0.28, 0.18 and 0.07. Random init stays at least 0.89 in every window there and ends at 0.98 (each group 0.91 or above). PDE-KD's other 21 groups end between 0.13 and 1.00. Power falls first at the lowest Ra and Pr, the groups with the shortest IC memory and the largest thermal diffusivity (Appendix F.1). Diffusivity alone does not order the grid, since at equal $\mathrm{Ra}\,\mathrm{Pr}$ the higher-Pr group ends with at least as much power. For Pixels and PCA-white the rate can fall below $12/225$ (PCA-white's mean at $\mathrm{Pr} \geq 5$ is 0.01 at frames 40–59), so for them it is not a monotone measure of separation.

Coefficient controls. The coefficient controls of Table 44 move one grid step in Ra or Pr at equal ΔT_{init}. Adjacencies share groups, so their bootstrap intervals are approximate and likely too narrow. PDE-KD separates a one-step change of Pr with mean power 0.25 over the 30 adjacencies in the first window and 0.71 at frames 20–39. At the final window, ImageNet and Spectrum lead on adjacent Ra (0.90 and 1.00; PDE-KD 0.57, Random init 0.41), and Random init, Spectrum and PDE-KD are level on adjacent Pr (0.77, 0.76 and 0.75). PDE-KD's final-window lead is therefore on the cross-ΔT_{init} probe, not on the one-step coefficient controls. Spectrum-native, which Table 44 omits, reaches power 1.000 on adjacent Ra in every window from frames 20–39 on (28 of 28 adjacencies), as Spectrum does. For the spectra, a rate of 1.000 is also what an under-dispersed same-cell null would produce, and no held-out calibration was run.

Further controls. (i) *Training exposure*: 80% of the trajectories come from the RB training split, on which the PDE-KD feature map and the PCA basis were fit. Restricted to the 1,277 pairs with both trajectories held out, every space's power, pooled over those pairs, changes by at most 0.037 in any window (0.038 averaged over groups), and the ordering of PDE-KD and Poseidon is unchanged. The thresholds still come from the full null pool, which contains training-split trajectories. (ii) *Graded contrast*: PDE-KD's KD rank-correlates with $|\Delta\Delta T_{\mathrm{init}}|$ at $+0.73$ at frames 40–59 and $+0.48$ at 180–199 ($+0.78$ in the IC-containing window 0–19). Poseidon's correlation is higher

Table 44: **Coefficient controls: exceedance power, mean over adjacencies** (28 in Ra, 30 in Pr), six spaces and Random init. 95% CI (last window): bootstrap over adjacencies, 1,000 draws, approximate. Horizon as in Table 43; >199: not below .80 within the record; never reached: the mean never reaches .80. Bold: $\geq .80$. Random init: mean over three initialization seeds (its CI: the mean of the seeds' bounds).

frames	PDE-KD	ImageNet	Poseidon	Random init	Pixels	Spectrum	PCA-white
adjacent Ra *(same* Pr, ΔT*)*							
0–19	.650	**.876**	.250	**.800**	.707	**.994**	**.800**
20–39	**.999**	**1.000**	.696	**.990**	.415	**1.000**	.171
40–59	**.993**	**1.000**	.671	.625	.040	**1.000**	.027
60–79	**.853**	**1.000**	.363	.265	.049	**1.000**	.030
80–99	.494	**.998**	.352	.338	.044	**1.000**	.038
100–119	.526	**.981**	.388	.400	.038	**1.000**	.038
120–139	.541	**.933**	.356	.416	.034	**1.000**	.041
140–159	.559	**.925**	.321	.403	.035	**1.000**	.041
160–179	.574	**.898**	.311	.393	.036	**1.000**	.043
180–199	.570	**.897**	.290	.406	.035	**1.000**	.038
95% CI, 180–199	.489–.663	.817–.954	.227–.368	.293–.531	.027–.042	1.000–1.000	.030–.048
horizon	80–99	>199	never reached	40–59	never reached	>199	20–39
adjacent Pr *(same* Ra, ΔT*)*							
0–19	.250	.658	.200	.605	.484	.165	.715
20–39	.713	.486	.462	.737	.250	.492	.207
40–59	.507	.359	.143	.344	.046	.564	.042
60–79	.363	.393	.073	.246	.045	.616	.039
80–99	.311	.237	.081	.496	.037	.589	.035
100–119	.509	.268	.120	.679	.027	.686	.032
120–139	.590	.289	.142	.766	.024	.760	.029
140–159	.667	.321	.141	.765	.025	.769	.026
160–179	.711	.321	.152	.750	.022	.767	.028
180–199	.750	.313	.153	.769	.019	.756	.024
95% CI, 180–199	.661–.853	.214–.429	.124–.186	.677–.857	.012–.026	.672–.838	.019–.029
horizon	never reached	never reached	never reached	never reached	never reached	never reached	never reached

in every reported window ($+0.82$ and $+0.61$ at the same windows), so the persistence cannot be attributed to coefficient supervision alone.

F.3 Unseen physics: the turbulent radiative layer

The Well's turbulent radiative layer (TRL2D) is not in the training corpus. It has 9 cooling times with the same 10 ICs in each, which permits exactly paired coefficient contrasts. We ran an analogous design on all splits, with paired coefficient contrasts but without an IC-family probe. There Spectrum separates cooling times best (unpaired power 0.90 at frames 0–19 and 0.85 at 80–99, versus 0.74 and 0.62 for PDE-KD). By single-frame nearest-neighbor recovery of IC identity across cooling times (chance 0.10), the two physics-trained spaces lose IC identity fastest over frames 1–10 and Pixels keeps it longest. At frame 5, recovery is 0.44 for PDE-KD and 0.41 for Poseidon, against 0.58 for Spectrum, 0.97 for ImageNet and 1.00 for Pixels; at frame 20, Pixels still recovers 0.70. This probe contrasts IC identity across a coefficient change, with ICs that differ only by a small band-limited velocity perturbation, whereas RB's IC families differ in their initial temperature gradient at fixed coefficients. A single scalar, the domain-mean pressure, tracks $\log t_{\mathrm{cool}}$ at $|\rho| \geq 0.95$ and orders the cells better than any KD, so on this dataset the coefficient axis is confounded with a bulk scalar.

F.4 Structure of the RB reference

Same-cell structure. In this subsection a *coefficient cell* is a group, one (Ra, Pr) pair, with its five test ICs. The five test ICs of a cell mostly differ in ΔT_{init} (289 of the 350 same-cell pairs), so the rejection rate below includes the ΔT_{init} differences of Appendices F.1 and F.2 as well as any multistability. The roll-number contrast below is not stratified by ΔT_{init}. In the PDE-KD space, ground truth at a fixed coefficient cell is not a single mode. A test of equality between two ICs of one coefficient cell rejects at $p \leq 0.05$ for 52.6% of pairs, more often than for cells one Rayleigh decade apart (37.9%). These single-trajectory tests relabel frames, which are autocorrelated within a trajectory, so they are not exact: the two rates are comparable with each other but not with 0.05. Within each trajectory, KD between the first and second thirds of frames 61–199 is $5.4\times$ that between the second

and third (KD $\times 10^3$: 21.3 versus 3.9 at the same lag), consistent with a slow transient. The dominant roll wavenumber differs between ICs of the same cell in 24 of 35 cells, and pairs with different roll numbers carry more KD (KD $\times 10^3$: 29.4 versus 16.3; within-cell permutation $p = 0.0024$). The reference at fixed (Ra, Pr) is therefore multimodal, with roll-number multistability in a large minority of ICs, so a KD gap against pooled reference ICs is not automatically model error.

Separation of the coefficient grid. Between coefficient cells, ImageNet rejects more cell pairs than PDE-KD. We pooled each cell's five test ICs (frames 61–199) and tested every pair of the 35 cells by exact permutation. ImageNet rejects equality for 98% of the 595 cell pairs at $p \leq 0.05$, against 66% for PDE-KD and 56% for Poseidon, and the cell nearest in KD is a grid neighbor for 97%, 83% and 57% of cells. Along each axis, ImageNet's KD tracks grid distance about as closely as PDE-KD's (Spearman correlation 0.91 along Ra and 0.82 along Pr, against 0.88 and 0.82), but on the joint grid less closely (0.60 against 0.85). Its rejection and nearest-neighbor rates on this in-distribution grid are nevertheless the highest of the three spaces.

G Physical diagnostics and model ladders

G.1 Model roster

The RB physical-alignment roster has nine models, scored on the 175 test ICs:

- Lola latent diffusion models at three latent compressions (Rozet et al., 2025), with 535M parameters each on RB. Their labels count latent channels: on RB, Lola-c4, c16 and c64 compress $1024\times$, $256\times$ and $64\times$ (on Euler, $1280\times$, $320\times$ and $80\times$), so c4 is the most compressed. Their parameter counts differ by 0.24% on RB (0.28% on Euler), so their ladder is a pure compression axis.
- Walrus (McCabe et al., 2026), pretrained on a mixture of Well and FlowBench (Tali et al., 2024) scenarios (Walrus-PT) and fine-tuned on the target physics (Walrus-FT; 1.29B parameters). Walrus does not condition on the PDE coefficients (Lola does), so it infers them from the conditioning frames. It is not deterministic: its patch jitter stays active at evaluation, and on RB the difference between two seeds is 45% of its error against the reference, so every Walrus score carries unquantified draw variance. Two RB draws of each Walrus model appear in this appendix. Appendices G.3, G.4 and G.5 use one draw, and Table 45 and the model zoo of Appendices D.5 and G.7 use the other. For Walrus-FT, vRMSE is 0.950 in the first draw and 0.962 in the second. In the PDE-KD space, paired KD $\times 10^3$ is 2.40 against 2.36 for Walrus-FT and 67.5 against 67.27 for Walrus-PT.
- four Well operators (Ohana et al., 2024) trained from scratch on RB for 32 epochs with the Well's recipe: FNO (Li et al., 2021), TFNO (Kossaifi et al., 2024), U-Net (UNetClassic; Ronneberger et al., 2015) and ConvNeXt-U-Net (UNetConvNext; Liu et al., 2022), with 17.5M–19.3M parameters each. All use the final epoch of their first cycle, with no rollout-based selection.

The 11-checkpoint model zoo of Appendices D.5 and G.7 adds the second and third warm-restart cycles of FNO to these nine models. All RB rollouts share one conditioning anchor, native frame 16, so lead 1 is native frame 17. The number of conditioning frames differs by model: Lola 1, Walrus-PT 10, Walrus-FT 6. On RB, KD scores two stored members per IC for each Lola model and one rollout per IC for Walrus and the Well operators. The exception is Table 51, which scores the first Lola member only (Appendix G.5). CRPS uses all 16 members of the Lola ensembles. On Euler, KD scores one member per IC. The Well operators were trained on RB only, so neither Euler ladder includes them. The Euler-open ladder has five models (the three Lola models, Walrus-PT and Walrus-FT) on a common set of 500 ICs. On Euler, Lola advances four native frames per step, so KD uses the leads 64, 68, …, 88 that every model reaches. Unless stated otherwise, a model's errors over the window from lead 61 average its own leads: 64–88 for Lola and 61–91 for Walrus. Table 45 gives the RB ladder under every metric. Its Walrus spectral columns come from the first Walrus draw, because the second records no spectra on the full IC pool (none for Walrus-PT, 35 ICs for Walrus-FT). Four of its KD values (Lola-c4, c16, c64 and Walrus-FT) are unresolved estimates; the paired contrasts of Appendix G.7 establish the ordering of these models.

Well baselines. The released checkpoints of these operators reproduce neither the published VRMSE nor persistence on RB, because of upstream changes to a normalization flag and to the data orientation, and none exists for Euler. We therefore retrained them with the Well's own training code. The budget differs: we train for 32 epochs, whereas the published RB runs reached 32, 31, 29 and 12 epochs within their 12-hour limit (Ohana et al., 2024, Table 6). Each model uses its model card's learning rate for RB (10^{-4}; 5×10^{-4} for UNetConvNext). The checkpoint rule is ambiguous upstream: the paper selects hyperparameters on validation, and the released trainer scores the final model, so we report both. On the Well's one-step VRMSE over 1,024 windows of the test split, our final-epoch checkpoints score 0.52 (FNO), 0.76 (TFNO), 0.52 (U-Net) and 0.17 (ConvNeXt-U-Net) (standard errors 0.04, 0.06, 0.04, 0.01), and our best-validation checkpoints 0.91, 0.89, 0.86 and 0.25 (0.06, 0.07, 0.06, 0.02). The published test-set values are 0.84, 0.66, 1.49 and 0.67 (Ohana et al., 2024, Table 2) and carry no error bars. Under either rule, both U-Nets score better and TFNO worse. FNO scores better at the final epoch and about one standard error worse at the best-validation checkpoint (0.91 against 0.84).

Euler ladder values. In the PDE-KD space (KD $\times 10^3$, paired, 500 ICs, leads 64–88), the periodic-boundary values are those of Table 50, against a split-half floor of 0.55 (s.d. 0.94). On open boundaries, Lola-c64, c16 and c4 score 0.04, 0.06 and 0.45, within one floor s.d. of the floor (-1.45, s.d. 2.08). Walrus-FT and Walrus-PT score 3.48 and 42.0. On open boundaries, the PDE-KD point estimates of the Lola models follow the expected order but are not separated by magnitude, and every other space scored there inverts at least one Lola pair. ImageNet, whose Lola scores are resolved from its floor, gives Lola-c64 3.48 and Lola-c16 3.36.

Table 45: **RB model ladder**, leads 61 onward, 175 test ICs. KD $\times 10^3$: regime-conditional (paired, same-IC-excluded, averaged over the 35 (Ra, Pr) cells), on individual realizations (two per IC for the ensemble models). vRMSE: on the ensemble mean. CRPS (fair): on all members of the ensemble models. b_0–b_3: spectral band errors from large to small scales. Bold: column minimum. The KD values of Lola-c4/c16/c64 and Walrus-FT are unresolved under the resolution flag (mean $-$ 1 s.d. across the 35 regimes > 0).

				spectral band error			
model	KD / 10^{-3}	vRMSE	CRPS	b_0	b_1	b_2	b_3
ensemble, K=16; 2 members/IC persisted for KD							
Lola-c4	4.51	**0.659**	**0.344**	**0.360**	0.247	0.311	0.898
Lola-c16	1.73	0.678	0.353	0.377	0.182	0.245	0.511
Lola-c64	**1.00**	0.704	0.367	0.390	**0.169**	**0.217**	0.277
deterministic							
Walrus-PT	96.69	0.980	–	5.686	0.963	0.521	**0.232**
Walrus-FT	12.41	0.962	–	1.753	1.488	0.567	0.738
FNO	283.37	1.367	–	13.284	2.075	114.070	236.219
TFNO	124.87	1.398	–	9.333	13.253	104.572	74.621
UNetClassic	90.21	1.700	–	10.679	124.736	92.206	4.344
UNetConvNext	68.97	1.617	–	7.882	180.040	100.753	2.955

Training budget. The only training-budget comparison in our RB data is the FNO warm-restart cycles. Their regime-conditional KD $\times 10^3$ (the estimator of Table 45) at leads 61 onward is 283.4, 140.6 and 174.1 after cycles 1, 2 and 3. All three are resolved under the resolution flag, and the sequence is not monotone. TFNO's later cycles diverged and are excluded.

G.2 Physical observables

This subsection defines the physical diagnostics of Appendices G.3, G.4 and G.5, and the two ways we reduce them over leads.

Rayleigh–Bénard (7 functionals, 37 components). The Nusselt number is $\mathrm{Nu} = 1 + \langle v_z b \rangle / (\kappa \Delta B / L_z)$ with $\kappa = (\mathrm{Ra}\,\mathrm{Pr})^{-1/2}$, where z is the wall-normal coordinate (stored as y in the Well files). The vertical average uses Chebyshev–Gauss quadrature on the solver's wall-clustered grid; on one reference frame, an unweighted row mean mis-states Nu by 12%. The other functionals are the horizontally averaged heat-flux and temperature profiles (16 height bins each), mean kinetic-energy density, enstrophy, divergence RMS, and wall-boundary violation. The last two have nonzero

reference floors on the discrete data, so errors are paired against the reference, not measured from zero.

Euler (15 scalars). The scalars are mass; two momentum components; total, kinetic and internal energy; entropy; equation-of-state (EOS) residual (median and RMS); unphysical fraction (cells with non-positive density or pressure); minimum density; minimum pressure; density total variation; shock fraction; and enstrophy. The reference EOS residual has median 2.6×10^{-8}, so the absolute EOS error is effectively the generated residual. Table 46 gives the formulas.

Table 46: **Physical observables.** $\langle\cdot\rangle$: area average (RB: Chebyshev–Gauss quadrature in z); $\sum \cdot\, dA$: plain sum over cells with $dA = h^2$, $h = 1/511$ (Euler). RB: b buoyancy, v velocity, $\kappa = (\mathrm{Ra}\,\mathrm{Pr})^{-1/2}$, $\Delta B = L_z = 1$. Euler: ρ density, m momentum, E energy, p pressure, $u = m/\rho$, $p_{\mathrm{eos}} = (\gamma - 1)(E - |m|^2/2\rho)$.

System	Observable	Definition
RB	Nusselt number	$1 + \langle v_z b\rangle / (\kappa\, \Delta B / L_z)$
	heat-flux / temperature profiles	$\langle v_z b\rangle_x(z)$, $\langle b\rangle_x(z)$ in 16 height bins
	kinetic energy	$\frac{1}{2}\langle v_x^2 + v_z^2\rangle$
	enstrophy	$\langle \omega^2\rangle$, $\omega = \partial_x v_z - \partial_z v_x$
	divergence RMS	RMS of $\partial_x v_x + \partial_z v_z$ (nonzero on the discrete reference)
	wall-boundary violation	$\langle |b(0) - 1|\rangle_x + \langle |b(L_z)|\rangle_x + \langle |v|(0)\rangle_x + \langle |v|(L_z)\rangle_x$
Euler	mass, momentum, energy	$\sum \rho\, dA$, $\sum m\, dA$, $\sum E\, dA$ (invariant only on periodic boundaries)
	kinetic / internal energy	$\sum_{\rho>0} |m|^2/(2\rho)\, dA$; $\sum p/(\gamma - 1)\, dA$
	entropy	$\sum_{\rho>0, p>0} \rho(\ln p - \gamma \ln \rho)\, dA$
	EOS residual (median / RMS)	$\mathrm{median}\, |p_{\mathrm{eos}} - p|/|p|$; $\mathrm{rms}(p_{\mathrm{eos}} - p)/\,\mathrm{rms}(p)$
	unphysical fraction	fraction of cells with $\rho \le 0$ or $p \le 0$
	minimum density / pressure	per-frame minimum
	density total variation	$(\sum |\Delta_x \rho| + \sum |\Delta_y \rho|)\, h$
	shock fraction	fraction of cells with $|\nabla \rho|\, h/|\rho| > 0.1$
	enstrophy	$\sum (\partial_x u_y - \partial_y u_x)^2\, dA$ over $\rho > 0$

We use two reductions. The signed *bias* is the difference of window means, and the *absolute error* is the per-lead absolute difference averaged over the window. The position of the absolute value can change the estimand by $20\times$: in one example, Nusselt on the reference solver (Dedalus) restarted from the same frames gives 1.19 versus 24.42. The per-lead error is large even for this restart once trajectories decorrelate: its per-lead Nusselt error reaches 45% of the reference level at $\mathrm{Ra} = 10^{10}$ in the late window. Divergence RMS needs care. As a finite difference of a spectrally divergence-free field, it also measures grid-scale content: on the reference, its ratio to the RMS of $\partial_x v_x$ is 0.002 at $\mathrm{Ra} = 10^6$ and 0.275 at 10^{10}, so a smoother rollout scores as more incompressible. We therefore do not base a ladder conclusion on divergence RMS alone, although it carries PDE-KD's largest pooled correlation (Appendix G.4).

G.3 Agreement with physical diagnostics

This subsection asks whether KD tracks physical error across the fixed roster of trained simulators of Appendix G.1. No checkpoint was trained or selected to optimize the reported evaluations. For each physical diagnostic, we rank the models by their absolute error against the reference. A model's family score is its mean rank over the diagnostics of a predefined family. Each entry of Table 47 is the Spearman correlation between KD and that mean rank, so it asks whether KD tracks the family's diagnostics in aggregate.

The RB family has five scalars: the Nusselt number (heat transport), mean kinetic-energy density and enstrophy (flow intensity and vortical content), and divergence RMS and wall-boundary violation (incompressibility and the imposed boundary conditions) (Ahlers et al., 2009; Ohana et al., 2024). Euler with open boundaries has four families (Toro, 2009; Takamoto et al., 2022). Structure (density total variation, shock fraction, and enstrophy) separates compressive from vortical structure. EOS consistency (the median and RMS equation-of-state residuals) measures thermodynamic closure. Bulk totals (paired mass, two-component momentum, and total-energy errors) measure the balance under boundary flux. Energy components (kinetic and internal energy) measure the split between mechanical and thermodynamic energy. Under open boundaries, mass, momentum and energy change through boundary flux, so the bulk-total errors compare paired trajectories rather than invariant totals. Registered RB profiles and the Euler entropy and positivity diagnostics are excluded

rather than assigned post hoc to a family. Euler with periodic boundaries is excluded, because in the PDE-KD space three of its fifteen (model, window) cells, all at leads 1–20, have KD below the floor, against none on open boundaries. RB uses leads 61–183. On Euler, KD and the diagnostic errors share the leads 64–88 (Appendix G.1). The reduction and tie handling are given at the end of this subsection (*Reduction and caveats*).

Table 47: **Agreement of feature-space KD with physical diagnostics.** Spearman ρ between KD and each model's mean rank over the named family's scalars (RB: five scalars, nine models; Euler-open: the indicated family, five models); positive when KD (or the replacing error) grows with physical error. vRMSE and spectral-error rows: the same correlation with KD replaced by those errors (Appendix G.3). Bold: exact two-sided $p \leq 0.05$ ($|\rho| \geq 0.70$ on RB, $|\rho| = 1$ on Euler). The Euler-open PCA-white basis is fit in-sample on 350 of the 500 test ICs (150 held out). Random init: the untrained PDE-KD network, mean over three initialization seeds of each seed's ρ (Appendix C.2), bold when that mean reaches the exact critical $|\rho|$.

	Rayleigh–Bénard	Euler, open boundaries			
space	physical scalars (5)	structure (3 scalars)	EOS consistency (2 scalars)	bulk totals (4 scalars)	energy components (2 scalars)
PDE-KD	**0.84**	**1.00**	**1.00**	0.70	0.70
ImageNet	0.35	0.90	0.90	0.90	0.90
Poseidon	**0.77**	0.70	0.70	**1.00**	**1.00**
Random init	**0.91**	0.73	0.73	0.83	0.83
Pixels	**0.87**	0.90	0.90	0.60	0.60
Spectrum	**0.79**	0.80	0.80	0.80	0.80
PCA-white	**0.83**	0.60	0.60	0.90	0.90
vRMSE	**0.92**	0.70	0.70	**1.00**	**1.00**
Spectral error	**0.93**	0.40	0.40	–0.10	–0.10

Bold entries of Table 47 pass the exact test of its caption. On RB every row except ImageNet passes; on Euler only perfect orderings pass ($p = 2/120$, before any multiplicity correction), so the Euler columns rank representations only descriptively. Among learned RB representations, PDE-KD has the highest aggregate agreement with the physical diagnostics ($\rho = 0.84$, $p = 0.007$). Among the non-learned spaces, Pixels reaches 0.87. The two reference errors on the same rollouts, the pointwise vRMSE and the spectral error, reach 0.92 and 0.93, at least as high as every feature-space row of the table. The untrained network (Random init), scored on the same nine models, reaches 0.91, above every other feature-space row. The nine models form two families: five Lola and Walrus models and four Well operators. For the single functionals, the pooled correlation mostly reflects the separation between these families, and no within-family correlation reaches $p = 0.05$ (Appendix G.4, caveat (i)). On this nine-model ladder, agreement with the physical scalars therefore distinguishes neither a distributional score from pointwise error nor trained from untrained features.

Euler includes only five models, so each correlation is coarse, and a family rank built from several diagnostics adds no independent models. PDE-KD reaches 1.00 on structure and EOS consistency, while Poseidon reaches 1.00 on bulk totals and energy components. These perfect orderings are the only Euler entries of any feature space that pass this uncorrected test. In every row, the structure and EOS-consistency columns coincide, as do the bulk-total and energy-component columns. vRMSE follows Poseidon's pattern (0.70, 0.70, 1.00, 1.00). The IC-mean spectral error agrees weakly or not at all (0.40 on structure and EOS consistency, -0.10 on bulk totals and energy components). This row is sensitive to how ICs are aggregated (*Reduction and caveats* below). No representation leads across all families. Normalization, architecture, and input-interface differences remain possible confounders, including the omission of density from the Poseidon adapter.

Reduction and caveats. For a physical functional h, the error is

$$e_h = |h(x_{\mathrm{gen}}) - h(x_{\mathrm{ref}})| \,. \tag{12}$$

We form this absolute member–reference error at each selected lead and average it over leads, ensemble members and ICs, in that order. Models are ranked by e_h with tie-aware midranks, and the Spearman correlation of each family entry (defined above) is tie-aware. Averaging ranks gives every scalar one vote without mixing unlike physical units; the generated record also lists every per-scalar

correlation. Paired KD uses the unbiased cubic-polynomial U-statistic with $\gamma = 1/d$ and excludes every kernel pair sharing an IC. KD floors are not subtracted. The model roster (nine models for RB, five for Euler) was fixed before any feature space was inspected. The vRMSE and spectral-error rows apply the same rule with KD replaced by the model's frame-normalized RMSE and by its spectral power error averaged over the four isotropic bands (Appendix C.6). Both are reduced from the same rollouts' per-IC metric stores on the same leads. The spectral error is the canonical IC mean per band, which near-quiescent ICs can dominate. With the per-band IC median instead, its row reads 0.88 on RB, 1.00 on structure and EOS consistency, and 0.70 on bulk totals and energy components. The weak Euler agreement of the IC-mean row is therefore not robust to that choice.

The Spectrum entries of Tables 32 and 47 mix two feature stagings. They use canonical-resize features for perceptual severity, native-grid features for spectral severity and RB physical alignment, and canonical features for Euler, the only Spectrum staging measured there. No common canonical staging of the Spectrum row was scored. EOS RMS is tail-sensitive, and Euler enstrophy can be dominated by near-vacuum cells. Averaging ranks can hide disagreement among a family's scalars, and it does not make a family mechanism-specific.

G.4 Per-functional correlations

Table 48 expands Table 47 (Appendix G.3) to single functionals. For each space and functional, it gives Spearman's ρ between KD and that functional's absolute generated–reference error across models, using paired KD at leads 61 onward. On Euler, KD uses the shared leads 64–88, whereas the observable errors use each model's own leads (Appendix G.1), so the Euler entries of Table 49 and Figure 11 are not lead-matched as the Euler columns of Table 47 are. With nine RB models, $|\rho| \geq 0.70$ is the exact $p \leq 0.05$ cutoff (Table 22). With five Euler models and no ties, only $\rho = 1$ reaches it (Table 49). Figure 11 summarizes two flagship diagnostics per system. The RB ρ is pooled over two model families and mostly reflects the separation between them (caveat (i) below). On RB, PDE-KD reaches the cutoff for 5 of 7 functionals, ImageNet for 0, Poseidon for 2, Random init for 5 (seed mean), Pixels for 5, Spectrum for 4, Spectrum-native for 6, PCA-white for 5 and PCA-raw for 6. The correlations keep models whose KD is below its floor. In the spaces marked ^{u} in Table 48, scoring each model with one ensemble member instead of two moves KD further than the gaps between the three near-neighbor Lola models, in every testable window of both designs (paired and disjoint). A low ρ in these spaces therefore reads "unresolved at this separation", not "disagrees". PDE-KD's largest pooled ρ are divergence RMS ($+0.95$), the mean-temperature profile ($+0.88$), Nusselt and the heat-flux profile ($+0.75$), and wall-boundary violation ($+0.72$). On Euler, all 450 correlations of the five spaces scored on both boundaries (PDE-KD, ImageNet, Poseidon, Pixels and Spectrum; every observable, window and boundary condition) are positive, the weakest $+0.20$. On open boundaries, PCA-white reaches 0.0 (unphysical fraction at leads 1–20), and every correlation of the three Random-init seeds is positive (weakest $+0.10$). Table 49 shows open boundaries at leads 61 onward. Poseidon, Pixels and PCA-white carry the RB within-family-unresolved flag, which was not re-measured on Euler, so their bold Euler cells do not show that they order the three Lola models.

Caveats on the correlations. Four caveats apply. (i) *Within* model families, no space's ρ reaches $p = 0.05$ in any of 48 cells. The four Well operators cannot reach it even with a perfect ordering (smallest $p = 0.083$), and the five Lola and Walrus models only with a perfect ordering ($p = 0.017$), which none attains (best $+0.90$, $p = 0.083$). The pooled nine-model ρ therefore mostly reflects the separation *between* families, not agreement within them. (ii) Leads 1–20 are the only window in which a physical oracle, the relative Nusselt bias, orders all 20 cross-family model pairs; it orders 17 and 18 of the 20 in the later windows. At leads 1–20 the comparison uses the disjoint design, because the paired design used for the correlations above cannot adjudicate the learned spaces in that window. A space's verdict is decided by an inverting model when it does not separate the families, and by the boundary pair (the highest-KD Lola or Walrus model and the lowest-KD Well model) when it does. In the paired design, the deciding model of each learned space is flagged at its floor. In the disjoint design, KD orders 13, 15 and 18 of the 20 pairs in PDE-KD, ImageNet and Poseidon, and all 20 in both spectra and PCA-white. Sixteen of PCA-white's 20 pairs involve a below-floor model, although its deciding pair is clean. Pixels and PCA-raw cannot be scored in this window, because their deciding model sits at its own floor. This comparison rests on one oracle on RB and leaves Pixels and PCA-raw unscored, so it does not establish a general ordering of non-learned over

Table 48: **Spearman ρ between KD and single-functional error**, RB, nine models, leads 61 onward. Bold: exact tie-aware two-sided $p \leq 0.05$ ($|\rho| \geq 0.70$ at $n = 9$). *Profile functional: median ρ over its 16 height bins, bold only if every bin reaches the cutoff. ^{u}Within-family unresolved: a low ρ in this space reads "unresolved at this separation", not "disagrees" (see text). Random init: the untrained PDE-KD network, mean over three initialization seeds (Appendix C.2); its count rows are means of the per-seed counts. The heat-flux-profile row equals the Nusselt row in every space, and at $n = 9$ the two count rows coincide.

functional	role	PDE-KD	ImageNet	Poseidon^u	Random init	Pixels^u	Spectrum	Spectrum-native	PCA-white^u	PCA-raw^u
Nusselt number	transport	**0.75**	0.17	0.68	**0.88**	**0.90**	**0.85**	**0.72**	**0.85**	**0.98**
enstrophy	structure	0.68	0.18	0.67	**0.96**	**0.85**	**0.97**	**0.83**	0.67	**0.87**
kinetic energy	structure	0.65	0.10	0.63	**0.92**	**0.83**	**0.93**	**0.80**	0.68	**0.88**
divergence RMS	admissibility	**0.95**	0.55	**0.90**	**0.73**	**0.77**	0.65	**0.70**	**0.93**	**0.80**
wall-BC violation	admissibility	**0.72**	0.40	0.62	0.44	0.48	0.35	0.30	**0.73**	0.57
heat-flux profile*	profile	**0.75**	0.17	0.68	**0.88**	**0.90**	**0.85**	**0.72**	**0.85**	**0.98**
mean-temperature profile*	profile	**0.88**	0.50	**0.92**	0.64	0.67	0.60	**0.80**	**0.93**	**0.73**
# $\rho \geq .70$		5	0	2	5.0	5	4	6	5	6
# reaching exact cutoff		5	0	2	5.0	5	4	6	5	6

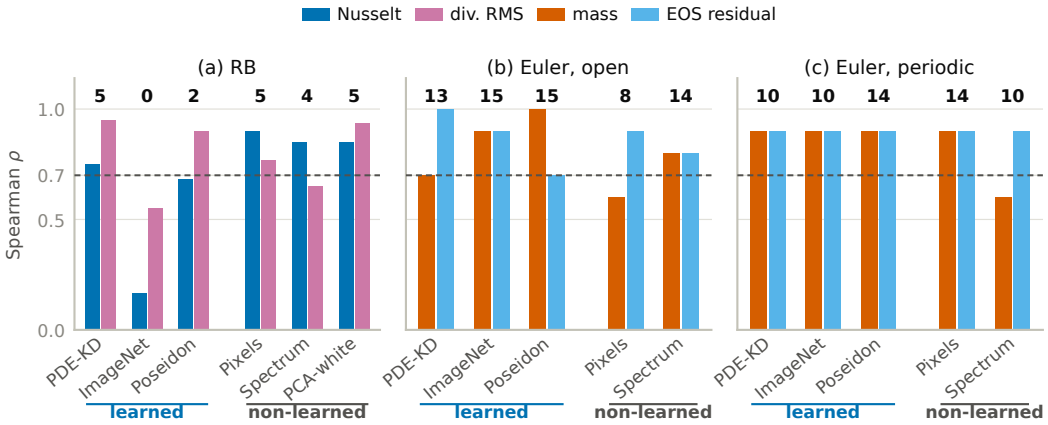


Figure 11: **Pooled rank correlation between KD and two physical diagnostics per system.** Spearman ρ between paired KD and the error of two diagnostics across models (RB: Nusselt number and divergence RMS; Euler: mass and EOS residual). (a) RB, nine models pooled over two model families, leads 61 onward; (b) Euler with open and (c) with periodic boundaries, five models, KD on the shared leads 64–88 (observable errors on each model's own leads, Appendix G.1; not lead-matched). Dashed line: the nine-model RB significance cutoff (0.70). Numbers above the bars: functionals (of 7 on RB, 15 on Euler) with $\rho \geq 0.70$; on Euler these counts are descriptive.

learned spaces. (iii) On Euler, many observables rank the five models identically, so agreement with several of them is one confirmation, not several. This is a property of the five-model ladder: across the 500 reference ICs on open boundaries, most observable pairs are nearly uncorrelated (median Pearson $+0.006$; 122 of 182 pairs below 0.2 in magnitude; enstrophy versus entropy, for example, $+0.067$). RB is the opposite case: its five non-profile observables order the nine models in five different ways, while their reference levels co-vary across ICs (median Pearson $+0.56$ over 175 ICs). (iv) The reduction of an observable matters. On the five-model Lola and Walrus ladder at leads 61 onward, Nusselt read as a signed bias (the error of the model's mean) gives $\rho(\mathrm{KD}) = +0.90$ against $\rho(\mathrm{vRMSE}) = +0.70$. Read as the per-lead absolute error used here, it gives $+0.60$ against $+1.00$ ($p = 0.017$). The two reductions differ by the decorrelation of chaotic trajectories, which the signed bias averages out and the per-lead error, like a pixel error, does not.

Physical findings on Euler. Several findings hold on the common 500-IC set. The EOS residual itself rises monotonically with Lola's compression in every (boundary, window) cell; this concerns its value, not its rank agreement with KD. Walrus models leave the hyperbolic domain: non-positive

Table 49: **Spearman ρ between KD and single-functional error**, Euler with open boundaries, five models on a common set of 500 ICs, paired KD on the shared leads 64–88 (observable errors: leads as in Appendix G.1). Bold: exact tie-aware two-sided $p \leq 0.05$; without ties only $\rho = 1$ reaches it, and with tied errors a ρ below 1 can. Poseidon, Pixels and PCA-white carry the RB within-family-unresolved flag of Table 48 (not re-measured on Euler). Random init: mean over three initialization seeds of each seed's ρ, bold when the mean reaches the exact critical $|\rho|$; its count rows are means of the per-seed counts.

functional	role	PDE-KD	ImageNet	Poseidon	Random init	Pixels	Spectrum	PCA-white
enstrophy	structure	0.90	**1.00**	0.90	0.77	0.70	0.90	0.70
shock fraction	structure	**1.00**	0.90	0.70	0.73	0.90	0.80	0.60
TV of density	structure	**1.00**	0.90	0.70	0.73	0.90	0.80	0.60
EOS residual	admissibility	**1.00**	0.90	0.70	0.73	0.90	0.80	0.60
EOS residual (RMS)	admissibility	**1.00**	0.90	0.70	0.73	0.90	0.80	0.60
minimum density	admissibility	0.80	0.90	0.80	0.67	0.60	**1.00**	0.60
minimum pressure	admissibility	0.60	0.80	0.90	0.73	0.50	0.90	0.80
unphysical fraction	admissibility	**0.97**	**0.97**	0.82	0.77	0.82	0.87	0.67
total energy	conservation	0.70	0.90	**1.00**	0.83	0.60	0.80	0.90
entropy	conservation	0.90	**1.00**	0.90	0.77	0.70	0.90	0.70
internal energy	conservation	0.70	0.90	**1.00**	0.83	0.60	0.80	0.90
kinetic energy	conservation	0.70	0.90	**1.00**	0.83	0.60	0.80	0.90
mass	conservation	0.70	0.90	**1.00**	0.83	0.60	0.80	0.90
momentum x	conservation	0.70	0.90	**1.00**	0.83	0.60	0.80	0.90
momentum y	conservation	0.60	0.70	0.90	0.87	0.70	0.60	**1.00**
# $\rho \geq .70$		13	15	15	10.7	8	14	9
# reaching exact cutoff		5	3	5	2.0	0	1	1

density or pressure occurs in 24–62% of their ICs (both boundaries, every window), against 0–2% for Lola. On periodic boundaries, the reference never loses entropy (0 of 44,000 increments over leads 1–91), whereas Lola loses entropy in a share of ICs that grows with compression (Table 52). The median per-IC density total-variation ratio shows ringing for Walrus (1.33–3.24 in every boundary condition and window); Lola's smoothing on periodic boundaries is in Table 52.

G.5 Scoring trained neural simulators: details

This subsection gives the protocol and full results behind Table 4 of Section 4.3.

Models and expected order. The models are defined in Appendix G.1. The three Lola models (Rozet et al., 2025) differ only in latent compression, and Walrus-FT is Walrus-PT (McCabe et al., 2026) fine-tuned on the target physics. The expected order (c64 before c16 before c4; FT before PT) is a prior from how the models were built, not a measurement. It was fixed before any score was computed, and no checkpoint was trained or selected using these metrics. The four Well operators have no expected order and are reported in Appendix G.1.

Protocol. RB uses the 175 test ICs and leads 61–183. Euler with periodic boundaries uses the 500 test ICs. There, KD and the shares of ICs with an entropy decrease or an unphysical state use the leads 64–88 that all models reach, and the vRMSE, TV and EOS values use each model's own leads (Appendix G.1). PCA-white was not scored on Euler-periodic and PCA-raw on neither Euler ladder, and the spectral error is shown for RB only (Table 51). CRPS covers the three Lola pairs. KD is the paired, same-IC-excluded estimator with one rollout per IC for every model. For Lola, whose stored rollouts hold two ensemble members per IC, the first member is used; the snapshots in Tables 4 and 51 show that same member. vRMSE is taken on the ensemble mean for Lola and CRPS over its 16 members. A KD value is shown in parentheses when it lies within two standard deviations of its split-half floor, so it is not resolved from the reference on its own. Each Walrus entry is one inference draw; Appendix G.1 gives the draw variance and the second RB draw of Table 45.

Snapshots. Each snapshot is the first test IC of one coefficient cell, not selected: RB buoyancy at $\mathrm{Ra} = 10^{10}$, $\mathrm{Pr} = 10$ at lead 183, in the 128×128 window of largest reference variance; Euler density on the full periodic domain at lead 88. Colors are centered on the reference frame with

a half-width set by its 90th percentile deviation, computed from the reference alone, so a rollout cannot rescale itself.

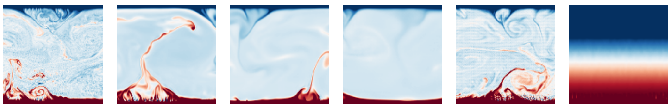
Raw values. Table 50 gives the raw Euler values behind the rescaled rows of Table 4. The rescaling is linear per row, so it changes no ordering and no expected-order count.

Table 50: **Raw values of Table 4** (Euler, periodic boundaries; 500 test ICs; KD on leads 64–88). KD $\times 10^3$ (Spectrum: raw KD); columns, parentheses, highlighting and expected order as in Table 4.

		Lola			Walrus		
	reference	c64	c16	c4	FT	PT	expected order
KD, learned features							
PDE-KD		(**1.43**)	5.09	36.7	5.63	141	4/4
ImageNet		**37.1**	86.6	584	297	930	4/4
Poseidon		(0.101)	(**0.0189**)	(0.157)	(1.00)	(126)	3/4
KD, non-learned features							
Pixels		(0.243)	(**0.110**)	(0.270)	(1.70)	(5.94)	3/4
Spectrum (raw KD)		(**7.11**)	(10.4)	63.3	972	793	3/4
error against the paired reference							
vRMSE		0.389	**0.386**	0.424	0.753	1.210	3/4
CRPS		0.104	**0.103**	0.115	–	–	2/3
physical consensus (mean rank over distribution-level diagnostics)							
mean rank		**1.25**	2.00	3.25	4.25	4.25	3/3†

Rayleigh–Bénard. Table 51 repeats Table 4 on the 175 RB test ICs, with raw values. KD in all three learned feature spaces orders all four expected pairs, and so does the spectral error. The per-lead scores reverse the compression order: vRMSE orders one pair and CRPS none, because both prefer the most compressed Lola model, by about 6%. The snapshots suggest why: compression removes the thin plumes of the reference. Walrus-PT has relaxed to the conductive profile, a collapse that vRMSE barely registers (0.980 against 0.950 for Walrus-FT). The spectral bands locate the reversal (Table 45): the three smaller-scale bands order the Lola models as KD does, whereas the largest-scale band orders them as vRMSE does. Among the non-learned spaces, Pixels orders all four pairs (see Resolution below), Spectrum three and PCA-white two.

Table 51: **The Rayleigh–Bénard half of Table 4.** 175 test ICs, leads 61–183; snapshots show buoyancy at $\mathrm{Ra} = 10^{10}$, $\mathrm{Pr} = 10$, lead 183. KD $\times 10^3$ (Spectrum: raw KD); parentheses, highlighting and expected order as in Table 4; spectral error is the relative band-power error against the paired reference.

		Lola, best → worst			Walrus, best → worst		
	reference	c64	c16	c4	FT	PT	expected order
lead 183							
KD, learned features							
PDE-KD		(**0.308**)	(0.719)	2.15	2.40	67.5	4/4
ImageNet		204	445	1,044	**73.2**	864	4/4
Poseidon		(**−0.265**)	(−0.130)	(0.799)	23.7	955	4/4
KD, non-learned features							
Pixels		(**−2.67**)	(11.8)	(16.8)	(25.7)	1,049	4/4
PCA-white		(0.110)	(0.609)	(**−0.0393**)	4.17	48.8	2/4
Spectrum (raw KD)		(8.60)	(**8.52**)	(30.9)	(30.0)	1,557	3/4
error against the paired reference							
vRMSE		0.704	0.678	**0.659**	0.950	0.980	1/4
CRPS		0.367	0.353	**0.344**	–	–	0/3
spectral error		**0.263**	0.329	0.454	1.14	1.85	4/4
physical consensus (mean rank over distribution-level diagnostics)							
mean rank		**1.50**	**1.50**	3.00	4.00	5.00	3/3†

Resolution. The expected-order counts compare point estimates. On RB, ImageNet resolves all five models from its floor, whereas PDE-KD resolves Lola-c4 and both Walrus models, and Poseidon only the two Walrus models. PDE-KD's Lola-c64 and c16 are ordered by the regime-conditional paired contrasts of Appendix G.7. Every Pixels score for Lola and Walrus-FT lies within two standard deviations of its floor, so its four ordered RB pairs carry no evidence. Counting only pairs whose two models are both resolved from the floor, PDE-KD resolves one RB pair and two Euler pairs, ImageNet four of each.

Physical consensus. The last row of Tables 4, 50, 51 and 52 is each model's mean rank, with ties at their mean, over diagnostics fixed in advance that compare the statistics of the rollout with those of the reference. On Euler these are the share of ICs with an entropy decrease, the TV(ρ) ratio (distance from 1), the EOS residual and the share of ICs with an unphysical state; on Rayleigh–Bénard, the window-mean Nusselt and enstrophy biases. Per-lead errors are excluded, because after decorrelation they behave like a pixel error. On Euler, the entropy decrease, the TV ratio and the EOS residual each rank Lola-c64 ahead of Lola-c16 and Lola-c16 ahead of Lola-c4, and unphysical states never occur for Lola, so the consensus supports the compression order. The two Walrus models tie (4.25), since entropy and EOS favor Walrus-FT while the TV ratio and unphysical states favor Walrus-PT. On Rayleigh–Bénard, the consensus ties Lola-c64 and Lola-c16 (1.50): the enstrophy bias favors c64 and the Nusselt bias c16. The spectral error, which the consensus does not use, orders c64 ahead of c16, as do the three smaller-scale spectral bands (Table 45). Every other within-family pair agrees with the expected order.

Why the pointwise scores reverse the compression order. The reversal is an instance of the distortion–perception tradeoff (Blau & Michaeli, 2018): after decorrelation, a smoother field pays less for being out of phase with the reference. A buoyancy field at its spatial mean has vRMSE 1.0, whereas pairs of independent reference trajectories have a median channel-mean vRMSE of 1.05 (range 0.74–1.20; Appendix C.6). CRPS scores every member rather than the ensemble mean and still prefers compression, so the reversal does not come from scoring a mean.

Physical diagnostics. Table 52 scores the same models with physical diagnostics. On RB, the per-lead Nusselt error behaves like vRMSE and orders one pair, whereas the window-mean biases of the Nusselt number and enstrophy order three and four. On Euler, two checks that need no reference realization agree with the expected order: total entropy decreases, a violation of the second law, in 1.2%, 14.4% and 57.6% of ICs from least to most compressed and in more ICs for Walrus-PT than for Walrus-FT, and the EOS residual orders all four pairs. A third such check disagrees: unphysical states (non-positive density or pressure) never occur for Lola, so its pairs are tied, and they occur in slightly more ICs for Walrus-FT (39.4%) than for Walrus-PT (36.8%).

Disagreements outside the expected pairs. On Euler, PDE-KD and ImageNet both place Walrus-FT between Lola-c16 and Lola-c4 (PDE-KD 5.09, 5.63 and 36.7; ImageNet 86.6, 297 and 584), although Walrus-FT reaches non-positive density or pressure in 39% of ICs and Lola in none. Spectrum and vRMSE place it behind every Lola model, as do Poseidon and Pixels with values within two standard deviations of their floors (Table 50). On RB, ImageNet and Poseidon rank both U-Nets of the Well ahead of Walrus-PT, and so does PDE-KD in the regime-conditional score of Table 45 (KD $\times 10^3$ 90.2 and 69.0 against 96.7), whereas vRMSE, the spectral error and the physical diagnostics rank Walrus-PT ahead (Appendix G.1). In the paired score, PDE-KD places UNetClassic just behind Walrus-PT (69.5 against 67.5).

Limits. The evidence is three compressions and one fine-tuning pair per system, and the expected order is a prior. We do not estimate the power of the tests on the expected pairs.

G.6 Euler compression ladder

Appendices G.3 and G.5 score the Euler model ladders pooled over the adiabatic index. This subsection conditions on it: within each adiabatic-index cell, it asks whether a space recovers the designed order of the Lola compression ladder.

Table 52: **Physical diagnostics of the neural simulators in Table 4.** Nu |err|: per-lead absolute Nusselt error in Nusselt units (Ahlers et al., 2009). |bias|: median over ICs of the relative difference between rollout and reference window means. Entropy decrease and unphysical state: shares of ICs on leads 64–88. TV(ρ) ratio: median over ICs of rollout over reference, best at 1. EOS residual: window mean of each frame's median relative residual (Toro, 2009). Columns, highlighting and expected order as in Table 4; † tied pairs are excluded. Mean rank: the physical consensus defined in the text.

		Lola			Walrus		
	reference	c64	c16	c4	FT	PT	expected order
(a) Rayleigh–Bénard							
physical diagnostics							
Nu |err| (per lead)		22.9	22.3	**20.8**	30.5	39.6	1/4
Nu |bias|		0.058	**0.044**	0.104	0.212	0.758	3/4
enstrophy |bias|		**0.061**	0.072	0.124	0.425	0.960	4/4
physical consensus (mean rank over distribution-level diagnostics)							
mean rank		**1.50**	**1.50**	3.00	4.00	5.00	3/3†
(b) Euler, periodic							
physical diagnostics							
entropy decrease	0%	**1.2%**	14.4%	57.6%	33.6%	82.4%	4/4
unphysical state	0%	**0.0%**	**0.0%**	**0.0%**	39.4%	36.8%	0/1†
TV(ρ) ratio	1	**0.91**	0.85	0.64	1.77	1.52	3/4
EOS residual	2.6×10^{-8}	**0.0011**	0.0025	0.0045	0.048	0.059	4/4
physical consensus (mean rank over distribution-level diagnostics)							
mean rank		**1.25**	2.00	3.25	4.25	4.25	3/3†

Design and the pre-specified comparison. We asked whether each space recovers the designed ordering of the Lola compression ladder (latent compression $80\times < 320\times < 1280\times$) within each of 10 adiabatic-index cells, with 50 ICs per cell, in each of five 20-lead windows. With three models, chance recovers the ordering in 8.3 of the 50 (cell, window) combinations. On periodic boundaries, PDE-KD and ImageNet recover it in all 50 combinations, Spectrum in 45, Poseidon in 19 and Pixels in 9. On open boundaries the counts are 38, 35, 22, 10 and 11 in the same order. The PCA spaces were not scored in this analysis. Pixels on both boundaries and Poseidon on open boundaries are therefore not distinguishable from chance (one-sided binomial tails treating the 50 combinations as independent, $p = 0.46$ and 0.20 for Pixels, 0.32 for Poseidon). At 50 ICs per cell, this pre-specified comparison cannot rank PDE-KD against ImageNet. Both recover the ordering in all 50 periodic combinations, and on open boundaries an exact McNemar test on the 8 combinations recovered only by PDE-KD and the 5 recovered only by ImageNet does not reject (two-sided $p = 0.58$).

Five ICs per cell: ImageNet is more reliable on periodic boundaries. With only 5 ICs per cell on periodic boundaries, ImageNet recovers the ordering more reliably than PDE-KD in all ten paired draws: 0.95 versus 0.71, with complete separation (ImageNet's worst draw 0.86, PDE-KD's best 0.84), sign test $p = 0.002$. This comparison was found post hoc. Adding models cannot change the three-model statistic, so repeating it with more models would not be a replication. The confirmation instead uses the consistency of the full five-model ordering (the three Lola models, Walrus-PT and Walrus-FT), i.e. the share of (cell, window) combinations that return the most frequent ordering: 0.59 versus 0.28, 10 of 10 draws, $p = 0.002$. Spectrum is also more consistent than PDE-KD on this statistic (0.54). On periodic boundaries the gap persists at leads 61–100 (0.975 versus 0.795, 10 of 10, $p = 0.002$), a post-hoc slice that refines the result rather than confirming it. On open boundaries the evidence for ImageNet's advantage is weaker. The sign test on the designed-ordering gap does not reject (0.33 versus 0.27, 8 of 10 draws, $p = 0.11$), while the one on ordering consistency rejects (0.21 versus 0.16, 9 of 10, $p = 0.021$); neither separates the draws completely.

G.7 Paired inversions and CRPS against climatology

This subsection compares KD with pointwise and ensemble scores on the RB models: the checkpoint pairs on which they disagree, the per-frame KD witness, and the CRPS of the Lola ensembles against a no-skill forecaster.

Where KD and pointwise errors disagree. We tested every pair of the 11 RB model checkpoints in three windows under vRMSE, MSE and CRPS, 339 tests in total. An inversion is counted when both KD and the error resolve the pair at ± 2 standard errors with opposite signs. We find 61 inversions. Twenty-nine of them involve a checkpoint that is unresolved under its own resolution flag; a pairwise contrast can still resolve a pair that contains such a checkpoint. Within the Lola compression ladder, all 16 resolved comparisons invert at leads 21 onward: KD prefers less compression, whereas vRMSE and CRPS prefer more (Appendix G.5 discusses why). Inversions concentrate within families: they occur in 20 of 27 comparisons between Lola models (74.1%) but in 2 of 108 between Lola and Well models (1.9%). KD and the pointwise errors therefore mostly agree on which family is better and disagree on which member is. Walrus illustrates the magnitudes: fine-tuning lowers its marginal paired KD at leads 61 onward 28.6-fold ($67.27 \rightarrow 2.36 \times 10^{-3}$), while its vRMSE moves by a factor of 1.02 (0.980 to 0.962).

Per-frame witness. The KD witness of a generated frame is its mean kernel value against generated frames of other ICs minus that against reference frames of other ICs. Its mean over generated frames is exactly the generated half of Eq. (10). Per frame, the witness is only weakly correlated with vRMSE within a lead window and Ra. Over the 11 checkpoints, the median $|\rho|$ is 0.13, 0.22 and 0.25 in the three windows (single checkpoints reach 0.50), and the median signed ρ is $+0.11$, $+0.15$ and -0.04. The ladder's KD separation is concentrated at $\mathrm{Ra} = 10^{10}$. On one reference trajectory there ($\mathrm{Ra} = 10^{10}$, $\mathrm{Pr} = 1$; buoyancy, leads 61–183), the most and least compressed models ($1024\times$ and $64\times$) retain 0.17% and 0.80% of the reference's highest-band power.

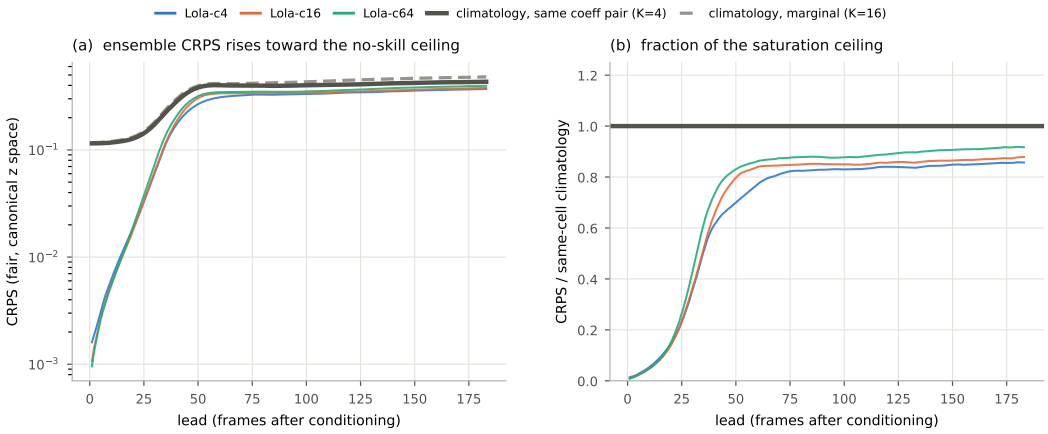


Figure 12: **CRPS against climatology after decorrelation.** (a) Fair CRPS of the three Lola ensemble models on RB against two no-skill climatology forecasters built from other reference trajectories (same coefficient cell, $K = 4$; marginal, $K = 16$). (b) Ratio to the same-cell climatology, per lead. Lines above 1 are worse than no skill.

CRPS against climatology. We compared each ensemble model's CRPS with a no-skill forecaster that draws other reference trajectories of the same (Ra, Pr) cell, whose fair CRPS at leads 61 onward is $\mathrm{CRPS}_{\mathrm{clim}} = 0.412$. The three Lola models, the only $K = 16$ ensembles in the zoo, score 0.84, 0.86 and 0.89 times this climatology (Lola-c4, c16 and c64; their CRPS is in Table 45). None is worse than no skill, and none is far from it. Across them, CRPS spans a factor of 1.07 (Figure 12), against 6.6 for the marginal paired KD on their two stored members per IC (KD $\times 10^3$ 2.12, 0.72 and 0.32). That KD span is floor-driven. Under a descriptive rule (KD exceeds its split-half floor by more than two standard deviations of the floor across split seeds), only Lola-c4 is resolved. The smallest KD, Lola-c64's, lies 0.5 of these standard deviations above its floor. Table 51 scores one member per IC and applies the same rule: Lola-c4 (2.15) is resolved, whereas Lola-c16 (0.719) and Lola-c64 (0.308) are not. This rule differs from the ladder's resolution flag, so Lola-c4 is resolved here but not in Table 45.

H PROBES OF THE FEATURE MAP

We probe what the frozen features encode about physical functionals of the flow and about the swept coefficients, and we report the geometry of the feature space.

Physical functionals. We fit ridge (linear) and MLP probes from frozen features to the log energy, enstrophy and palinstrophy of RB velocity. The probes use three frames per trajectory (4,200 training, 525 validation) and a trajectory-level bootstrap. Raw R^2 saturates near 1 for the feature map and for a random network, because these functionals largely re-encode Ra and Pr. A linear probe on pooled velocity pixels scores about 0, as expected for targets quadratic in velocity (an MLP reaches 0.72–0.87). We therefore partial out a linear function of $\log$Ra and $\log$Pr (the normalized coefficients of Table 6), fit by least squares with an intercept, and probe the residual (Table 53). Coefficient dependence that is not linear in $\log$Ra and $\log$Pr stays in the residual, so part of the margins below may still be coefficient information.

Residual probes. The supervised features beat a random-initialization floor of the same network (5 seeds, calibrated BatchNorm) with non-overlapping confidence intervals for the linear probe; the MLP intervals overlap for all three functionals. A linear probe on 16×16-pooled velocity pixels recovers almost none of the residual (0.02–0.08), again by construction. The fair pixel comparison is an MLP on the same pixels, which reaches 0.58, 0.42 and 0.40 for energy, enstrophy and palinstrophy, against 0.96–0.98 for an MLP on the PDE-KD features and 0.94–0.97 on the random floor's. The pixel probe sees velocity only, whereas the feature map sees all four RB channels. An end-to-end network trained from scratch on the residual targets lands *below* the random floor, so it is not an upper bound.

Floor and checkpoint. The floor uses the initialization the feature map was trained from, with zero-initialized residual scales (Appendix C.2); it is not the Random init control. The probed checkpoint is a sibling of the released feature map: it shares training up to step 15,120 and then took its own cooldown (step 19,152). A generation-2 feature map (Appendix B.5; one channel at a time, per-component features concatenated) reaches similar residual R^2 (0.947 / 0.906 / 0.922) above a much lower random floor (0.845 / 0.678 / 0.699), so its margin over its own floor is two to three times larger. Between these generations, which also differ in training budget and feature construction, the random floor rises more than the trained features do, so the probed checkpoint's smaller margin reflects a higher floor rather than weaker trained features.

Table 53: **Residualized linear-probe** R^2 on RB validation frames (functionals with Ra and Pr linearly partialled out; 95% trajectory-bootstrap CI). PDE-KD: the sibling checkpoint at step 19,152. Random floor: the feature map's own initialization, not the Random init control of Appendix C.2; mean $\pm$ s.d. over 5 seeds. Pixels: 16×16-pooled velocity.

Target	PDE-KD (sibling)	random floor	Pixels
Energy	0.966 [0.959, 0.972]	0.922 ± 0.003	0.020
Enstrophy	0.920 [0.901, 0.935]	0.822 ± 0.011	0.075
Palinstrophy	0.912 [0.892, 0.926]	0.837 ± 0.012	0.081

From functionals to coefficients. The converse probe regresses the swept coefficients on exact functionals of velocity, on held-out trajectories. Energy, enstrophy and palinstrophy predict the Rayleigh number with $R^2 = 0.977$ (MLP). For shear flow they predict the Reynolds number with $R^2 = 0.688$, rising to 1.000 when time derivatives are added. The Schmidt number, a tracer property invisible to velocity, stays at $R^2 \approx 0$ and serves as a negative control. The feature map's coefficient regression (Table 11) is somewhat lower, for Ra (R^2 0.927 validation, 0.934 test) and for the single-frame Reynolds number (0.66 and 0.62). The feature map's Schmidt R^2 differs (0.74 / 0.69), because it also sees the tracer.

Geometry. On RB test frames, the 512-dimensional features have participation ratio 6.1 and entropy-effective rank 11.0. Ten directions carry 90% of the variance. The sample-varying part is 32% of the feature norm (about 10% of its square); the rest is a common offset. For comparison,

the participation ratios of the other frame spaces are 2.8 (ImageNet), 3.2 (Poseidon), 12.2 (Pixels), 1.9 and 1.8 (Spectrum, Spectrum-native), 311 (PCA-white) and 11.3 (PCA-raw).

I Out-of-distribution generalization

We used five sources of out-of-distribution (OOD) data, none in the training corpus: the fluid–structure interaction (FSI) cylinder wake of RealPDEBench (Hu et al., 2026), in both CFD and particle-image-velocimetry (PIV) arms; HRRR surface weather (Dowell et al., 2022); ERA5 (Hersbach et al., 2020); WeatherBench-2 forecasts (Rasp et al., 2024); and the Well's turbulent radiative layer (TRL2D), used in Appendix I.2.

Channel routing. HRRR and ERA5 carry mean-sea-level pressure, 2 m temperature and the two 10 m wind components. We route them by physical role in the channel dictionary (Appendix B.2): pressure to slot 0, temperature to the primary-scalar slot 1 and wind to the vector slots 3–4. Surface weather then occupies slots $\{0,1,3,4\}$, the slot pattern of RB (temperature in the place of buoyancy) and shear flow. The FSI CFD arm ($\{0,3,4\}$) and PIV arm ($\{3,4\}$) use patterns the feature map never saw. No other routing was tried, so the results carry no measure of routing sensitivity.

Grids and splits. HRRR is the 3 km analysis over the contiguous United States, bilinearly resized to 512^2 with open boundaries and cut into one trajectory per UTC day (24 hourly frames from 00 UTC; days with any non-finite value are dropped). Its days are split chronologically: training September 2018 to June 2025, test January to April 2026 (120 days). ERA5 is the global reanalysis, bilinearly regridded from 0.25° to a 256×512 latitude–longitude grid, with one trajectory per calendar month at 6-hourly cadence (112–124 frames). It is split by year: training 1979–2017, test 2020–2021 (24 months). Each field is z-scored with its dataset's per-channel constants. The HRRR axes of Appendix I.1 use the last 12 training months (July 2024 to June 2025; 363 days), because the test period spans only four months. The probes of Appendix I.5 are fit on the training period, and all other results on these two grids use the test period. WeatherBench-2 forecasts and their ERA5 truth are scored on WeatherBench-2's own 1.5° (121×240) grid and are never pooled with the ERA5 data above (Appendix I.4).

I.1 Reference-versus-reference axes and their triviality

We asked whether KD distinguishes reference populations along physical axes (Table 54, Figure 13). On FSI, KD in PDE-KD between Reynolds-number cells rises monotonically with the Reynolds step, spanning a factor of 40, and nearest-neighbor cell recovery is 100% (ImageNet 53%, Poseidon 82%). However, no single pair is significant, because three replicate simulations per cell floor the exact p at 0.10. On HRRR, the relabeling test in PDE-KD and ImageNet rejects all 276 hour pairs and all 66 month pairs (Poseidon: all hour pairs and 98.5% of month pairs). The month test relabels days within one calendar-month instance, so it is anti-conservative under month-level effects. On ERA5 the test rejects 0 of 66 month pairs, with a p-floor of $1/3$.

Two kinds of control test whether simpler signals explain these results: probes and non-learned KD spaces under the same null, and amplitude normalization on FSI. Under them, the two HRRR axes are trivial and the ERA5 and FSI axes undetermined (Table 55), so none of these axes isolates a property of the learned feature space.

Verdict rule. Before running the controls of Table 55, we fixed a verdict rule per axis. An axis is *trivial* if a two-class linear probe on 8×8-pooled pixels, on the same rows and under the same null as KD, rejects a fraction of pairs comparable to that of KD in PDE-KD, and KD in Pixels or a spectrum space does too. It is *non-trivial* if KD in PDE-KD rejects while the probe and every non-learned KD space do not. It is *undetermined* if the permutation floor exceeds 0.05 (ERA5 and FSI).

HRRR. On HRRR, the pixel probe also rejects 100% of the hour and month pairs, and a probe on the domain-mean 2 m temperature alone rejects 91.3% and 97.0% (Table 55). As a regime *readout*, however, that scalar is nearly useless for the hour of day (mean angular error 81.4° against 90° at chance, while pooled pixels reach 14.6°). The diurnal signal is therefore carried by the coarse

spatial pattern rather than the bulk level, whereas season is carried by both. Both axes are trivial. Restricting to a fixed region leaves 24 of 36 cells trivial and 12 undetermined. A non-trivial verdict was unreachable under the rule, because the pooled-pixel probe rejects every pair in all 24 regional HRRR cells.

FSI. On FSI, the Reynolds number sets the bulk speed (Pearson $+0.998$ with mean $|U|$). On the PIV arm, normalizing amplitude lowers PDE-KD's step-to-replicate separation from 3.45 to 1.20–1.31, which is ImageNet's unnormalized level (1.26), while ImageNet under the same transforms is unchanged at 1.19–1.26. This ratio and its tolerance were set after the raw result, not pre-registered. On the CFD twin, whose unnormalized separation is only 1.38, the ladder partially survives normalization, so the two arms are read separately. On raw fields, a trajectory-conditional test with a pooled replicate null rejects 83% of 1,224 cross-cell pairs in the PDE-KD space, at a held-out calibration of 8.3%. Its gap is 15.7 null interquartile ranges, against 0.3–0.6 for pixels and PCA. After amplitude normalization the same statistic falls to 1.3, 1.3 and 0.3 under the three normalizations, so there is no evidence that the gap survives amplitude control. That null is approximate (its 51 members are dependent), and the raw gap drops to 1.7 on the CFD twin.

Table 54: **Reference-versus-reference OOD axes** (KD $\times 10^3$; no model involved). "KD (1 step)" and "KD (max step)": median over cell pairs that far apart (max step in parentheses). "NN rec.": nearest-neighbor cell recovery. "Rejecting": fraction of all cell pairs with relabeling $p \leq 0.05$; $p_{\min}$: smallest attainable p.

axis	representation	pairs	KD (1 step)	KD (max step)	NN rec.	rejecting	$p_{\min}$
FSI (RealPDEBench, real): 17 (Re, m^*) cells, 3 replicates each							
	PDE-KD	136	19.8	800.4 (8)	1.00	0.000	0.10
	ImageNet	136	28.0	209.3 (8)	0.53	0.000	0.10
	Poseidon	136	64.7	10077.5 (8)	0.82	0.000	0.10
HRRR: UTC hour of day (steps in hours)							
	PDE-KD	276	1.17	45.1 (12)	–	1.000	–
	ImageNet	276	4.58	155.3 (12)	–	1.000	–
	Poseidon	276	3.43	254.8 (12)	–	1.000	–
HRRR: calendar month (steps in months)							
	PDE-KD	66	43.8	371.5 (6)	–	1.000	–
	ImageNet	66	12.1	71.3 (6)	–	1.000	–
	Poseidon	66	379.4	3079.7 (6)	–	0.985	–
ERA5: calendar month (steps in months)							
	PDE-KD	66	5.9	83.5 (6)	–	0.000	1/3
	ImageNet	66	2.5	25.7 (6)	–	0.000	1/3
	Poseidon	66	47.1	472.5 (6)	–	0.000	1/3

Table 55: **Triviality controls**: fraction of axis pairs rejected by linear probes and by KD in each space, under the same null. On ERA5 and FSI no pair can reject ($p_{\min} = 1/3$ and 0.10). Verdict: the rule of Appendix I.1.

arm	HRRR hour	HRRR month	ERA5 month	FSI Re
linear probes under the same null				
one scalar (d=1)	91.3%	97.0%	0%	0%
pooled statistics (d=34/18)	100.0%	100.0%	0%	0%
8×8 pooled pixels (d=256/128)	100.0%	100.0%	0%	0%
KD, relabelling permutation test (exhaustive on ERA5/FSI, sampled on HRRR)				
PDE-KD	100.0%	100.0%	0%	0%
ImageNet	100.0%	100.0%	0%	0%
Poseidon	100.0%	98.5%	0%	0%
Pixels	99.6%	100.0%	0%	0%
Spectrum	98.6%	81.8%	0%	0%
Spectrum-native	98.6%	81.8%	0%	0%
PCA-white	99.6%	100.0%	0%	0%
PCA-raw	99.3%	100.0%	0%	0%
verdict	trivial	trivial	undetermined	undetermined

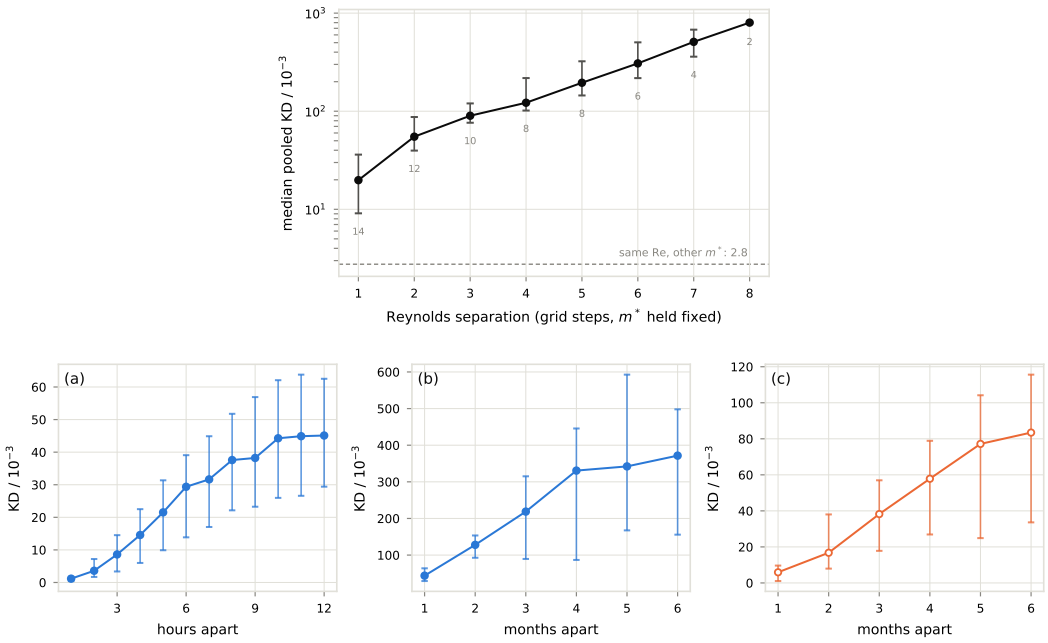


Figure 13: **Reference-versus-reference KD along OOD axes** in the PDE-KD space (KD $\times 10^3$, as in Table 54). Markers: median KD over the cell pairs at each separation; whiskers: their interquartile range. Top: FSI, KD between Reynolds cells against the Reynolds step at fixed mass ratio (numbers: cell pairs per step; dashed: same Reynolds number, other mass ratio); these markers carry no test outcome, because no FSI pair can reject ($p_{\min} = 0.10$). Bottom: (a) HRRR hours apart, (b) HRRR months apart, (c) ERA5 months apart; filled: every pair at that separation rejects at $p \leq 0.05$; open: at least one does not (every ERA5 marker, $p_{\min} = 1/3$). Both HRRR axes are trivial under the verdict rule of Appendix I.1, so these panels do not isolate a property of the PDE-KD feature space. On the FSI PIV arm, amplitude normalization lowers the step-to-replicate separation of PDE-KD to ImageNet's unnormalized level (Appendix I.1).

I.2 Controlled disturbances out of distribution

We applied the perceptual and spectral ladders of Appendix E.2 to FSI, HRRR and ERA5, with each dataset's IC pool as the n_{80} ceiling (26 / 60 / 12 perceptual, 51 / 120 / 24 spectral).

Detection counts. In the pre-registered roster of four spaces (PDE-KD, ImageNet, Pixels, Spectrum), 38 of the 48 perceptual (dataset, space, family) cells are detected, 4 are rejected at L5 without reaching 80% power within the pool, and 6 are neither (Table 56; per-family values in Tables 57–60). Blur accounts for five of the six misses. ImageNet detects all 12 of its cells, PDE-KD and Spectrum 11, and Pixels 4. Of the 132 spectral cells, 99 are detected, 10 are rejected at L5 without reaching 80% power within the pool, and 23 are neither. The other rows of Table 56 (Poseidon, Spectrum-native, PCA-white and PCA-raw) are a roster extension with its own denominator and are never pooled with these counts. These counts do not rank the spaces. For 33 of the 38 perceptual detections, the fitted power curve crosses 0.80 between the sweep's grid points of 6 and 8 ICs. Because 8 is the smallest IC count at which the exact test can reject at all, the fitted n_{80} values in that step do not separate these cells.

Spectral rejections per space. Table 3 counts instead the spectral families whose single L5 test rejects, i.e. the detected families plus the (+k) families of Table 56. By that count, PDE-KD rejects the most spectral families on HRRR (Table 3) and on ERA5 (all 11, against 10 for ImageNet), but not on FSI or TRL2D. On FSI, ImageNet rejects 10 of the 11 spectral families, and Spectrum and Spectrum-native 9 each, against 8 for PDE-KD. On TRL2D, ImageNet rejects 9, against 6 for PDE-KD.

Blur on FSI. On FSI, ImageNet KD rejects blur at every level, but cubic-kernel KD never rejects it in the PDE-KD space ($p = 0.08$–0.12), nor in Pixels or Spectrum. The energy distance on the same features rejects in the PDE-KD space at L5 (uncorrected $p = 0.036$, in a campaign of 240 cells without multiplicity control) and in Spectrum from L3. This suggests that the blur information survives the feature map while the cubic kernel does not see it. On FSI, the energy distance in the PDE-KD space rejects in 1, 3 and 3 of 40 clean-versus-clean draws in the disjoint, paired and conditional designs (the disjoint rate is also in Table 21). Its L5 rejection, however, is a single uncorrected test. The energy-distance result therefore withdraws only the reading that the PDE-KD features lack the blur information. It establishes no verdict of its own, and the KD verdict stands. We did not test whether ImageNet resolves physical field structure or only the structure of the $4\times$ bilinear upsampling. The kernel reading is also narrow: on the 24 high-wavenumber-removal cells of the spectral half, the pre-registered rule returns 2 cells that support it, 16 that complicate it and 6 in which neither estimator rejects.

Calibration. The disjoint tests are randomization tests, exact when the permuted units are exchangeable (Appendix D.6). The test code passes its exactness check on every out-of-distribution system (Table 19), and Table 61 reports each space's clean-versus-clean false-positive rate. On HRRR, the 95% intervals of PDE-KD (6 of 40, 0.057–0.298) and Poseidon (8 of 40, 0.091–0.356) lie above 0.05; neither is significant after Holm correction (Appendix D.6). The identity rung of FSI ImageNet, a single clean-versus-clean test on the published split, rejects at $p = 0.047$ with $B = 2{,}000$ and does not at $B = 20{,}000$ ($p = 0.051$); it is reported, not used to exclude the space. On ERA5 the spectral identity rung rejects at 0.060–0.075 in all four spaces over 200 re-partitions (nominal 0.05; no cell more than 1.6 binomial standard errors above it).

TRL2D. In a separate campaign, we ran both ladders on a Well system outside the corpus, the turbulent radiative layer (TRL2D), with the multi-physics protocol of Appendix E.4: all 90 trajectories (10 ICs $\times$ 9 cooling times, all splits), frames 20–100, pools 45 perceptual and 90 spectral. The spaces are PDE-KD, ImageNet, Poseidon and Pixels, with Spectrum on the perceptual and Spectrum-native on the spectral suite (no PCA basis exists for TRL2D). PDE-KD detects 3 of 4 perceptual and 5 of 11 spectral families, against 4 and 9 for ImageNet, 3 and 2 for Poseidon, and 2 and 3 for Pixels (Tables 56 and 60). Spectrum detects all 4 perceptual families, and Spectrum-native 5 of the 11 spectral families. The campaign's low-wavenumber band holds no grid mode for density or velocity, so low-k injection and removal change pressure only. These units are not independent ICs, so exactness rests on their exchangeability (Appendix D.6). The test code rejects 198 of 4,000 fresh clean-versus-clean splits on TRL2D (Table 19), and the per-space rates are 14–18 of 200, with the 95% intervals of the three learned spaces above 0.05 (Table 61).

Table 56: **Controlled disturbances on OOD data**: families (of 4 perceptual, 11 spectral) detected under the pre-registered rule, i.e. the L5 full-pool disjoint test rejects ($B = 2{,}000$) and the fitted n_{80} lies within the pool; (+k): k more families rejected at L5 without reaching 80% power within the pool. –: not run (Spectrum-native has no perceptual suite; TRL2D has no PCA basis, and its spectral suite uses Spectrum-native in place of Spectrum).

	FSI		HRRR		ERA5		TRL2D	
	perc.	spec.	perc.	spec.	perc.	spec.	perc.	spec.
representation	(pool 26)	(pool 51)	(pool 60)	(pool 120)	(pool 12)	(pool 24)	(pool 45)	(pool 90)
PDE-KD	3/4	5/11 (+3)	4/4	11/11	4/4	10/11 (+1)	3/4	5/11 (+1)
ImageNet	4/4	9/11 (+1)	4/4	9/11	4/4	10/11	4/4	9/11
Poseidon	0/4 (+1)	0/11 (+5)	3/4	7/11 (+1)	2/4 (+1)	8/11	3/4	2/11
Pixels	0/4 (+2)	5/11 (+3)	2/4 (+1)	6/11 (+2)	2/4 (+1)	8/11	2/4 (+1)	3/11 (+1)
Spectrum	3/4	9/11	4/4	8/11	4/4	9/11	4/4	–
Spectrum-native	–	7/11 (+2)	–	8/11	–	7/11	–	5/11 (+1)
PCA-white	1/4 (+1)	6/11 (+2)	3/4 (+1)	8/11	2/4	8/11	–	–
PCA-raw	0/4	0/11	1/4 (+1)	5/11	0/4 (+1)	5/11	–	–

I.3 SIM-TO-REAL GAP

RealPDEBench pairs PIV measurements with CFD twins in 17 (Reynolds number, mass ratio) cells. Each cell has three replicates per arm, which gives a balanced relabeling null with a p-floor of 0.10. Six spaces were scored (PDE-KD, ImageNet, Poseidon, Pixels, Spectrum, Spectrum-native),

Table 57: **FSI (RealPDEBench, PIV arm)**: n_{80} at L5 | power at L2, censored at the pool (26 perceptual, 51 spectral ICs). Format and colors as in Table 26.

	learned						non-learned								
disturbance	PDE-KD		ImageNet		Poseidon		Pixels		Spectrum		Spectrum-native		PCA-white		PCA-raw	
Perceptual: pool 26 ICs																
Occlusion	9	.07	6	**1.00**	26+	.05	26+	.06	16	**.89**	–		8	.09	26+	.06
Gaussian noise	6	**1.00**	6	**1.00**	26+	.08	26+	.06	6	**1.00**	–		26+	.08	26+	.06
Salt-and-pepper	6	**1.00**	6	**1.00**	26+	.05	26+	.06	7	**1.00**	–		26+	.08	26+	.06
Gaussian blur	26+	.07	6	**1.00**	26+	.05	26+	.06	26+	.09	–		26+	.07	26+	.05
Spectral (isotropic): pool 51 ICs																
Low-k injection	51+	.05	51+	.05	51+	.07	51+	.05	51+	.04	51+	.05	51+	.15	51+	.05
Mid-k injection	51+	.06	19	.26	51+	.13	51+	.06	51+	.07	51+	.07	51+	.10	51+	.05
High-k injection	51+	.06	10	**.98**	51+	.04	51+	.05	27	.45	13	.79	51+	.07	51+	.04
Low-k removal	51+	.08	51+	.06	51+	.06	38	.09	9	.04	6	.05	10	.16	51+	.05
Mid-k removal	21	.10	8	.50	40	.20	40	.07	8	.07	6	.14	18	.09	51+	.05
High-k removal	51+	.04	6	**1.00**	51+	.03	51+	.04	6	.76	6	**1.00**	51+	.07	51+	.04
Independent phase	8	.49	6	**1.00**	51+	.05	15	.06	6	.18	51+	.05	8	.09	51+	.05
Shared phase	8	.40	6	**1.00**	51+	.04	16	.06	6	.18	51+	.05	8	.10	51+	.05
Small-scale tilt	6	**.92**	6	**1.00**	51+	.06	22	.07	6	**1.00**	6	**1.00**	9	.13	51+	.05
Large-scale tilt	16	.21	6	**1.00**	51+	.06	51+	.06	6	**1.00**	6	**1.00**	35	.15	51+	.05
Truncation	51+	.05	6	**1.00**	51+	.03	51+	.05	6	**1.00**	6	**1.00**	51+	.07	51+	.04

Table 58: **HRRR**: as above, pools 60 / 120.

	learned						non-learned								
disturbance	PDE-KD		ImageNet		Poseidon		Pixels		Spectrum		Spectrum-native		PCA-white		PCA-raw	
Perceptual: pool 60 ICs																
Occlusion	6	**.84**	6	**1.00**	9	.12	14	.09	8	**.85**	–		8	.12	30	.08
Gaussian noise	7	**.99**	6	**1.00**	25	.21	60+	.11	6	**1.00**	–		41	.12	60+	.08
Salt-and-pepper	6	**1.00**	6	**1.00**	39	.25	12	.61	7	**1.00**	–		8	**.94**	60+	.06
Gaussian blur	6	**1.00**	6	**1.00**	60+	.08	60+	.09	6	**1.00**	–		60+	.15	60+	.05
Spectral (isotropic): pool 120 ICs																
Low-k injection	97	.14	120+	.02	69	.23	120+	.12	120+	.03	120+	.03	33	.59	120+	.07
Mid-k injection	39	.30	8–	**1.00**	52	.28	120+	.10	45	.19	45	.19	30	.75	120+	.07
High-k injection	40	.39	8–	**1.00**	120+	.07	120+	.07	8	**1.00**	8	**1.00**	120+	.04	120+	.07
Low-k removal	37	.17	120+	.02	28	.31	16	.14	8–	.03	8–	.03	8–	**.99**	57	.12
Mid-k removal	21	.37	8–	**1.00**	20	.41	17	.16	8–	**1.00**	8–	**1.00**	8–	**1.00**	74	.11
High-k removal	12	.68	8–	**1.00**	120+	.06	120+	.06	8–	**1.00**	8–	**1.00**	120+	.05	120+	.08
Independent phase	26	.60	8–	**1.00**	49	.13	12	.07	120+	.03	120+	.03	8–	.70	36	.08
Shared phase	25	.65	8–	**1.00**	67	.11	12	.07	120+	.03	120+	.03	8–	**.81**	35	.08
Small-scale tilt	8–	**1.00**	8–	**1.00**	8–	.59	8	**1.00**	8–	**1.00**	8–	**1.00**	8–	**1.00**	30	.13
Large-scale tilt	8–	**1.00**	8–	**1.00**	120+	.12	43	.24	8–	**1.00**	8–	**1.00**	10	**1.00**	120+	.08
Truncation	8	**1.00**	8–	**1.00**	120+	.06	120+	.09	8–	**1.00**	8–	**1.00**	120+	.08	120+	.08

Table 59: **ERA5**: as above, pools 12 / 24.

	learned						non-learned								
disturbance	PDE-KD		ImageNet		Poseidon		Pixels		Spectrum		Spectrum-native		PCA-white		PCA-raw	
Perceptual: pool 12 ICs																
Occlusion	6	.12	6	**1.00**	6	.06	6	.05	7	.33	–		6	.09	12+	.05
Gaussian noise	6	.18	6	**1.00**	8	.13	12+	.06	6	**1.00**	–		9	.12	12+	.04
Salt-and-pepper	6	**1.00**	6	**1.00**	12+	.08	12	.06	7	**1.00**	–		12+	.04	12+	.05
Gaussian blur	6	**1.00**	6	**1.00**	12+	.04	12+	.05	6	**1.00**	–		12+	.04	12+	.04
Spectral (isotropic): pool 24 ICs																
Low-k injection	11	.23	24+	.09	10	.28	11	.23	24+	.07	24+	.07	12	.33	24+	.07
Mid-k injection	10	.22	6	**1.00**	9	.30	18	.13	24+	.07	24+	.09	9	.56	24+	.07
High-k injection	24+	.10	6	**1.00**	24+	.09	24+	.06	8	.76	8	**.87**	24+	.06	24+	.05
Low-k removal	6	.36	12	.10	6	**.83**	6	.58	6	.06	6	.07	6	.65	15	.05
Mid-k removal	7	.26	6	**1.00**	7	.57	6	.23	6	.07	6	.11	6	**.93**	9	.07
High-k removal	9	.12	6	**1.00**	24+	.06	24+	.06	6	**1.00**	6	**1.00**	24+	.05	24+	.06
Independent phase	6	**1.00**	6	**1.00**	6	.17	6	.17	7	.12	24+	.06	6	.43	6	.06
Shared phase	6	**1.00**	6	**1.00**	6	.18	6	.17	7	.12	24+	.06	6	.49	6	.06
Small-scale tilt	6	**1.00**	6	**1.00**	6	.79	6	.52	6	**1.00**	7	**1.00**	6	.74	8	.06
Large-scale tilt	6	**1.00**	6	**1.00**	7	.23	8	.26	6	**1.00**	7	**1.00**	7	.52	24+	.07
Truncation	8	.08	6	**1.00**	24+	.07	24+	.06	6	**1.00**	6	**1.00**	24+	.06	24+	.06

Table 60: **TRL2D (turbulent radiative layer)**: as Table 57, pools 45 / 90. –: not run (no PCA basis exists for TRL2D, and its spectral suite uses Spectrum-native in place of Spectrum).

	learned						non-learned								
disturbance	PDE-KD		ImageNet		Poseidon		Pixels		Spectrum		Spectrum-native		PCA-white	PCA-raw
Perceptual: pool 45 ICs															
Occlusion	8	.14	8–	**1.00**	13	.07	11	.08	8–	**1.00**	–		–	–
Gaussian noise	8–	.37	8–	**1.00**	43	.07	45+	.08	8–	**1.00**	–		–	–
Salt-and-pepper	8–	**1.00**	8–	**1.00**	21	.41	8–	**1.00**	8–	**1.00**	–		–	–
Gaussian blur	45+	.06	8–	**1.00**	45+	.07	45+	.06	8–	**1.00**	–		–	–
Spectral (isotropic): pool 90 ICs															
Low-k injection	90+	.04	90+	.04	90+	.04	90+	.06	–		90+	.05	–	–
Mid-k injection	16	.75	33	.27	90+	.09	45	.16	–		90+	.05	–	–
High-k injection	90+	.05	60	.17	90+	.04	90+	.06	–		90+	.15	–	–
Low-k removal	90+	.04	90+	.04	90+	.05	90+	.06	–		90+	.05	–	–
Mid-k removal	8–	**1.00**	8–	.24	14	.20	10	.23	–		8	.05	–	–
High-k removal	90+	.04	8–	**1.00**	90+	.04	90+	.05	–		8–	.29	–	–
Independent phase	34	.07	8–	**1.00**	90+	.04	90+	.05	–		90+	.05	–	–
Shared phase	36	.07	8–	**1.00**	90+	.04	90+	.05	–		90+	.05	–	–
Small-scale tilt	8	**1.00**	8–	**1.00**	23	.09	10	.20	–		8–	**1.00**	–	–
Large-scale tilt	90+	.11	8–	**1.00**	90+	.08	90+	.10	–		8–	**1.00**	–	–
Truncation	90+	.05	8–	**1.00**	90+	.04	90+	.05	–		8–	**1.00**	–	–

Table 61: **False-positive rates on OOD data.** Top: k of R clean-versus-clean replicates of the design calibration rejected at $\alpha = 0.05$ (disjoint KD; the run with the most replicates). Bottom: the spectral identity rung's rejection rate over the sweep's 200 re-partitions at the pool. Intervals: Clopper–Pearson 95%. –: not run, as in Table 56.

	FSI		HRRR		ERA5		TRL2D	
representation	k/R	95% CI	k/R	95% CI	k/R	95% CI	k/R	95% CI
Clean vs clean, design calibration (disjoint KD)								
PDE-KD	2/40	.006–.169	6/40	.057–.298	1/40	.001–.132	17/200	.050–.133
ImageNet	2/40	.006–.169	3/40	.016–.204	1/40	.001–.132	18/200	.054–.139
Poseidon	4/40	.028–.237	8/40	.091–.356	2/40	.006–.169	18/200	.054–.139
Pixels	3/40	.016–.204	5/40	.042–.268	0/40	.000–.088	16/200	.046–.127
Spectrum	1/40	.001–.132	3/40	.016–.204	1/40	.001–.132	15/200	.043–.121
Spectrum-native	1/40	.001–.132	3/40	.016–.204	1/40	.001–.132	14/200	.039–.115
PCA-white	3/40	.016–.204	4/40	.028–.237	4/40	.028–.237	–	–
PCA-raw	2/40	.006–.169	2/40	.006–.169	1/40	.001–.132	–	–
Identity rung, spectral sweep at the pool								
PDE-KD	11/200	.028–.096	10/200	.024–.090	15/200	.043–.121	7/200	.014–.071
ImageNet	10/200	.024–.090	4/200	.005–.050	14/200	.039–.115	8/200	.017–.077
Poseidon	6/200	.011–.064	12/200	.031–.102	14/200	.039–.115	8/200	.017–.077
Pixels	9/200	.021–.084	11/200	.028–.096	12/200	.031–.102	10/200	.024–.090
Spectrum	8/200	.017–.077	6/200	.011–.064	12/200	.031–.102	–	–
Spectrum-native	9/200	.021–.084	6/200	.011–.064	12/200	.031–.102	9/200	.021–.084
PCA-white	14/200	.039–.115	8/200	.017–.077	10/200	.024–.090	–	–
PCA-raw	8/200	.017–.077	14/200	.039–.115	11/200	.028–.096	–	–

each under two z-score normalizations: each arm's own constants, or constants shared by both arms. Three gates were pre-registered: registration, channel and amplitude. Registration passed. Following the channel gate, the CFD arm's pressure, which PIV does not record, is dropped from the inputs of the learned spaces (leaving slot 0 at exactly zero, as in the PIV arm) and sliced away in Pixels and both spectra. Following the amplitude gate, the shared constants remove the instrumental amplitude offset between the arms.

The no-gap null. Every space places at least 14 of 17 cells above the median of its own null, and four place all 17 there. Combined over cells (Fisher; indicative, because the cells are dependent), every space rejects the no-gap null under shared constants. Under the arms' own constants Spectrum does not ($p = 0.088$), so five of six do. Yet the physical gap is small: 1.15 times the within-arm replicate floor on the mean-velocity-profile oracle.

Cell ordering. With Reynolds number partialled out, we compare each space's cell ordering with two readings of the oracle's gap: the raw gap, and the amplitude-controlled gap (the raw gap divided

by the larger of the two arms' replicate floors). Benjamini–Hochberg correction (Benjamini & Hochberg, 1995) runs over the 12 (space, reading) tests of each normalization. The ordering agrees with the oracle's raw gap for Pixels under both normalizations ($p = 0.029$ and 0.030) and for ImageNet under one of them (0.029 with own constants, 0.070 with shared). The PDE-KD ordering does not significantly agree under either normalization (0.228 and 0.078). No space agrees with the amplitude-controlled gap in that family. The one-sided convention and the 12-test family were both settled after the numbers were seen. Two-sided, the smallest adjusted p is 0.049 (own) and 0.050 (shared), each attained by an agreement; over the six non-learned tests alone, Spectrum-native would agree with the amplitude-controlled gap under shared constants (0.046). ImageNet and PDE-KD give Reynolds trends of opposite sign: $\rho(\mathrm{KD}, \mathrm{Re})$ is $+0.62$ for ImageNet and -0.54 for PDE-KD with own constants ($+0.79$ and -0.79 with shared constants).

I.4 Ranking weather forecasts

WeatherBench-2 provides the only out-of-distribution model ladder. We ask whether KD orders its forecast systems in agreement with their published scores and with a physical arbiter fixed in advance.

Design. We scored 8 forecast systems against ERA5 at 40 lead times: HRES, GraphCast (Lam et al., 2023), Pangu (Bi et al., 2023), FuXi (Chen et al., 2023), the ENS mean, ENS members, persistence and climatology, as provided by WeatherBench-2 (Rasp et al., 2024). The systems were initialized 117 times in 2020 (the ENS members 24 times), and we used four surface variables and the paired design.

KD against the published scores. KD in PDE-KD disagrees with both published scores. Its median Spearman ρ over leads is -0.40 against the published RMSE and -0.89 against the published ACC (seven systems; ACC is negated so that positive ρ means agreement, as for RMSE). Among the four deterministic forecasters, the KD and published-RMSE orders are exact reverses at 36 of the 40 leads (every lead from 5 onward). With four systems no exact test can reach 0.05 (floor 0.083), so the four-system $\rho = -1$ is descriptive. Our pipeline reproduces the published RMSE ordering ($\rho = +0.99$) when pixel error is used. The reversal is not specific to learned features: Spectrum also anti-correlates with the published RMSE. Across the systems whose KD is resolved, PDE-KD and ImageNet prefer forecasts with more high-wavenumber power. On this set of systems that preference cannot be separated from a spectrum closer to the truth's, because the two summaries are near rank-mirrors (median $\rho = -0.84$). KD in Pixels and the published RMSE instead prefer smoother forecasts.

Physical arbiter. As a third ordering, fitted to neither, a physical arbiter was fixed in advance. It uses scalars computed from the fields that are insensitive to blur. At leads 1–20 it orders the systems like the published error and not like KD, with 7 of its 8 pre-registered scalars significant; its far-window reading was withdrawn. Three limits apply to this verdict.

1. The disagreement between the published-error and KD orderings is not itself established in this window: their Spearman correlation, $\rho = -0.24$, is not significantly different from zero ($p = 0.58$ at eight systems), so the arbiter decides between two orderings whose opposition is not established.
2. Persistence's rank on two scalars is fixed by construction, since a frozen analysis has zero diurnal amplitude and zero pressure fall: the diurnal amplitude is one of the eight deciding scalars, and the 24-hour pressure fall is a blur-sensitive scalar that does not decide.
3. The ENS members were initialized 24 times, against 117 times for the other systems.

In addition, three of the deciding scalars carry unresolved validity flags, and the analysis departs from its pre-registration in two ways. A scalar-selection rule, used beside the pre-registered scalar set, was added after pre-registration; the near-window verdict rests on the pre-registered set alone. The pre-registered jackknife rule for excluding orderings that rest on unresolved systems was not implemented.

Scope of the verdict. The result holds for this ladder only: eight systems, four of them deterministic forecasters, one year, four surface variables and one grid. It is also narrower than a verdict on KD. The arbiter is blur-neutral by construction: paired spectral summaries are excluded from it, and its blur-sensitive scalars are reported but do not decide. It can therefore say which forecasts get the blur-insensitive statistics right, but it cannot judge the other reading of the reversal, that KD prefers the forecasts whose spectrum is closer to the truth's. At leads 1–20, KD in PDE-KD therefore does not order these eight forecasts by the blur-insensitive statistics, so on this evidence it should not be used to rank them on accuracy. Whether its preference for high-wavenumber power tracks spectral fidelity cannot be resolved on this ladder.

Further observations. Adding the two 12-hourly GenCast forecast systems (Price et al., 2025) on the even leads leaves the median ρ at -0.44, and the GenCast members sit at or below their own split-half floor at 18 of 20 leads. On the IFS ensemble alone, the ensemble mean beats a member on pixel error at every lead but loses on KD at 78–98% of leads in four of seven spaces. Persistence's KD oscillates with a 24-hour period, with a minimum at every whole-day lead, because a held 00 UTC analysis matches the diurnal phase of the truth once a day.

I.5 Frozen features as regression inputs

We also fit linear (ridge) and MLP probes from frozen features to the phase of year and the phase of day on ERA5 and HRRR, fit on the training period and scored on the test period. The random-initialization floor is one seed of the same network with BatchNorm statistics calibrated on training frames. Like the floor of Appendix H, it uses the feature map's own initialization, with zero-initialized residual scales (Appendix C.2); it is not the Random init control. On ERA5 the frozen features have lower error than that floor for both targets and both heads, but the 95% month-grouped bootstrap intervals separate only for the phase of day with an MLP head (angular error 19.5° [18.2, 21.0] versus 24.7° [23.1, 26.6]; phase of year 11.0° [10.1, 11.9] versus 12.8° [11.7, 13.9]). The intervals do not include the floor's seed-to-seed spread, which one seed cannot measure. On HRRR the intervals overlap for every comparison. Raw pixels pooled to 8×8 beat the frozen features in all eight comparisons (two datasets, two targets, two heads). Fine-tuning the trunk changes this on HRRR, where the diurnal error falls from 22.4° to 5.6°, below pooled pixels' 15.7°. On ERA5 it falls from 19.5° to 0.4°, but pooled pixels still do better (0.1°). Each of these last figures is the better of the two heads. On the phase of day, the fine-tuned trunk therefore beats pooled pixels only on HRRR. These probes measure how well frozen features regress a target; the KD results above instead ask whether the frozen space separates *populations*.

I.6 A held-out-family feature map

We retrained the feature map without shear flow as a supervised-only ablation. We did not use it for the transfer test it would enable: scoring shear flow in a space that never saw that family. The recipe deviates from the released one: a uniform 1/8 family mix, and all 132 epochs at batch 1024 from scratch. On the remaining eight classes, compared with the released feature map re-scored on the same eight, its pooled coefficient MSE and family accuracy are close (0.00151 versus 0.00142; 0.9987 versus 0.9994). Its active-matter coefficient error, however, doubles (0.00426 versus 0.00208), and its regime heads are over-confident (cross-entropy 0.197 versus 0.031). It was exported and passed the same PyTorch-port parity check as the released feature map (Appendix B.2), but it was not used as a KD space, so these supervised numbers are its only evaluation.

J Qualitative examples

Figure 14 illustrates the compression ladder of Appendix G.7. Latent compression removes small-scale structure, and the paired contrasts there show KD preferring less compression where vRMSE and CRPS prefer more (the Lola population KD values in Table 45 are unresolved). The figure's lead 54 lies inside the convective transient, in the 21–60 window of the paired contrasts, before the window of Table 45 (leads 61 onward). Its panels carry no scores; the models' KD and vRMSE are in Table 45. Figure 15 shows single frames of the out-of-distribution data of Appendix I. Figure 16 shows every model of Table 45 at two Rayleigh numbers. In it, Walrus-PT collapses to the conductive profile at $\mathrm{Ra} = 10^{10}$, FNO and TFNO lose the plume structure, and the Lola rollouts become

smoother as compression increases. The disturbance operators are illustrated at L5 in the top row of Table 2.

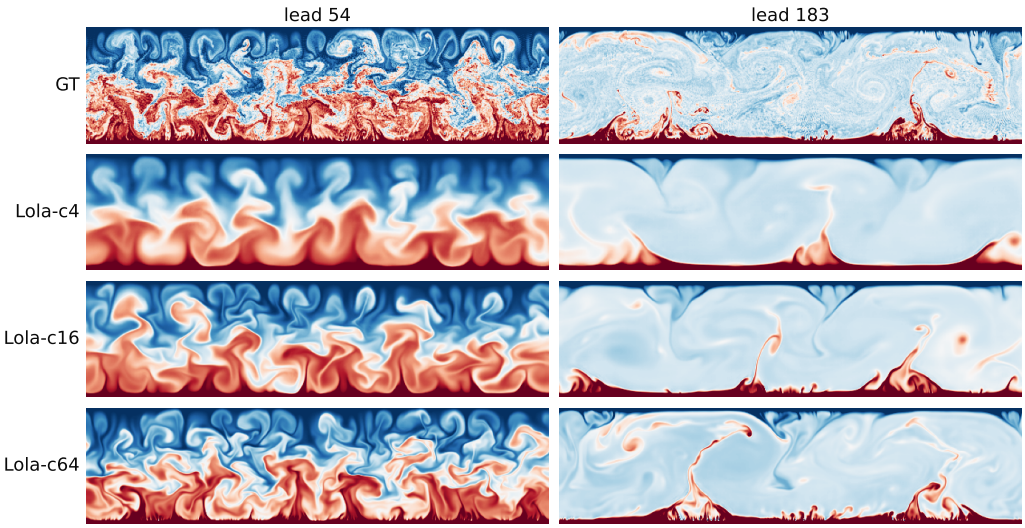


Figure 14: **Latent compression removes small scales.** Reference RB buoyancy (GT; $\mathrm{Ra} = 10^{10}$, $\mathrm{Pr} = 10$, one test IC) and Lola rollouts at three compressions (c4 the most compressed), at leads 54 and 183.

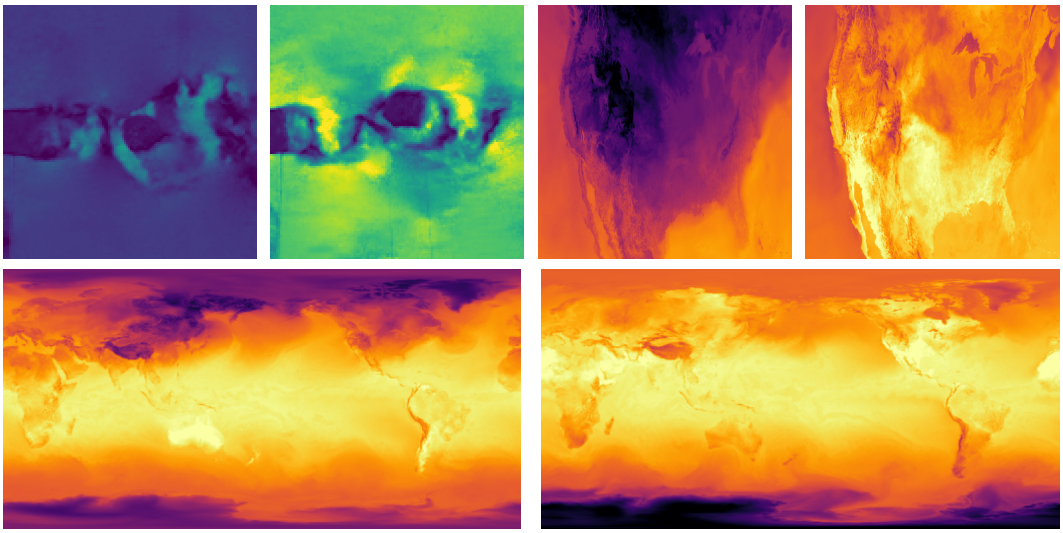


Figure 15: **Out-of-distribution data.** Top, left to right: FSI cylinder wake (RealPDEBench PIV, velocity, frame 130) at $\mathrm{Re} = 3{,}272$ and $9{,}068$; HRRR 2 m temperature, January 06 UTC and July 18 UTC (shown on the square canonical grid). Bottom: ERA5 2 m temperature, January and July 2020, 00 UTC. Color scales are per panel and illustrative.

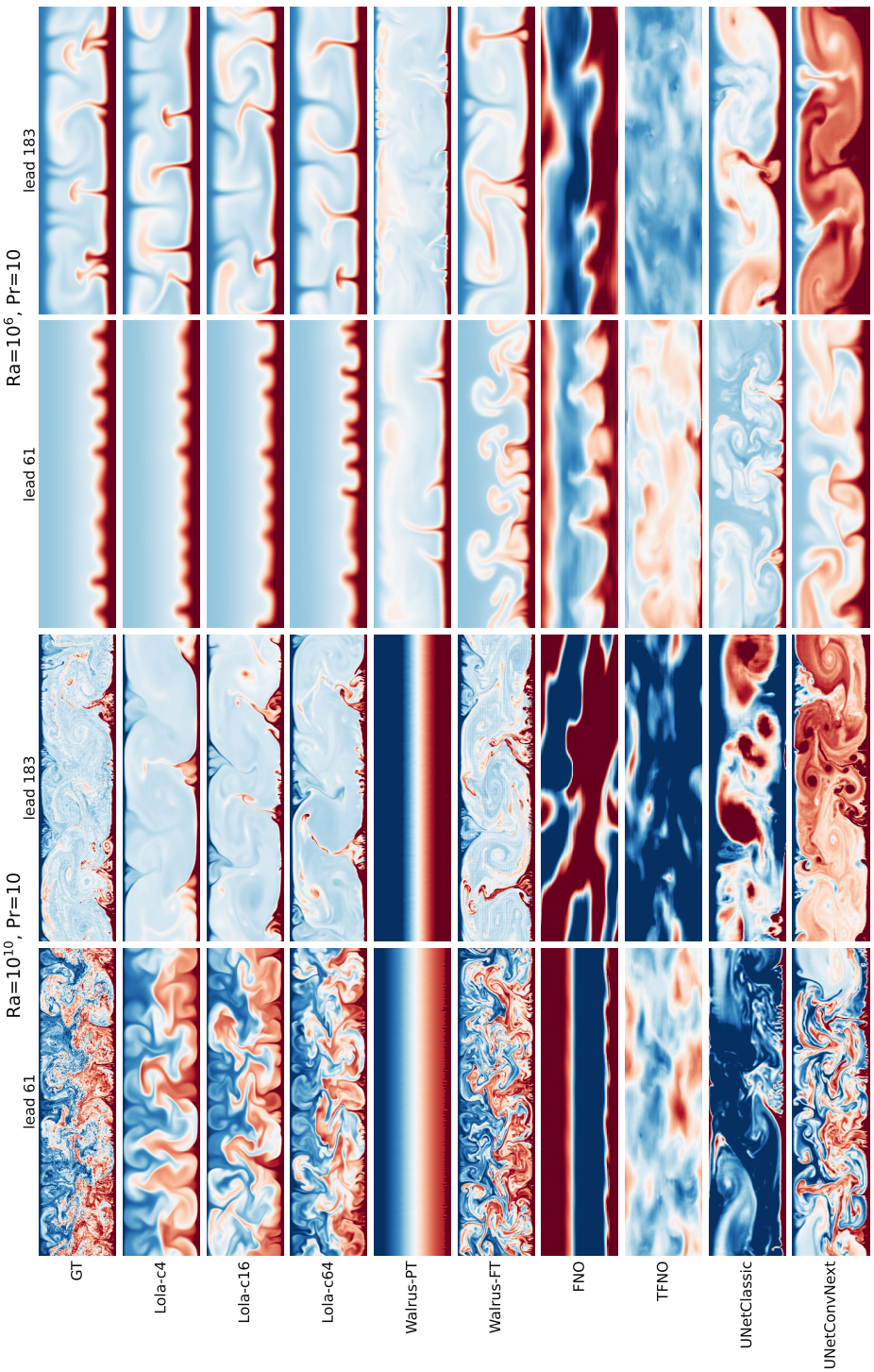


Figure 16: **RB model ladder, qualitatively.** Buoyancy of the reference (GT) and of every model in Table 45, for one test IC at $\mathrm{Ra} = 10^{10}$ (left pair of columns) and $\mathrm{Ra} = 10^{6}$ (right pair), both at $\mathrm{Pr} = 10$, at leads 61 and 183.