LATENT PREDICTION HELPS PHYSICAL REPRESENTATIONS ONLY WHEN IT PREDICTS THE FUTURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Joint-embedding predictive architectures (JEPAs), which predict in a learned latent space, are reported to learn better representations of physical systems than masked autoencoders, which reconstruct pixels. Previous studies, however, change the space of the prediction target together with its time horizon, the architecture and the training pipeline, making it difficult to determine which factor contributes to improved representations. With this motivation, we separate prediction space from prediction horizon. We train ViT-S encoders on three physical simulations from The Well. These encoders are identical except for their objectives. We train on four distinct objectives, varying whether they predict (a) pixels or latent representations and (b) the current eight-frame clip or the next eight-frame clip. Across all encoder layers, we probe representations for regime parameters and for pooled and token-level physical quantities at current and future times. We find that regime probes and future-time probes favour representations trained on the Latent-Future objective, suggesting that latent prediction helps physical representations only when paired with future prediction. Probes for same-time physical quantities provide similar results for all models. We further analyse representations after adding noise to the inputs and find there is not one objective that produces representations more robust to noising. This paper analyses how prediction space and prediction horizon affect learned representations while holding the other training components fixed.

1 INTRODUCTION

Large collections of physical simulations such as The Well (Ohana et al., 2024) have enabled self-supervised training on physics data. Most of this work trains surrogates for next-state prediction (McCabe et al., 2024; Herde et al., 2024; McCabe et al., 2026), but a pretrained encoder is also useful if simple probes can recover physical quantities from its features such as the regime parameters specific to a trajectory, the energy, the locations of sharp gradients, and the future state. Two families of objectives account for most self-supervised pretraining on video-like data; masked autoencoders reconstruct hidden pixels (He et al., 2022; Tong et al., 2022). Joint-embedding predictive architectures instead predict representations of hidden content in a learned latent space (Assran et al., 2023; Bardes et al., 2024). Several studies report that encoders trained with the latent-space prediction objective give more accurate probes of physical quantities: Qu et al. (2026) find that a JEPA trained on simulations from The Well recovers governing parameters better than VideoMAE; Garrido et al. (2025) find that video models predicting in latent space show intuitive physical understanding of the system over pixel-prediction models; and Punzo et al. (2026) show that latent-prediction models have representations that are more layer-wise probeable for physical variables. Typically, the explanation for the performance difference is that a pixel-space reconstruction loss is dominated by uninformative or unpredictable high-variance input components, which the encoder must represent, while a latent-prediction loss does not (Balestriero & LeCun, 2024; Littwin et al., 2024). But these comparisons do not isolate the prediction space. The models they compare differ in architecture, training pipeline and data, and (often in a less visible way) the time that the prediction target refers to. How can we determine which factor contributes to a better model representation if so many components of the training pipeline are changed?

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.

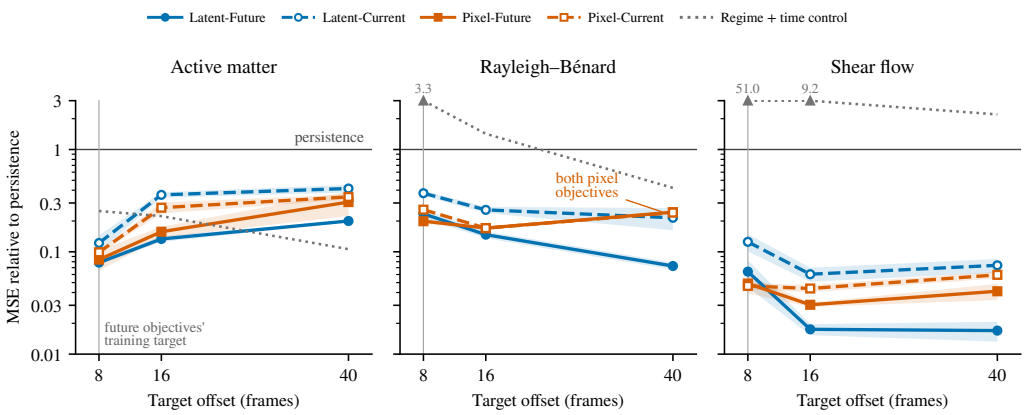

Figure 1: Probe error relative to persistence against how far ahead the probed quantity lies. The y-axis is $1 - S = \mathrm{MSE}_{\mathrm{model}}/\mathrm{MSE}_{\mathrm{persistence}}$ (log scale; the line at 1 is the error of copying the current value forward, lower is better) for pooled probes at offsets 8, 16 and 40 frames, averaged over each system's physical targets (four, five and four) and over three encoder seeds; bands show ± 1 standard deviation over seeds. Each line is one pretraining objective: colour gives the prediction space (blue latent, vermillion pixel), line style the horizon (solid future, dashed current). The grey vertical line marks offset 8, the clip the two future objectives were trained to predict; the grey dotted line is a validation-selected Ridge/MLP regime-and-time control that sees only the regime parameters and clip time (values off scale are printed at the top). At offset 8 the objectives are close and a pixel objective leads in Rayleigh–Bénard and shear flow. Beyond offset 8, Latent-Future has lower error than the other three objectives; at offset 40 its error relative to persistence is 23 to 41% of theirs in Rayleigh–Bénard and shear flow, where the control's error is far higher than every encoder's. In active matter the control has lower error than every encoder at offset 40, so at that offset the targets are largely determined by the regime parameters and the elapsed time.

VideoMAE reconstructs masked parts of the clip it sees. Many latent-prediction models, including V-JEPA 2-AC and latent world models (Assran et al., 2025; Zhou et al., 2024; Hafner et al., 2023), predict representations of the future. If the advantage comes from predicting the future, or the combination of both future prediction and latent prediction, then attributing it to latent-space prediction alone is wrong.

In this paper, we hold everything fixed except the objective, and vary it along two axes (Figure 2): whether the model predicts pixels or latent representations, and whether it predicts the eight-frame clip that is in its context, or the next eight frames. Thus, we have models trained on four objectives: Latent-Current, Latent-Future, Pixel-Current and Pixel-Future. We keep the same: a ViT-S encoder, spacetime patching, 90% tube masking, data and probes. We train three seeds of each on active matter, Rayleigh–Bénard convection and shear flow, and probe the frozen encoders for regime parameters and for physical quantities at the current time and 8, 16 and 40 frames ahead. Future probes are scored by how they perform against a persistence model and compared with a regime-and-time control probe that only has the regime parameters and clip time as inputs. Regime probes favour Latent-Future, whereas on current-time physical quantities the probe performance of the four objectives does not show any significant differences. For probes predicting future quantities, Latent-Future has lower probe error than the other three objectives, and neither changing only the prediction space (Latent-Current) nor changing only the horizon (Pixel-Future) produces this effect (Figure 1). On these systems, the benefit of latent prediction depends on the prediction horizon. Our contributions are:

1. **A controlled comparison of prediction space objectives.** We train self-supervised models with four different objectives — Latent-Current, Pixel-Current, Latent-Future, Pixel-Future — on three physical systems whilst keeping all other components of the training pipeline fixed. This enables us to quantify the contribution of differences in prediction space on model representations.

2. **Latent prediction helps physical representations when paired with future prediction.** Latent-Future has the lowest regime-probe error in all three systems and the best future-state probe performance at long horizons, but neither latent prediction alone nor future prediction alone produces this advantage.

2 Related work

Latent and pixel objectives on physical data. Mialon et al. (2023) learn joint-embedding representations of PDE solutions with Lie point symmetries as augmentations (Brandstetter et al., 2022) and evaluate them by regressing equation coefficients. Garrido et al. (2025) and Punzo et al. (2026) compare pretrained video models that differ in scale, architecture and training data as well as objective. Closest to our work, Qu et al. (2026) compare a JEPA with VideoMAE on The Well and find that JEPA embeddings recover the governing parameters more accurately; their two models differ in architecture, training setup and prediction target at once. Potapov et al. (2026) ask whether the reconstruction loss itself is responsible for the gap. We keep one encoder and one pipeline and change the objective along two factors, so that each factor's effect can be read off directly.

Why latent prediction might help. Balestriero & LeCun (2024) show that reconstruction objectives prioritise high-variance components of the input, which can be uninformative for downstream tasks, and Littwin et al. (2024) show that deep linear JEPAs are implicitly biased towards predictive, low-noise features. Both analyses predict an advantage for latent prediction whatever the horizon. Our same-time physical-target results test that prediction in a setting where only the objective changes.

Predicting the future. Forecasting future observations is an old self-supervised signal (Srivastava et al., 2015), motivated by the idea that prediction favours slowly varying, predictable structure (Wiskott & Sejnowski, 2002; Sobal et al., 2022). Contrastive predictive coding (van den Oord et al., 2018), V-JEPA 2-AC (Assran et al., 2025) and latent world models (Hafner et al., 2023; Zhou et al., 2024) predict future latents using differently trained encoders. SiamMAE (Gupta et al., 2023) used a related objective that reconstructs masked future patches using an unmasked past frame and visible future patches. Our design varies the two independently.

Probing. We read representations with linear and shallow nonlinear probes (Alain & Bengio, 2016). Because an expressive probe can learn a target without the representation encoding it (Hewitt & Liang, 2019; Pimentel et al., 2020), we compare against control predictors that never see the encoder; they play the role of Hewitt & Liang's control tasks for continuous physical targets.

3 Setup

Shared encoder and training. All encoders are 12-block ViT-S models with six attention heads and 384-dimensional tokens (Dosovitskiy et al., 2021; Touvron et al., 2021). An input clip $\mathbf{x} \in \mathbb{R}^{C\times 8\times H\times W}$ of C z-normalised channels is cut into non-overlapping $2\times 16\times 16$ spacetime patches, embedded, and given positional encodings. We apply 90% tube masking: the same spatial mask at every time step, so a masked location stays hidden throughout the clip (Feichtenhofer et al., 2022; Tong et al., 2022). Let V index the visible tokens and M the masked ones. A linear map from 384 to 192 dimensions separates the encoder from every objective-specific head, as in I-JEPA and V-JEPA (Assran et al., 2023; Bardes et al., 2024), and the heads are sized for comparable parameter counts. Each encoder is trained for 100,000 steps with batch size 64; we train three seeds per objective and system.

The four objectives. Latent-Current follows V-JEPA (Bardes et al., 2024): an online encoder f_θ sees the visible tokens, a six-block ViT predictor g_ϕ predicts the target-encoder representation at the masked positions, and the target encoder $f_{\bar\theta}$ is an exponential moving average of f_θ applied to the full clip,

$$\hat{\mathbf{z}}_i = \big[g_\phi\big(f_\theta(\mathbf{x}^{(V)}), V, M\big)\big]_i, \qquad \mathbf{z}_i = \mathrm{LN}\left([f_{\bar\theta}(\mathbf{x})]_i\right), \qquad i \in M,$$

with a smooth-L_1 loss and gradients stopped through $\mathbf{z}_i$. Pixel-Current follows VideoMAE (Tong et al., 2022): a four-block ViT decoder h_ψ predicts the masked patches, each standardised by its own

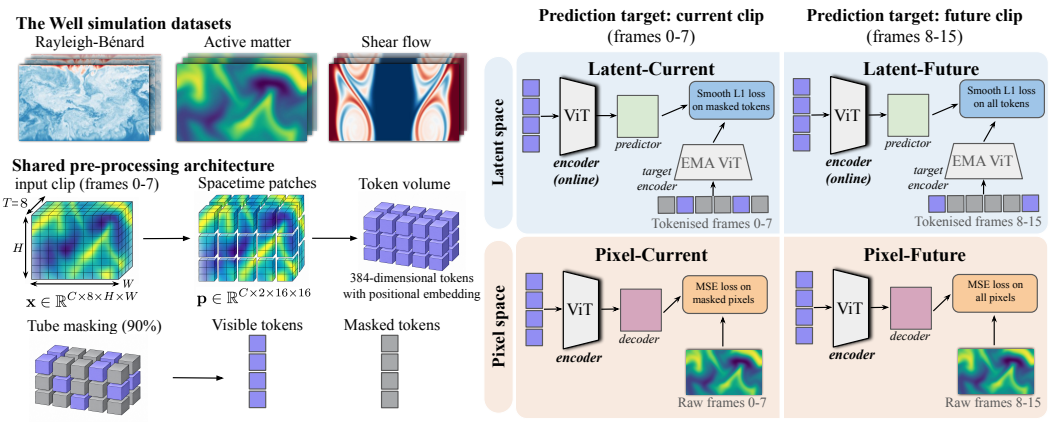

Figure 2: The four objectives. All share the simulation data, the spacetime patching of an eight-frame clip into $2 \times 16 \times 16$ patches, 90% tube masking and a ViT-S encoder (left). They differ only in the prediction head and target (right). Rows set the prediction space: a latent target from an exponential-moving-average (EMA) encoder with a smooth-L_1 loss, or standardised pixels with an MSE loss. Columns set the prediction horizon: the masked tokens of the input clip (frames 0–7), or every token of the next clip (frames 8–15). A difference between two cells can only come from the objective.

mean and variance, $\widetilde{\mathbf{p}}_i = (\mathbf{p}_i - \mathrm{mean}(\mathbf{p}_i))/\sqrt{\mathrm{var}(\mathbf{p}_i) + 10^{-6}}$, under an MSE loss. The two future objectives keep these architectures and losses but move the target to the next clip. The encoder still sees the masked frames 0–7, and the head predicts every token $i \in F$ of frames 8–15: Latent-Future regresses $\mathrm{LN}([f_{\bar{\theta}}(\mathbf{x}_{8:15})]_i)$ from $g_\phi(f_\theta(\mathbf{x}^{(V)}_{0:7}), V, F)$, and Pixel-Future regresses the standardised patches of frames 8–15 from $h_\psi(f_\theta(\mathbf{x}^{(V)}_{0:7}), V, F)$. Neither future objective has any loss on the input clip.

Checkpoints. We probe the final checkpoint for the two latent objectives and the checkpoint with the lowest pretraining-validation loss for the two pixel objectives, with the validation set formed by every eighth training trajectory. The rules differ because the losses differ. The latent loss is measured against an EMA target that changes during training, so its value at one step is not comparable with its value at another and cannot rank checkpoints; the pixel loss has a fixed target, where selecting on held-out loss is standard. In active matter this selects Pixel-Future checkpoints from steps 18,000 to 24,000; all other selected pixel checkpoints lie between steps 90,000 and 100,000.

Data. We use three systems from The Well (Ohana et al., 2024), detailed in Section A: an active nematic suspension (Maddu et al., 2024) with 11 channels, and Rayleigh–Bénard convection and periodic shear flow (Burns et al., 2020) with four channels each. Each system has a grid of regimes set by two parameters: dipole strength α and alignment strength ζ in active matter, Rayleigh and Prandtl numbers in Rayleigh–Bénard, Reynolds and Schmidt numbers in shear flow. Encoders pretrain on The Well training split. Probes are fit on the training split, select every setting on the validation split, and are scored once on the test split. Validation and test regimes also appear in training, so we evaluate on held-out trajectories and initial conditions within the seen regime grid.

Probes. At each of the 13 layer outputs (indices 0–11 follow transformer blocks 1–12, and index 12 follows final normalisation) we fit a Ridge probe and a one-hidden-layer MLP probe, whose prediction is the mean of three initialisations. Pooled probes use the mean over all tokens of a clip; token probes use the representation of one token and predict the value of the target averaged over that token's $2 \times 16 \times 16$ patch, at 64 sampled positions per clip. The layer and hyperparameters of each probe family, and then the family itself, are chosen by validation VRMSE, separately for each encoder and target; test data are not used for any of these choices. The variance-scaled root mean squared error used by The Well is $\mathrm{VRMSE} = \left(\sum_n (y_n - \hat{y}_n)^2 / \sum_n (y_n - \bar{y})^2\right)^{1/2} = \sqrt{1 - R^2}$, so 0 is perfect and 1 is the error of predicting the test mean. At offset 0 we use VRMSE in the main

text and provide both VRMSE and the equivalent R^2 in the appendix. For future offsets the score is skill against persistence,

$$S = 1 - \frac{\mathrm{MSE}_{\text{model}}}{\mathrm{MSE}_{\text{persistence}}} = 1 - \Big(\frac{\mathrm{VRMSE}_{\text{model}}}{\mathrm{VRMSE}_{\text{persistence}}}\Big)^2,$$

where persistence predicts the future target by its value in the input clip; $S = 1$ is perfect, $S = 0$ is no better than persistence and $S < 0$ is worse. Persistence uses the current value as its forecast, so a probe has positive skill only if it predicts the future value with lower MSE than this baseline. At offset 0 the persistence forecast equals the target, so skill is undefined there, and we report mean-referenced VRMSE instead. Persistence can also be a weak reference: in active matter at offset 40 its target-averaged VRMSE is 2.35, far worse than predicting the test mean, so skill alone would overstate how well every encoder predicts there. We therefore report the regime-and-time control below as its own skill beside every future score. All scores are computed per encoder seed and reported as the mean and sample standard deviation over the three seeds; figures plot $1 - S$, the error relative to persistence, on a log axis.

Targets and offsets. For each system we probe the two regime parameters (logarithms of the dimensionless numbers) and four or five derived quantities that combine gradients, second derivatives and products of channels: kinetic energy, enstrophy, nematic order and nematic gradient energy in active matter; enstrophy, buoyancy gradient energy, convective flux, pressure gradient magnitude and buoyancy Laplacian magnitude in Rayleigh–Bénard; cross-stream kinetic energy, enstrophy, tracer variance and tracer gradient energy in shear flow (Section B). The encoder always sees frames 0–7 of a context. A physical target is computed on the eight frames starting at offset 0, 8, 16 or 40 frames after the context start. Offset 8 is the clip the future objectives were trained to predict; offsets 16 and 40 lie beyond every objective's training target, and we use probes at these offsets to measure how much information about the future state a representation contains. Pooled probes use three contexts per trajectory and token probes one.

Controls. We compare encoder probes with controls that do not use encoder representations. *Persistence*, the reference for skill, predicts a future target by its value at offset 0. The pooled *regime-and-time* control uses the two regime parameters and clip time (context start divided by trajectory length minus one); the token-level *regime, time, and position* control also uses the token's time and spatial coordinates. For each target, offset and resolution, we select a Ridge or MLP control by validation VRMSE before test scoring, using the same selection procedure as for the encoder probes (Section C). These controls measure how much can be predicted without the state fields or encoder features. An encoder that does not beat these controls has not shown that it represents the state of the particular trajectory.

Noise. To test the common argument that latent-prediction encoders do not represent noisy, unpredictable detail, we add Gaussian noise of standard deviation $\sigma \in \{0.05, 0.1, 0.2, 0.5, 1\}$ to the input clips at test time, with three corruption seeds, and score the probes fitted on clean data against the clean targets. We use pooled probes at each objective's own training horizon (offset 0 for current and 8 for future objectives), so the two groups are compared at different horizons.

4 Results

4.1 Regime probes favour Latent-Future; same-time physical probes do not

Table 1 reports the individual-clip regime probes. Changing only the prediction space gives no consistent benefit: Latent-Current is worse than Pixel-Current in active matter (MSE 0.054 against 0.039) and within seed spread of it in Rayleigh–Bénard and shear flow. The combination of latent prediction and a future target separates instead. Latent-Future has the lowest MSE in all three systems: 0.016 in active matter, 0.067 in Rayleigh–Bénard and 0.178 in shear flow. Relative to the next-best objective, its MSE is lower by 33%, 68% and 46%, respectively. The parameter-level VRMSEs show the same strong advantage in Rayleigh–Bénard and shear flow; in active matter, Pixel-Current has the lowest ζ VRMSE while Latent-Future has the lowest α VRMSE and joint MSE. The comparison with Qu et al. (2026) is not like-for-like because their encoders, inputs and attentive probes differ, but it puts the fixed-scale errors in context.

Table 1: Individual-clip regime-parameter test VRMSE and system-mean MSE (mean ± standard deviation over encoder seeds). Each eight-frame clip is a separate example. For each parameter and encoder seed, the layer, probe hyperparameters and probe family (Ridge or MLP) are selected by validation VRMSE before test scoring. The corresponding test R^2 values are reported in Table 5. The published JEPA and VideoMAE system means from Qu et al. (2026, Table 1) are included only as references: their encoders, 16-frame sampling and attentive probes differ from ours. Lower is better throughout; bold marks the best objective in each row.

System	Parameter VRMSE / MSE	LC	LF	PC	PF
Active matter					
	α VRMSE	0.332 ± 0.018	$\mathbf{0.124 \pm 0.020}$	0.294 ± 0.015	0.175 ± 0.006
	ζ VRMSE	0.144 ± 0.009	0.137 ± 0.034	$\mathbf{0.100 \pm 0.009}$	0.153 ± 0.015
	Our MSE	0.054 ± 0.004	$\mathbf{0.016 \pm 0.003}$	0.039 ± 0.004	0.024 ± 0.003
	Qu et al. MSE	JEPA 0.079		VideoMAE 0.160	
Rayleigh–Bénard					
	$\log_{10}$ Ra VRMSE	0.251 ± 0.064	$\mathbf{0.033 \pm 0.007}$	0.320 ± 0.000	0.320 ± 0.000
	$\log_{10}$ Pr VRMSE	0.514 ± 0.012	$\mathbf{0.302 \pm 0.007}$	0.516 ± 0.001	0.515 ± 0.001
	Our MSE	0.210 ± 0.021	$\mathbf{0.067 \pm 0.002}$	0.227 ± 0.002	0.225 ± 0.003
	Qu et al. MSE	JEPA 0.130		VideoMAE 0.180	
Shear flow					
	$\log_{10}$ Re VRMSE	0.498 ± 0.012	$\mathbf{0.263 \pm 0.001}$	0.475 ± 0.018	0.417 ± 0.013
	$\log_{10}$ Sc VRMSE	0.500 ± 0.004	$\mathbf{0.340 \pm 0.004}$	0.458 ± 0.006	0.429 ± 0.005
	Our MSE	0.439 ± 0.066	$\mathbf{0.178 \pm 0.006}$	0.424 ± 0.094	0.331 ± 0.011
	Qu et al. MSE	JEPA 0.380		VideoMAE 0.670	

By contrast, probes for the physical quantities at offset 0 show no latent advantage. Across the 13 pooled targets, every objective has VRMSE at most 0.160; Latent-Current and Pixel-Current differ by a mean absolute VRMSE of 0.023 (median 0.019), and no pair of objectives differs by more than 0.024 on average. Where differences exceed the seed spread they favour pixel prediction: Pixel-Current has the lowest error on all four active-matter targets and three of the four shear-flow targets, five of these seven by more than the seed spread. These probes use clip-specific information, since every encoder has lower error than the regime-and-time control on every offset-0 target (target-averaged VRMSE of the best objective 0.043, 0.082 and 0.022, against 0.285, 0.324 and 0.933 for the control, in active matter, Rayleigh–Bénard and shear flow). But all four objectives have uniformly low VRMSE on these targets, so they cannot rank the objectives.

4.2 Beyond the training horizon, Latent-Future separates

Moving the target forward in time separates the objectives (Figure 1). At offset 8, the clip both future objectives were trained to predict, a pixel objective leads: Pixel-Future in Rayleigh–Bénard (target-averaged skill 0.801, against 0.765 for Latent-Future) and Pixel-Current in shear flow (0.954 against 0.936). The latent-future lead appears beyond the training target. At offset 16 Latent-Future has the highest skill on 10 of the 13 pooled targets, six by more than the seed spread; at offset 40 it has the highest skill on 12 of 13, ten by more than the seed spread (Table 2; all offsets in Table 6). The lead grows with the horizon. Dividing the best other objective's error relative to persistence ($1 - S$) by Latent-Future's gives 0.85, 1.16 and 2.96 at offsets 8, 16 and 40 in Rayleigh–Bénard, 0.73, 1.74 and 2.42 in shear flow, and 1.07, 1.18 and 1.53 in active matter. At offset 40 Latent-Future's probe squared error is 7.3% of persistence's in Rayleigh–Bénard and 1.7% in shear flow, against 21 to 24% and 4.1 to 7.4% for the other objectives. Per target, it predicts the Rayleigh–Bénard buoyancy Laplacian magnitude with skill 0.971, against 0.847 for Latent-Current and 0.759 for both pixel objectives, and the shear-flow tracer gradient energy with 0.944, against 0.748, 0.799 and 0.858 for Latent-Current, Pixel-Current and Pixel-Future.

Token-level probes, which predict the future target averaged over one patch, show smaller differences between the objectives. At offset 40 Latent-Future has the highest skill on four of the five Rayleigh–Bénard token targets, but Pixel-Future has it on three of the four shear-flow token targets (Table 6). In active matter the four objectives' target-averaged token skill at offset 40 lies between

Table 2: Pooled physical-target skill $S = 1 - \mathrm{MSE}_{\mathrm{model}}/\mathrm{MSE}_{\mathrm{persistence}}$ at offset 40 frames (higher is better). Cells are mean ± standard deviation over three encoder seeds; the regime-and-time control is seed-independent. Bold marks the best objective, and † marks a lead larger than the greater seed standard deviation of the best two objectives. AM, RB, and SF denote active matter, Rayleigh–Bénard, and shear flow. Latent-Future has the highest skill on every Rayleigh–Bénard and shear-flow target, by more than the seed standard deviation, while in active matter the objectives differ less.

System	Target	LC	LF	PC	PF	Reg./time
AM	Kinetic E.	-0.061 ± 0.142	$\mathbf{0.476 \pm 0.051}^{\dagger}$	0.053 ± 0.074	0.079 ± 0.330	0.666
	Enstrophy	0.609 ± 0.028	$\mathbf{0.935 \pm 0.024}$	0.725 ± 0.023	0.911 ± 0.003	0.985
	Nematic order	0.994 ± 0.000	$\mathbf{0.995 \pm 0.001}$	0.995 ± 0.001	0.994 ± 0.000	0.996
	Nematic gradient E.	0.796 ± 0.034	0.794 ± 0.025	$\mathbf{0.851 \pm 0.019}^{\dagger}$	0.790 ± 0.056	0.929
RB	Enstrophy	0.546 ± 0.046	$\mathbf{0.842 \pm 0.023}^{\dagger}$	0.601 ± 0.006	0.596 ± 0.012	0.641
	Buoyancy gradient E.	0.848 ± 0.080	$\mathbf{0.953 \pm 0.000}^{\dagger}$	0.762 ± 0.000	0.762 ± 0.000	0.832
	Convective flux	0.810 ± 0.045	$\mathbf{0.932 \pm 0.002}^{\dagger}$	0.769 ± 0.002	0.769 ± 0.002	0.721
	Pressure gradient	0.873 ± 0.009	$\mathbf{0.938 \pm 0.005}^{\dagger}$	0.901 ± 0.002	0.894 ± 0.003	-0.177
	Buoyancy Laplacian	0.847 ± 0.083	$\mathbf{0.971 \pm 0.002}^{\dagger}$	0.759 ± 0.000	0.759 ± 0.000	0.874
SF	Cross-stream kinetic E.	0.996 ± 0.001	$\mathbf{0.998 \pm 0.000}^{\dagger}$	0.996 ± 0.000	0.997 ± 0.000	-0.805
	Enstrophy	0.988 ± 0.001	$\mathbf{0.995 \pm 0.001}^{\dagger}$	0.990 ± 0.001	0.994 ± 0.001	-1.370
	Tracer variance	0.974 ± 0.006	$\mathbf{0.995 \pm 0.001}^{\dagger}$	0.977 ± 0.004	0.986 ± 0.003	-2.482
	Tracer gradient E.	0.748 ± 0.042	$\mathbf{0.944 \pm 0.015}^{\dagger}$	0.799 ± 0.022	0.858 ± 0.028	-0.149

0.486 and 0.512. Nematic order is the exception: at offset 16 Latent-Future predicts it per token with skill 0.920, against 0.589 (Latent-Current), 0.617 (Pixel-Current) and 0.699 (Pixel-Future).

4.3 A Controlled Comparison of Objectives

The 2×2 design lets us change one factor at a time. We take Pixel-Current as the reference and measure the target-averaged pooled $1 - S$ at offset 40. Predicting the future in pixel space changes it by $+0.9\%$ in Rayleigh–Bénard, -31% in shear flow and -11% in active matter. Predicting the present in latent space changes it by -11%, $+24\%$ and $+21\%$. Predicting the future in latent space lowers it by 70%, 72% and 42%. If the two factors acted independently, the combined change would be close to the product of the single changes (-10%, -14% and $+7.6\%$) or to their sum (-10%, -6.9% and $+9.9\%$). In every system it is far larger, which is what an interaction between the factors means. In active matter the regime-and-time control beats all four objectives at this offset (Section 4.4), so any comparison there is among encoders that fall short of the control anyway.

Rayleigh–Bénard shows the pattern most clearly. The two pixel objectives agree to three decimals at offset 40 on buoyancy gradient energy (VRMSE 0.609 and 0.609) and buoyancy Laplacian magnitude (0.664 and 0.664), and so have the same skill, 0.762 and 0.759: on these targets, training to predict the next clip in pixel space did not change how well a probe on the encoder predicts the state 40 frames ahead. On these two targets both pixel objectives are below the regime-and-time control (skill 0.832 and 0.874), while Latent-Future reaches 0.953 and 0.971. Latent-Current's gain over the pixel objectives in this system comes mostly from the same two targets and is unstable across seeds: its skill has seed standard deviation 0.080 and 0.083 there, against under 0.003 for Latent-Future. Predicting the future in pixel space recovers part of the gap in two places, shear-flow tracer gradient energy (skill 0.858, against 0.799 for Pixel-Current and 0.944 for Latent-Future) and active-matter kinetic energy at offset 16 (0.611, against 0.217 for Pixel-Current and 0.661 for Latent-Future). Latent space alone, then, gives no consistent benefit; future prediction alone gives a partial benefit that depends on the system; the two together give the largest benefit in every system.

4.4 Controls show where the test is informative

A future-state probe measures trajectory-specific information only if the control's inputs, the regime parameters and clip time, do not already predict the target. In Rayleigh–Bénard and shear flow at offset 40, the encoders outperform this control on average across targets: Latent-Future's target-averaged skill is 0.927 and 0.983, against 0.578 and -1.202 for the regime-and-time control (Table 4). In these systems persistence is hard to beat: at offset 8 the regime-and-time control has skill -2.274 and -50.039. Every encoder has much lower error than persistence at offset 40 (target-

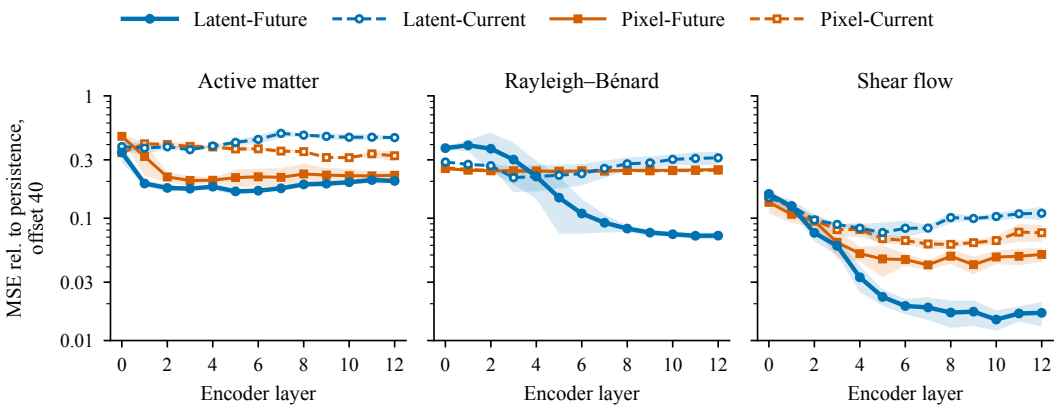

Figure 3: Where the offset-40 information sits in the encoder. Error relative to persistence, $1 - S$ (log scale, lower is better), of the pooled MLP probe at each layer , averaged over each system's physical targets and over three encoder seeds; bands show ± 1 seed standard deviation. Colours and line styles as in Figure 1. Latent-Future starts no better than the others at layer 0 and improves through the last layers in Rayleigh–Bénard and shear flow, while Latent-Current is best at layers 3 to 5 and then gets worse, so the objectives differ most in how probe error changes with layer, which a single best-layer score does not show.

averaged skill above 0.75), so the encoders contain information about state changes. Here the probes measure trajectory-specific information, and Latent-Future's lead is largest.

Active matter is different. At offset 40 the regime-and-time control has target-averaged skill 0.894, against 0.800 for Latent-Future, the best objective; Latent-Future does not beat the control on any of the four targets. On this same system, Latent-Future achieves lower regime-probe error than reported by Qu et al. (2026), suggesting that its failure to beat the regime-and-time control at offset 40 does not reflect a general failure to learn useful representations. For example, the control reaches 0.985 on enstrophy and 0.666 on kinetic energy, against 0.935 and 0.476 for Latent-Future. Forty frames span half of an 81-frame active-matter trajectory, and by then the pooled quantities are largely set by regime and elapsed time. This is why the objectives differ least here at offset 40, and why we compare them in active matter at offset 16. There the control's skill is 0.777, Latent-Future (0.867) beats it on two of four targets while Latent-Current and Pixel-Current beat it on one and two, respectively, and on kinetic energy Latent-Future reaches 0.661, against 0.150, 0.217 and 0.611 for Latent-Current, Pixel-Current and Pixel-Future and 0.448 for the control.

4.5 Future information sits in Latent-Future's last layers

Where in the encoder does the long-horizon information appear? Figure 3 traces the offset-40 probes through depth. At layer 0 Latent-Future is no better than the other objectives (target-averaged MLP $1 - S$ of 0.374, against 0.25 to 0.29 in Rayleigh–Bénard; 0.158 against 0.135 to 0.153 in shear flow). It then improves to its best at layer 11 in Rayleigh–Bénard (0.072) and layer 10 in shear flow (0.015). Latent-Current is best early, at layer 3 in Rayleigh–Bénard (0.215) and layer 5 in shear flow (0.076), and gets worse towards the output (0.312 and 0.110 at layer 12). The pixel objectives change little with depth in Rayleigh–Bénard (best 0.241 and 0.242 at layer 5, against 0.258 and 0.255 at layer 0) and are best at layers 7 to 8 in shear flow. With Ridge probes the late-layer pattern holds in Rayleigh–Bénard (Latent-Future best at layer 11) but not in shear flow, where Latent-Future's Ridge error is lowest at layer 5 and changes little afterwards (Figure 6); active matter has no consistent depth pattern. We do not test these descriptive profiles for significance. For MLP probes, the difference between objectives appears in the later encoder blocks, which a single best-layer score hides.

4.6 Noise: no uniform latent advantage

If latent prediction helped by ignoring noisy detail, latent representations should tolerate input noise better. The data do not show this uniformly (Figure 7). We report target-averaged VRMSE,

because each objective is scored at its own training offset against clean targets and we ask how far noise moves it from its clean score. In active matter latent objectives are more tolerant: at $\sigma = 1$ Latent-Current and Latent-Future reach 0.34 and 0.57, while Pixel-Current and Pixel-Future reach 2.25 and 1.27. In Rayleigh–Bénard every objective exceeds the test-mean error (VRMSE 1) by $\sigma = 0.5$, and the future objectives degrade first: at $\sigma = 0.05$ Latent-Future reaches 0.80 and Pixel-Future 0.74, against 0.29 and 0.33 for Latent-Current and Pixel-Current. In shear flow Pixel-Future is the most tolerant at small σ (VRMSE 0.06 at $\sigma = 0.05$ and 0.23 at $\sigma = 0.1$). Current and future objectives are scored at different horizons, so the cleaner comparison is latent against pixel at equal horizon. Latent prediction wins those by a wide margin only in active matter; elsewhere the order depends on the system and σ.

5 Discussion and limitations

What this means for choosing and evaluating objectives. On these systems, predicting latents instead of pixels improves probes of the future physical state, but only when the prediction target is the next clip rather than the current one. One possible explanation is that predicting future latents favours features that remain informative over time, whereas reconstructing the input rewards details that need not help predict its subsequent evolution. Our results do not test this directly, but they do demonstrate that current-state accuracy alone is probably insufficient when evaluating physical representations. For a representation of dynamics, the results from our study suggest that the objective should predict latent futures. On evaluation, it is also clear that probing regime parameters is not sufficient to determine if the models are really learning physics (as demonstrated by the results for active matter). Including simple baselines beyond persistence, like our regime-and-time control, can help differentiate these differences.

A hypothesis for why only latent-future helps. We have no direct evidence for why the combination works, and offer one hypothesis consistent with the theory above. Reconstruction weights input components by their variance (Balestriero & LeCun, 2024). Suppose that, in convection and shear flow, much of that variance sits at small scales that are poorly predictable eight frames ahead. Then a pixel-future loss is dominated by detail the encoder cannot predict, and the encoder gains little from it, which would explain why Pixel-Future matches Pixel-Current in Rayleigh–Bénard. A latent future target is itself learned, so training could reduce the loss by making the target consist of predictable features, such as slowly varying structure that still describes the state 40 frames later.

Limitations. The study uses one architecture (ViT-S), 100,000 training steps, three two-dimensional systems and three seeds per objective, and it is possible the results do not hold more generally. The checkpoint rule differs between latent and pixel objectives for the reason given in Section 3. We also use a fixed 90% tube masking strategy for every model. This makes the comparison more fair, but the decision to mask JEPA models with future prediction targets is somewhat unusual. The main result seems to be that the prediction horizon matters, but we only compare current clip and next clip objectives. It's possible that predicting even further ahead during training could improve representations.

6 Conclusion

We perform a controlled comparison of four self-supervised objectives on three physical systems, varying prediction space (pixels or latents) and prediction horizon (the current or next clip), and evaluate the learned representations through regime-parameter and physical-quantity probes. When only the objective changes, latent prediction of the present is no better than pixel reconstruction on same-time physical-quantity probes, and the advantage on future physical targets appears only when latent prediction targets the future and the probe looks beyond the training horizon. The individual-clip regime probes support the same interaction with Latent-Current performing worse than Pixel-Current in active matter and within seed spread of it in Rayleigh–Bénard and shear flow, whereas Latent-Future achieves the lowest MSE in all three systems. These results suggest that, while latent space prediction is often the most emphasized feature of JEPA models, the choice of prediction horizon can be just as important for physical representation learning.

REPRODUCIBILITY STATEMENT

Section 3 and Sections A and C give the full protocol: data, splits, architectures, objectives, checkpoint rule, probe families and hyperparameter grids, selection rule, targets, offsets, controls and the noise sweep. We will release the per-model score files for all 36 encoders, the aggregated summaries from which every number in the paper is computed, and the scripts that produce every figure and table from those files.

AI USE STATEMENT

In this work, we used generative AI tools (Claude Code and Codex) to provide feedback on our research methodology and experimental design, to implement methods, to clean and reformat datasets, and to support qualitative and thematic data analysis. The tools gave feedback only on methodology and experiments; the research design itself was developed by the authors. Additionally, we used generative AI tools to create and edit software code. We have reviewed all AI-assisted work. All AI-generated code was reviewed and tested for correctness by the authors, and dataset cleaning outputs were spot-checked against the raw data. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

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

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Randall Balestriero and Yann LeCun. How learning by reconstruction produces uninformative features for perception. *arXiv preprint arXiv:2402.11337*, 2024.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-JEPA: Latent video prediction for visual representation learning, 2024. URL https://openreview.net/forum?id=WFYbBOEOtv.

Johannes Brandstetter, Max Welling, and Daniel E. Worrall. Lie point symmetry data augmentation for neural PDE solvers. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Quentin Garrido, Nicolas Ballas, Mahmoud Assran, Adrien Bardes, Laurent Najman, Michael Rabbat, Emmanuel Dupoux, and Yann LeCun. Intuitive physics understanding emerges from self-supervised pretraining on natural videos. *arXiv preprint arXiv:2502.11831*, 2025.

Agrim Gupta, Jiajun Wu, Jia Deng, and Li Fei-Fei. Siamese masked autoencoders. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

Maximilian Herde, Bogdan Raonić, Tobias Rohner, Roger Käppeli, Roberto Molinaro, Emmanuel de Bézenac, and Siddhartha Mishra. Poseidon: Efficient foundation models for pdes, 2024. URL https://arxiv.org/abs/2405.19101.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Etai Littwin, Omid Saremi, Madhu Advani, Vimal Thilak, Preetum Nakkiran, Chen Huang, and Joshua Susskind. How JEPA avoids noisy features: The implicit bias of deep linear self distillation networks. *arXiv preprint arXiv:2407.03475*, 2024. URL https://arxiv.org/abs/2407.03475.

Suryanarayana Maddu, Scott Weady, and Michael J Shelley. Learning fast, accurate, and stable closures of a kinetic theory of an active fluid. *Journal of Computational Physics*, 504:112869, 2024.

Michael McCabe, Bruno Régaldo-Saint Blancard, Liam Holden Parker, Ruben Ohana, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Siavash Golkar, Geraud Krawezik, Francois Lanusse, Mariel Pettee, Tiberiu Tesileanu, Kyunghyun Cho, and Shirley Ho. Multiple physics pretraining for physical surrogate models, 2024. URL https://arxiv.org/abs/2310.02994.

Michael McCabe, Payel Mukhopadhyay, Tanya Marwah, Bruno Regaldo-Saint Blancard, Francois Rozet, Cristiana Diaconu, Lucas Meyer, Kaze W. K. Wong, Hadi Sotoudeh, Alberto Bietti, Irina Espejo, Rio Fear, Siavash Golkar, Tom Hehir, Keiya Hirashima, Geraud Krawezik, Francois 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, 2026. URL https://arxiv.org/abs/2511.15684.

Grégoire Mialon, Quentin Garrido, Hannah Lawrence, Danyal Rehman, Yann LeCun, and Bobak T. Kiani. Self-supervised learning with lie symmetries for partial differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina Agocs, Miguel Beneitez, Marsha Berger, Blakesly Burkhart, Stuart Dalziel, Drummond Fielding, et al. The well: a large-scale collection of diverse physics simulations for machine learning. *Advances in Neural Information Processing Systems*, 37:44989–45037, 2024.

Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

Alexey Potapov, Oleg Shcherbakov, and Ivan Kravchenko. Is the reconstruction loss culprit? an attempt to outperform JEPA. *arXiv preprint arXiv:2603.14131*, 2026. URL https://arxiv.org/abs/2603.14131.

Samuele Punzo, Niccolò Caselli, Ippokratis Pantelidis, Francesco Massafra, Salvatore Lo Sardo, and Mohammadreza Salehi. Do video foundation models understand intuitive physics? a layerwise probing analysis. *arXiv preprint arXiv:2606.09646*, 2026.

Helen Qu, Rudy Morel, Michael McCabe, Alberto Bietti, François Lanusse, Shirley Ho, and Yann LeCun. Representation learning for spatiotemporal physical systems. *arXiv preprint arXiv:2603.13227*, 2026. URL https://arxiv.org/abs/2603.13227.

Vlad Sobal, Jyothir S V, Siddhartha Jalagam, Nicolas Carion, Kyunghyun Cho, and Yann LeCun. Joint embedding predictive architectures focus on slow features. *arXiv preprint arXiv:2211.10831*, 2022.

Nitish Srivastava, Elman Mansimov, and Ruslan Salakhutdinov. Unsupervised learning of video representations using LSTMs. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked Autoencoders Are Data-Efficient Learners for Self-Supervised Video Pre-Training. In *Neural Information Processing Systems*, 2022. URL https://mlanthology.org/neurips/2022/tong2022neurips-videomae/.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pp. 10347–10357. PMLR, 2021.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Laurenz Wiskott and Terrence J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770, 2002.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.

A SIMULATION DATASETS

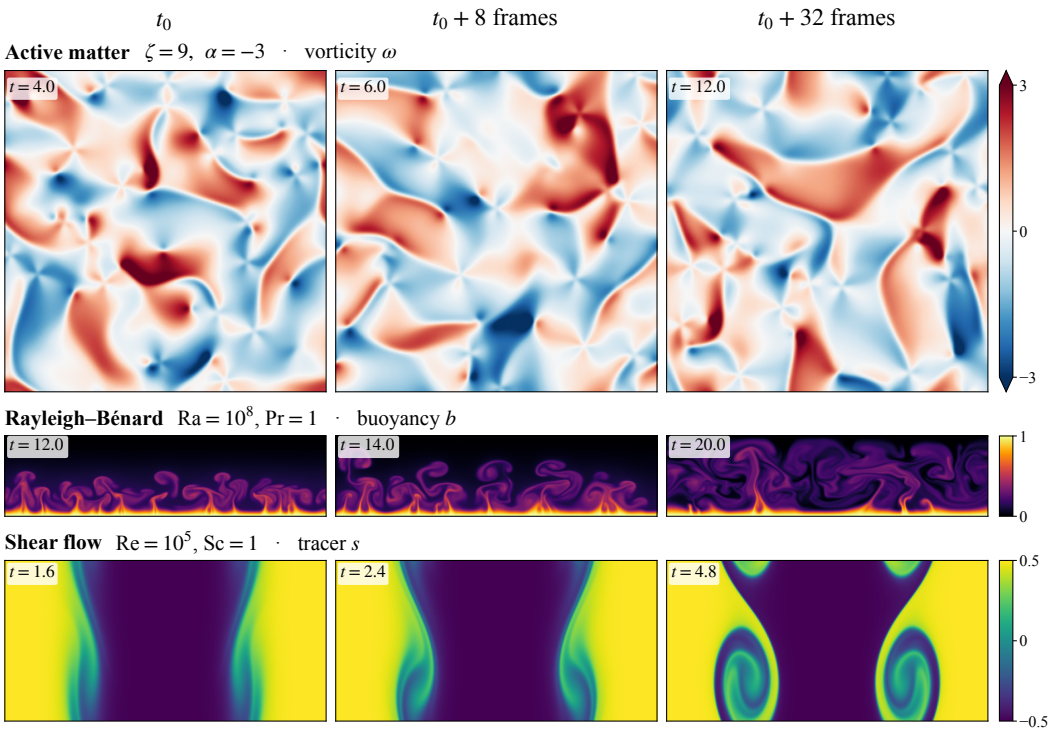

Figure 4: One test trajectory from each system at the start of a context clip (t_0) and 8 and 32 stored frames later; each panel gives its physical time. Active matter shows vorticity computed from the velocity field, Rayleigh–Bénard shows buoyancy, and shear flow shows the passive tracer. Because the systems are stored at different intervals, the same frame offset spans different physical times: 32 frames are 8 time units in active matter and Rayleigh–Bénard and 3.2 in shear flow.

We use three datasets from The Well (Ohana et al., 2024): an active suspension model (Maddu et al., 2024), Rayleigh–Bénard convection, and periodic shear flow, the latter two generated with Dedalus (Burns et al., 2020). They differ in governing physics, channel count, boundary conditions and grid geometry (Table 3; examples in Figure 4).

	Active matter	Rayleigh–Bénard	Shear flow
Channels C	11	4	4
Fields	c, $\mathbf{u}$, $\mathbf{D}$, $\mathbf{E}$	b, p, $\mathbf{u}$	s, p, $\mathbf{u}$
Grid points ($x \times y$)	256×256	512×128	256×512
Frames per rollout	81	200	200
Domain	$[0, 10]^2$	$x \in [0, 4)$, $y \in [0, 1]$	$x \in [0, 1)$, $y \in [-1, 1]$
Boundaries	periodic	periodic x, Dirichlet y	periodic
Grid geometry	uniform	uniform x, Chebyshev y	uniform
Stored Δt	0.25	0.25	0.1
Time range	$[0, 20]$	$[0, 50]$	$[0, 20]$
Regime parameters	α, ζ	Ra, Pr	Re, Sc
Regimes	45	35	28
Rollouts	225	1750	1120
Train / valid / test rollouts	175 / 24 / 26	1400 / 175 / 175	896 / 112 / 112
Regimes in train / valid / test	45 / 16 / 21	35 / 35 / 35	28 / 28 / 28
Tokens per clip	1024	1024	2048

Table 3: The three systems. Token counts are for $2 \times 16 \times 16$ spacetime patches over an eight-frame clip. Every validation and test regime also appears in training, so in all three systems the evaluation uses held-out trajectories from regimes seen in training.

Active matter. A continuum theory for rod-like active particles suspended in a Stokes fluid, following Equations (1)–(5) of Maddu et al. (2024). The state is a concentration field c, velocity $\mathbf{u}$, orientation tensor $\mathbf{D}$ and strain-rate tensor $\mathbf{E}$, giving 11 channels in 2D. Simulations start from uniform concentration $c(\mathbf{x}, 0) = 1$ and a plane-wave perturbation of the orientation tensor about the isotropic state, and are periodic in both directions. Snapshots are stored every 0.25 time units over $t \in [0, 20]$, giving 81 frames per rollout. Regimes are set by the dipole strength $\alpha \in \{-1, -2, -3, -4, -5\}$ and the alignment strength $\zeta \in \{1, 3, \ldots, 17\}$, with the particle-density parameter $\beta = 0.8$ fixed; each of the 45 regimes has five rollouts that differ only in the initialisation of $\{c, \mathbf{D}, \mathbf{u}\}$. The system passes from an isotropic to a nematic state as ζ increases.

Rayleigh–Bénard convection. A fluid layer heated from below and cooled from above, solving the nondimensional incompressible Boussinesq equations

$$\partial_t b + \mathbf{u} \cdot \nabla b = \kappa \nabla^2 b, \tag{1}$$

$$\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} = -\nabla p + \nu \nabla^2 \mathbf{u} + b\, \mathbf{e}_y, \tag{2}$$

$$\nabla \cdot \mathbf{u} = 0, \tag{3}$$

with $\kappa = (\mathrm{Ra}\,\mathrm{Pr})^{-1/2}$ and $\nu = (\mathrm{Ra}/\mathrm{Pr})^{-1/2}$. The four channels are buoyancy b, pressure p and the two velocity components; pressure uses a zero-mean gauge. The domain is periodic in x, with fixed buoyancy ($b = L_y$ at $y = 0$, $b = 0$ at $y = L_y$) and no-slip velocity at the walls $y = 0$ and $y = L_y$. Buoyancy starts as a linear background profile scaled by $\delta b_0 \in \{0.2, \ldots, 1.0\}$ plus Gaussian white noise of scale 10^{-3}; velocity and pressure start at zero. Regimes are the 35 combinations of $\mathrm{Ra} \in \{10^6, \ldots, 10^{10}\}$ and $\mathrm{Pr} \in \{0.1, 0.2, 0.5, 1, 2, 5, 10\}$, with 50 initial conditions each. The arrays are uniformly spaced in x and sampled at Chebyshev nodes in y.

Periodic shear flow. Two-dimensional incompressible flow with a passive tracer,

$$\partial_t \mathbf{u} + \mathbf{u} \cdot \nabla \mathbf{u} = -\nabla p + \nu \nabla^2 \mathbf{u}, \tag{4}$$

$$\partial_t s + \mathbf{u} \cdot \nabla s = D \nabla^2 s, \tag{5}$$

with $\nu = 1/\mathrm{Re}$ and $D = \nu/\mathrm{Sc}$, again with a zero-mean pressure gauge. The four channels are the tracer s, pressure p and the two velocity components, on a uniform grid periodic in both directions. The initial horizontal velocity consists of n_{shear} hyperbolic-tangent shear layers

spaced along y, with sinusoidal vertical perturbations that decay away from each layer; the tracer matches the shear at initialisation and the pressure is zero. Regimes are the 28 combinations of $\mathrm{Re} \in \{10^4, 5{\times}10^4, 10^5, 5{\times}10^5\}$ and $\mathrm{Sc} \in \{0.1, 0.2, 0.5, 1, 2, 5, 10\}$, with 40 initial conditions each, indexed by the number of shear layers $n_{\text{shear}} \in \{2, 4\}$, the number of perturbation blobs $n_{\text{blobs}} \in \{2, \ldots, 5\}$ and a width factor $w \in \{0.25, 0.5, 1, 2, 4\}$. At the larger Reynolds numbers the shear layers are unstable, so trajectories within one regime diverge quickly.

Splits and clips. For each system, encoders pretrain on The Well training split, with every eighth training trajectory reserved for pretraining validation and checkpoint selection. Probes are fit on the training split, select on the validation split, and are scored once on the test split (split sizes in Table 3). Channels are z-normalised with The Well statistics before entering the encoder; physical targets are computed from the raw fields. In Rayleigh–Bénard and shear flow, pooled contexts start at frames 0, 76 and 152 of each trajectory, and token contexts cycle through starts 0, 38, 76, 114 and 152. All four objectives use identical trajectories, clips, normalisation and partitions.

Temporal sampling. The systems are stored at different intervals and for different durations, so a fixed frame offset corresponds to different physical horizons. Offset 40 is 10 time units in active matter, half of the 20-unit rollout; 10 time units in Rayleigh–Bénard, a fifth of the rollout; and 4 time units in shear flow, also a fifth of the rollout. Active matter also allows the fewest contexts at offset 40: with 81 stored frames, a context whose target window starts 40 frames later has far fewer possible start positions than in the other two systems.

B PHYSICAL VARIABLES

Each target is computed from the raw simulation fields at the original resolution and then averaged over the eight-frame target window, either over the whole domain (pooled) or over one token's $2 \times 16 \times 16$ patch (token). We chose quantities that depend on gradients, second derivatives or products of channels, so that none is a linear function of the input channels.

Active matter. *Kinetic energy*: the energy of the fluid velocity field, set by how strongly the active stresses drive flow. *Enstrophy*: the squared vorticity, which measures the strength of the swirling flows that active nematics generate. *Nematic order*: the degree of local alignment of the particles, as read from the orientation tensor $\mathbf{D}$; it separates isotropic from nematic states. *Nematic gradient energy*: the spatial variation of the orientation field, which is large near defects and domain walls.

Rayleigh–Bénard convection. *Enstrophy*: the squared vorticity of the convective flow. *Buoyancy gradient energy*: the squared buoyancy gradient, which is large in thermal boundary layers and plume edges. *Convective flux*: the vertical transport of buoyancy by the flow, the quantity behind heat transport. *Pressure gradient magnitude*: the size of the pressure forces that redirect the flow. *Buoyancy Laplacian magnitude*: how quickly the buoyancy gradient changes, which is largest at the edges of thin thermal features and sets the diffusive term of the buoyancy equation. Derivatives in y respect the Chebyshev grid.

Shear flow. *Cross-stream kinetic energy*: the energy of the velocity component across the shear layers, which grows as the layers become unstable. *Enstrophy*: the squared vorticity, concentrated in the shear layers and the vortices they roll up into. *Tracer variance*: how far the passive tracer is from being fully mixed. *Tracer gradient energy*: the squared tracer gradient, which is large along the thin filaments that stirring produces and falls as diffusion smooths them.

Regime parameters. Active-matter probes predict α and ζ directly; the other systems' probes predict $\log_{10}$ of the Rayleigh and Prandtl numbers and of the Reynolds and Schmidt numbers. Each trajectory contributes three separate eight-frame clips, beginning at frames 0, 16 and 33 in active matter and at frames 0, 76 and 152 in Rayleigh–Bénard and shear flow. This gives train/validation/test sample counts of 525/72/78, 4200/525/525 and 2688/336/336, respectively. For the comparison with Qu et al. (2026) in Table 1, predictions are rescored as MSE on fixed label scales: α 1.41 and ζ 5.16; $\log_{10}\mathrm{Ra}$ 1.41 and physical Pr 3.38; $\log_{10}\mathrm{Re}$ 0.61 and physical Sc 3.38. Prandtl and Schmidt predictions are exponentiated for this score only, and this rescoring never determines probe selection.

Table 4: Pooled target-averaged persistence skill $S = 1 - \mathrm{MSE}_{\mathrm{model}}/\mathrm{MSE}_{\mathrm{persistence}}$ at future offsets (higher is better). "Best" is the objective with the highest skill averaged equally over targets; the LF column counts the targets on which Latent-Future's skill exceeds the regime-and-time control's. At offset 40, the regime-and-time control has higher skill than the best objective in active matter, while in Rayleigh–Bénard and shear flow Latent-Future has the highest skill.

System	Offset	Best S	LF S	Reg./time S	Best objective	LF beats control
Active matter	8	0.922 ± 0.004	0.922 ± 0.004	0.750	LF	3/4
	16	0.867 ± 0.008	0.867 ± 0.008	0.777	LF	2/4
	40	0.800 ± 0.006	0.800 ± 0.006	0.894	LF	0/4
Rayleigh–Bénard	8	0.801 ± 0.011	0.765 ± 0.023	-2.274	PF	5/5
	16	0.853 ± 0.011	0.853 ± 0.011	-0.424	LF	5/5
	40	0.927 ± 0.006	0.927 ± 0.006	0.578	LF	5/5
Shear flow	8	0.954 ± 0.003	0.936 ± 0.019	-50.039	PC	4/4
	16	0.983 ± 0.002	0.983 ± 0.002	-8.242	LF	4/4
	40	0.983 ± 0.004	0.983 ± 0.004	-1.202	LF	4/4

C PROBE AND CONTROL DETAILS

Probes. Ridge probes choose the regularisation strength from $\{10^{-5}, 10^{-4}, \ldots, 10^{2}\}$. MLP probes have one hidden layer of 128 units with dropout 0.1, learning rate 0.01 and weight decay 10^{-4}, and train for 150 to 2000 steps; each MLP prediction is the mean of three initialisations. Within each family the layer and hyperparameters are chosen by validation VRMSE. For each encoder and target, the family with the lower validation VRMSE is then fixed before test scoring, with ties favouring Ridge. This selection is performed separately for every regime parameter as well as every physical target.

Controls. The regime-and-time and regime, time, and position controls use both regime parameters (α and ζ directly in active matter, and $\log_{10}$ parameters in Rayleigh–Bénard and shear flow) and clip time, defined as context start divided by trajectory length minus one. The regime, time, and position control additionally uses token time, y and x coordinates scaled to $[-1, 1]$. Each target, offset and resolution has Ridge and MLP candidates, fitted with the same routines and hyperparameter settings as the encoder probes. Ridge uses a quadratic expansion, including squares and all pairwise interactions; the MLP uses the unexpanded control inputs, standardised using training data only. Validation VRMSE selects the family, with ties favouring Ridge, before test scoring. One fitted control is shared across all encoders and seeds, so it has no encoder-seed standard deviation. Persistence uses the target's value in the offset-0 window as the prediction for a later window; it is the reference for the skill S, which is undefined at offset 0 because persistence has zero error there. The regime-and-time control is scored by the same skill so that the two can be read side by side. Table 4 compares the pooled regime-and-time control with the best objective and with Latent-Future for every system and future offset.

D ADDITIONAL RESULTS

Table 5 gives the test R^2 corresponding to the regime-parameter VRMSEs in Table 1. Table 6 gives the skill against persistence of every pooled and token cell at offsets 8, 16 and 40, with seed standard deviations and the control's skill. Tables 24 and 36 give every pooled and token cell at all offsets in VRMSE and R^2. Figure 5 shows the VRMSE results as heatmaps. Figure 6 repeats Figure 3 with Ridge probes, and Figure 7 shows the noise sweep behind Section 4.6.

Table 5: Individual-clip regime-parameter test R^2 corresponding to Table 1 (mean $\pm$ standard deviation over encoder seeds). For each parameter and encoder seed, the probe is the one selected by validation VRMSE before test scoring. Here R^2 uses the test-set mean as its reference; higher is better, and bold marks the best objective in each row.

System	Parameter	LC	LF	PC	PF
Active matter					
	α	0.889 ± 0.012	$\mathbf{0.984 \pm 0.005}$	0.913 ± 0.009	0.969 ± 0.002
	ζ	0.979 ± 0.002	0.980 ± 0.010	$\mathbf{0.990 \pm 0.002}$	0.977 ± 0.005
Rayleigh–Bénard					
	$\log_{10}\mathrm{Ra}$	0.934 ± 0.034	$\mathbf{0.999 \pm 0.001}$	0.897 ± 0.000	0.898 ± 0.000
	$\log_{10}\mathrm{Pr}$	0.736 ± 0.012	$\mathbf{0.909 \pm 0.004}$	0.734 ± 0.001	0.735 ± 0.001
Shear flow					
	$\log_{10}\mathrm{Re}$	0.752 ± 0.012	$\mathbf{0.931 \pm 0.001}$	0.774 ± 0.017	0.826 ± 0.011
	$\log_{10}\mathrm{Sc}$	0.750 ± 0.004	$\mathbf{0.884 \pm 0.003}$	0.790 ± 0.005	0.816 ± 0.005

Table 6: Active matter pooled persistence skill S at target-start offset 8 frames. Skill is one minus model MSE divided by persistence MSE. Values are test mean $\pm$ standard deviation across three encoder seeds; the control is seed-independent. Bold marks the best objective and † a lead beyond the larger of the best two seed standard deviations. The persistence reference has skill zero; higher values indicate better prediction of future change.

Target	LC	LF	PC	PF	Reg./time
Kinetic energy	0.752 ± 0.049	0.832 ± 0.019	0.775 ± 0.022	$\mathbf{0.845 \pm 0.053}$	0.357
Enstrophy	0.907 ± 0.034	$\mathbf{0.972 \pm 0.012}$	0.955 ± 0.011	0.971 ± 0.008	0.980
Nematic order	0.970 ± 0.013	$\mathbf{0.994 \pm 0.001}$	0.987 ± 0.004	0.991 ± 0.003	0.900
Nematic gradient energy	0.884 ± 0.018	$\mathbf{0.888 \pm 0.008}$	0.887 ± 0.001	0.857 ± 0.028	0.761

Table 7: Active matter pooled persistence skill S at target-start offset 16 frames. Skill is one minus model MSE divided by persistence MSE. Values are test mean $\pm$ standard deviation across three encoder seeds; the control is seed-independent. Bold marks the best objective and † a lead beyond the larger of the best two seed standard deviations. The persistence reference has skill zero; higher values indicate better prediction of future change.

Target	LC	LF	PC	PF	Reg./time
Kinetic energy	0.150 ± 0.069	$\mathbf{0.661 \pm 0.010}^{\dagger}$	0.217 ± 0.091	0.611 ± 0.033	0.448
Enstrophy	0.739 ± 0.033	$\mathbf{0.957 \pm 0.016}$	0.874 ± 0.036	0.945 ± 0.022	0.988
Nematic order	0.852 ± 0.025	$\mathbf{0.993 \pm 0.001}^{\dagger}$	0.945 ± 0.016	0.979 ± 0.004	0.799
Nematic gradient energy	0.818 ± 0.024	0.855 ± 0.023	$\mathbf{0.885 \pm 0.015}^{\dagger}$	0.837 ± 0.038	0.871

Table 8: Active matter pooled persistence skill S at target-start offset 40 frames. Skill is one minus model MSE divided by persistence MSE. Values are test mean $\pm$ standard deviation across three encoder seeds; the control is seed-independent. Bold marks the best objective and † a lead beyond the larger of the best two seed standard deviations. The persistence reference has skill zero; higher values indicate better prediction of future change.

Target	LC	LF	PC	PF	Reg./time
Kinetic energy	-0.061 ± 0.142	$\mathbf{0.476 \pm 0.051}^{\dagger}$	0.053 ± 0.074	0.079 ± 0.330	0.666
Enstrophy	0.609 ± 0.028	$\mathbf{0.935 \pm 0.024}$	0.725 ± 0.023	0.911 ± 0.003	0.985
Nematic order	0.994 ± 0.000	$\mathbf{0.995 \pm 0.001}$	0.995 ± 0.001	0.994 ± 0.000	0.996
Nematic gradient energy	0.796 ± 0.034	0.794 ± 0.025	$\mathbf{0.851 \pm 0.019}^{\dagger}$	0.790 ± 0.056	0.929

Table 9: Rayleigh–Bénard pooled persistence skill S at target-start offset 8 frames. Skill is one minus model MSE divided by persistence MSE. Values are test mean ± standard deviation across three encoder seeds; the control is seed-independent. Bold marks the best objective and † a lead beyond the larger of the best two seed standard deviations. The persistence reference has skill zero; higher values indicate better prediction of future change.

Target	LC	LF	PC	PF	Reg./time
Enstrophy	-0.238 ± 0.129	0.218 ± 0.081	0.230 ± 0.046	$\mathbf{0.489 \pm 0.061}^{\dagger}$	-1.465
Buoyancy gradient energy	0.891 ± 0.006	$\mathbf{0.930 \pm 0.010}^{\dagger}$	0.864 ± 0.013	0.863 ± 0.007	0.518
Convective flux	0.899 ± 0.006	0.922 ± 0.011	0.919 ± 0.006	$\mathbf{0.952 \pm 0.004}^{\dagger}$	0.719
Pressure gradient magnitude	0.708 ± 0.029	0.832 ± 0.024	0.839 ± 0.010	$\mathbf{0.851 \pm 0.009}^{\dagger}$	-11.814
Buoyancy Laplacian magnitude	0.870 ± 0.012	$\mathbf{0.922 \pm 0.012}^{\dagger}$	0.857 ± 0.020	0.850 ± 0.010	0.672

Table 10: Rayleigh–Bénard pooled persistence skill S at target-start offset 16 frames. Skill is one minus model MSE divided by persistence MSE. Values are test mean ± standard deviation across three encoder seeds; the control is seed-independent. Bold marks the best objective and † a lead beyond the larger of the best two seed standard deviations. The persistence reference has skill zero; higher values indicate better prediction of future change.

Target	LC	LF	PC	PF	Reg./time
Enstrophy	0.188 ± 0.107	$\mathbf{0.573 \pm 0.041}$	0.536 ± 0.008	0.517 ± 0.011	0.116
Buoyancy gradient energy	0.953 ± 0.008	$\mathbf{0.987 \pm 0.001}^{\dagger}$	0.952 ± 0.001	0.948 ± 0.003	0.708
Convective flux	0.866 ± 0.005	0.876 ± 0.004	0.873 ± 0.001	$\mathbf{0.889 \pm 0.007}^{\dagger}$	0.811
Pressure gradient magnitude	0.758 ± 0.015	$\mathbf{0.842 \pm 0.019}$	0.831 ± 0.003	0.839 ± 0.008	-4.641
Buoyancy Laplacian magnitude	0.950 ± 0.004	$\mathbf{0.986 \pm 0.002}^{\dagger}$	0.957 ± 0.002	0.954 ± 0.001	0.884

Table 11: Rayleigh–Bénard pooled persistence skill S at target-start offset 40 frames. Skill is one minus model MSE divided by persistence MSE. Values are test mean ± standard deviation across three encoder seeds; the control is seed-independent. Bold marks the best objective and † a lead beyond the larger of the best two seed standard deviations. The persistence reference has skill zero; higher values indicate better prediction of future change.

Target	LC	LF	PC	PF	Reg./time
Enstrophy	0.546 ± 0.046	$\mathbf{0.842 \pm 0.023}^{\dagger}$	0.601 ± 0.006	0.596 ± 0.012	0.641
Buoyancy gradient energy	0.848 ± 0.080	$\mathbf{0.953 \pm 0.000}^{\dagger}$	0.762 ± 0.000	0.762 ± 0.000	0.832
Convective flux	0.810 ± 0.045	$\mathbf{0.932 \pm 0.002}^{\dagger}$	0.769 ± 0.002	0.769 ± 0.002	0.721
Pressure gradient magnitude	0.873 ± 0.009	$\mathbf{0.938 \pm 0.005}^{\dagger}$	0.901 ± 0.002	0.894 ± 0.003	-0.177
Buoyancy Laplacian magnitude	0.847 ± 0.083	$\mathbf{0.971 \pm 0.002}^{\dagger}$	0.759 ± 0.000	0.759 ± 0.000	0.874

Table 12: Shear flow pooled persistence skill S at target-start offset 8 frames. Skill is one minus model MSE divided by persistence MSE. Values are test mean ± standard deviation across three encoder seeds; the control is seed-independent. Bold marks the best objective and † a lead beyond the larger of the best two seed standard deviations. The persistence reference has skill zero; higher values indicate better prediction of future change.

Target	LC	LF	PC	PF	Reg./time
Cross-stream kinetic energy	0.996 ± 0.001	0.998 ± 0.000	0.997 ± 0.000	$\mathbf{0.999 \pm 0.000}^{\dagger}$	-5.454
Enstrophy	0.955 ± 0.008	$\mathbf{0.975 \pm 0.007}$	0.961 ± 0.001	0.973 ± 0.003	-22.374
Tracer variance	0.883 ± 0.017	0.924 ± 0.010	$\mathbf{0.945 \pm 0.006}^{\dagger}$	0.936 ± 0.008	-160.130
Tracer gradient energy	0.665 ± 0.076	0.847 ± 0.072	$\mathbf{0.911 \pm 0.010}$	0.896 ± 0.027	-12.198

Table 13: Shear flow pooled persistence skill S at target-start offset 16 frames. Skill is one minus model MSE divided by persistence MSE. Values are test mean ± standard deviation across three encoder seeds; the control is seed-independent. Bold marks the best objective and † a lead beyond the larger of the best two seed standard deviations. The persistence reference has skill zero; higher values indicate better prediction of future change.

Target	LC	LF	PC	PF	Reg./time
Cross-stream kinetic energy	0.997 ± 0.001	0.998 ± 0.000	0.997 ± 0.000	$\mathbf{0.999 \pm 0.000}$	-1.797
Enstrophy	0.977 ± 0.001	$\mathbf{0.989 \pm 0.003}$	0.979 ± 0.001	0.986 ± 0.001	-5.917
Tracer variance	0.948 ± 0.014	$\mathbf{0.983 \pm 0.002}^{\dagger}$	0.955 ± 0.003	0.975 ± 0.001	-23.580
Tracer gradient energy	0.837 ± 0.028	$\mathbf{0.960 \pm 0.010}^{\dagger}$	0.893 ± 0.013	0.919 ± 0.005	-1.674

Table 14: Shear flow pooled persistence skill S at target-start offset 40 frames. Skill is one minus model MSE divided by persistence MSE. Values are test mean $\pm$ standard deviation across three encoder seeds; the control is seed-independent. Bold marks the best objective and † a lead beyond the larger of the best two seed standard deviations. The persistence reference has skill zero; higher values indicate better prediction of future change.

Target	LC	LF	PC	PF	Reg./time
Cross-stream kinetic energy	0.996 ± 0.001	$\mathbf{0.998 \pm 0.000}^{\dagger}$	0.996 ± 0.000	0.997 ± 0.000	-0.805
Enstrophy	0.988 ± 0.001	$\mathbf{0.995 \pm 0.001}^{\dagger}$	0.990 ± 0.001	0.994 ± 0.001	-1.370
Tracer variance	0.974 ± 0.006	$\mathbf{0.995 \pm 0.001}^{\dagger}$	0.977 ± 0.004	0.986 ± 0.003	-2.482
Tracer gradient energy	0.748 ± 0.042	$\mathbf{0.944 \pm 0.015}^{\dagger}$	0.799 ± 0.022	0.858 ± 0.028	-0.149

Table 15: Active matter token persistence skill S at target-start offset 8 frames. Skill is one minus model MSE divided by persistence MSE. Values are test mean $\pm$ standard deviation across three encoder seeds; the control is seed-independent. Bold marks the best objective and † a lead beyond the larger of the best two seed standard deviations. The persistence reference has skill zero; higher values indicate better prediction of future change.

Target	LC	LF	PC	PF	Reg./time + position
Kinetic energy	0.337 ± 0.037	$\mathbf{0.370 \pm 0.008}$	0.304 ± 0.012	0.364 ± 0.060	0.342
Enstrophy	0.369 ± 0.010	0.360 ± 0.006	0.366 ± 0.005	$\mathbf{0.370 \pm 0.007}$	0.389
Nematic order	0.747 ± 0.058	$\mathbf{0.919 \pm 0.008}$	0.820 ± 0.024	0.908 ± 0.011	0.869
Nematic gradient energy	0.358 ± 0.017	$\mathbf{0.369 \pm 0.001}$	0.358 ± 0.012	0.366 ± 0.013	0.386

Table 16: Active matter token persistence skill S at target-start offset 16 frames. Skill is one minus model MSE divided by persistence MSE. Values are test mean $\pm$ standard deviation across three encoder seeds; the control is seed-independent. Bold marks the best objective and † a lead beyond the larger of the best two seed standard deviations. The persistence reference has skill zero; higher values indicate better prediction of future change.

Target	LC	LF	PC	PF	Reg./time + position
Kinetic energy	0.364 ± 0.022	$\mathbf{0.412 \pm 0.017}$	0.327 ± 0.014	0.385 ± 0.048	0.447
Enstrophy	0.457 ± 0.006	$\mathbf{0.461 \pm 0.003}$	0.442 ± 0.018	0.459 ± 0.004	0.482
Nematic order	0.589 ± 0.030	$\mathbf{0.920 \pm 0.004}^{\dagger}$	0.617 ± 0.003	0.699 ± 0.015	0.594
Nematic gradient energy	0.408 ± 0.011	0.412 ± 0.007	$\mathbf{0.423 \pm 0.010}^{\dagger}$	0.409 ± 0.009	0.455

Table 17: Active matter token persistence skill S at target-start offset 40 frames. Skill is one minus model MSE divided by persistence MSE. Values are test mean $\pm$ standard deviation across three encoder seeds; the control is seed-independent. Bold marks the best objective and † a lead beyond the larger of the best two seed standard deviations. The persistence reference has skill zero; higher values indicate better prediction of future change.

Target	LC	LF	PC	PF	Reg./time + position
Kinetic energy	0.220 ± 0.011	$\mathbf{0.259 \pm 0.023}^{\dagger}$	0.170 ± 0.007	0.187 ± 0.010	0.363
Enstrophy	$\mathbf{0.400 \pm 0.001}$	0.397 ± 0.015	0.390 ± 0.005	0.393 ± 0.011	0.428
Nematic order	0.947 ± 0.002	$\mathbf{0.962 \pm 0.001}^{\dagger}$	0.948 ± 0.001	0.952 ± 0.002	0.952
Nematic gradient energy	$\mathbf{0.444 \pm 0.016}$	0.430 ± 0.002	0.437 ± 0.010	0.426 ± 0.008	0.495

Table 18: Rayleigh–Bénard token persistence skill S at target-start offset 8 frames. Skill is one minus model MSE divided by persistence MSE. Values are test mean ± standard deviation across three encoder seeds; the control is seed-independent. Bold marks the best objective and † a lead beyond the larger of the best two seed standard deviations. The persistence reference has skill zero; higher values indicate better prediction of future change.

Target	LC	LF	PC	PF	Reg./time + position
Enstrophy	0.482 ± 0.025	0.518 ± 0.008	0.511 ± 0.009	$\mathbf{0.546 \pm 0.005}^{\dagger}$	0.427
Buoyancy gradient energy	0.382 ± 0.044	$\mathbf{0.582 \pm 0.011}^{\dagger}$	0.412 ± 0.018	0.420 ± 0.005	-0.301
Convective flux	0.390 ± 0.009	0.480 ± 0.081	0.528 ± 0.007	$\mathbf{0.609 \pm 0.012}^{\dagger}$	-1.940
Pressure gradient magnitude	0.422 ± 0.005	0.454 ± 0.001	0.458 ± 0.004	$\mathbf{0.463 \pm 0.002}^{\dagger}$	0.285
Buoyancy Laplacian magnitude	0.324 ± 0.048	$\mathbf{0.571 \pm 0.006}^{\dagger}$	0.387 ± 0.019	0.396 ± 0.021	-0.044

Table 19: Rayleigh–Bénard token persistence skill S at target-start offset 16 frames. Skill is one minus model MSE divided by persistence MSE. Values are test mean ± standard deviation across three encoder seeds; the control is seed-independent. Bold marks the best objective and † a lead beyond the larger of the best two seed standard deviations. The persistence reference has skill zero; higher values indicate better prediction of future change.

Target	LC	LF	PC	PF	Reg./time + position
Enstrophy	0.543 ± 0.021	$\mathbf{0.574 \pm 0.013}$	0.571 ± 0.019	0.565 ± 0.028	0.589
Buoyancy gradient energy	0.420 ± 0.023	$\mathbf{0.604 \pm 0.036}^{\dagger}$	0.458 ± 0.015	0.451 ± 0.005	-0.254
Convective flux	0.323 ± 0.008	0.379 ± 0.052	0.410 ± 0.009	$\mathbf{0.445 \pm 0.015}^{\dagger}$	-1.194
Pressure gradient magnitude	0.387 ± 0.003	0.411 ± 0.004	0.407 ± 0.002	$\mathbf{0.411 \pm 0.009}$	0.298
Buoyancy Laplacian magnitude	0.618 ± 0.016	$\mathbf{0.768 \pm 0.019}^{\dagger}$	0.627 ± 0.016	0.642 ± 0.003	0.421

Table 20: Rayleigh–Bénard token persistence skill S at target-start offset 40 frames. Skill is one minus model MSE divided by persistence MSE. Values are test mean ± standard deviation across three encoder seeds; the control is seed-independent. Bold marks the best objective and † a lead beyond the larger of the best two seed standard deviations. The persistence reference has skill zero; higher values indicate better prediction of future change.

Target	LC	LF	PC	PF	Reg./time + position
Enstrophy	0.447 ± 0.013	$\mathbf{0.465 \pm 0.007}$	0.461 ± 0.006	0.456 ± 0.005	0.487
Buoyancy gradient energy	0.542 ± 0.021	$\mathbf{0.624 \pm 0.003}^{\dagger}$	0.569 ± 0.023	0.591 ± 0.005	0.141
Convective flux	0.224 ± 0.007	0.259 ± 0.028	0.275 ± 0.003	$\mathbf{0.292 \pm 0.020}$	-0.835
Pressure gradient magnitude	0.376 ± 0.007	$\mathbf{0.395 \pm 0.005}^{\dagger}$	0.385 ± 0.007	0.386 ± 0.005	0.324
Buoyancy Laplacian magnitude	0.803 ± 0.005	$\mathbf{0.865 \pm 0.012}^{\dagger}$	0.835 ± 0.005	0.834 ± 0.006	0.772

Table 21: Shear flow token persistence skill S at target-start offset 8 frames. Skill is one minus model MSE divided by persistence MSE. Values are test mean ± standard deviation across three encoder seeds; the control is seed-independent. Bold marks the best objective and † a lead beyond the larger of the best two seed standard deviations. The persistence reference has skill zero; higher values indicate better prediction of future change.

Target	LC	LF	PC	PF	Reg./time + position
Cross-stream kinetic energy	0.455 ± 0.029	0.628 ± 0.048	0.634 ± 0.065	$\mathbf{0.645 \pm 0.002}$	-5.299
Enstrophy	0.442 ± 0.021	0.465 ± 0.043	0.536 ± 0.004	$\mathbf{0.611 \pm 0.030}^{\dagger}$	-2.695
Tracer variance	0.458 ± 0.044	0.583 ± 0.018	0.670 ± 0.008	$\mathbf{0.677 \pm 0.014}$	-12.074
Tracer gradient energy	0.349 ± 0.034	$\mathbf{0.563 \pm 0.021}^{\dagger}$	0.386 ± 0.014	0.416 ± 0.048	-0.634

Table 22: Shear flow token persistence skill S at target-start offset 16 frames. Skill is one minus model MSE divided by persistence MSE. Values are test mean ± standard deviation across three encoder seeds; the control is seed-independent. Bold marks the best objective and † a lead beyond the larger of the best two seed standard deviations. The persistence reference has skill zero; higher values indicate better prediction of future change.

Target	LC	LF	PC	PF	Reg./time + position
Cross-stream kinetic energy	0.602 ± 0.028	0.649 ± 0.010	0.666 ± 0.008	$\mathbf{0.684 \pm 0.029}$	-1.743
Enstrophy	0.399 ± 0.009	0.475 ± 0.018	0.507 ± 0.056	$\mathbf{0.522 \pm 0.035}$	-1.261
Tracer variance	0.533 ± 0.054	0.652 ± 0.008	$\mathbf{0.687 \pm 0.009}$	0.687 ± 0.022	-6.840
Tracer gradient energy	0.363 ± 0.042	$\mathbf{0.504 \pm 0.022}^{\dagger}$	0.472 ± 0.018	0.446 ± 0.008	-0.128

Table 23: Shear flow token persistence skill S at target-start offset 40 frames. Skill is one minus model MSE divided by persistence MSE. Values are test mean ± standard deviation across three encoder seeds; the control is seed-independent. Bold marks the best objective and † a lead beyond the larger of the best two seed standard deviations. The persistence reference has skill zero; higher values indicate better prediction of future change.

Target	LC	LF	PC	PF	Reg./time + position
Cross-stream kinetic energy	0.626 ± 0.039	0.686 ± 0.017	0.697 ± 0.014	$\mathbf{0.712 \pm 0.002}^{\dagger}$	-1.092
Enstrophy	0.351 ± 0.014	0.375 ± 0.026	0.395 ± 0.030	$\mathbf{0.505 \pm 0.023}^{\dagger}$	-0.580
Tracer variance	0.693 ± 0.022	0.774 ± 0.011	0.769 ± 0.006	$\mathbf{0.784 \pm 0.012}$	-2.374
Tracer gradient energy	0.310 ± 0.039	$\mathbf{0.480 \pm 0.017}^{\dagger}$	0.375 ± 0.019	0.377 ± 0.027	0.191

Table 24: Active matter pooled physical-target probes at target-start offset 0 frames. Each cell gives test VRMSE / R^2 (mean ± standard deviation across three encoder seeds); controls are seed-independent. Lower VRMSE is better; bold identifies the best objective, and † denotes a VRMSE lead exceeding the larger seed standard deviation of the two best objectives. Pixel-Current has the lowest VRMSE on all four targets.

Target	LC	LF	PC	PF	Reg./time	Persist.
	0.075 ± 0.007	0.076 ± 0.006	$\mathbf{0.059 \pm 0.006}^{\dagger}$	0.078 ± 0.011	0.506	
Kinetic energy	0.994 ± 0.001	0.994 ± 0.001	0.996 ± 0.001	0.994 ± 0.002	0.744	–
	0.040 ± 0.002	0.039 ± 0.009	$\mathbf{0.033 \pm 0.003}$	0.043 ± 0.002	0.073	
Enstrophy	0.998 ± 0.000	0.998 ± 0.001	0.999 ± 0.000	0.998 ± 0.000	0.995	–
	0.058 ± 0.006	0.053 ± 0.010	$\mathbf{0.027 \pm 0.006}^{\dagger}$	0.052 ± 0.019	0.286	
Nematic order	0.997 ± 0.001	0.997 ± 0.001	0.999 ± 0.000	0.997 ± 0.002	0.918	–
	0.072 ± 0.004	0.057 ± 0.009	$\mathbf{0.052 \pm 0.004}$	0.073 ± 0.015	0.275	
Nematic gradient energy	0.995 ± 0.001	0.997 ± 0.001	0.997 ± 0.000	0.995 ± 0.002	0.924	–

Table 25: Active matter pooled physical-target probes at target-start offset 8 frames. Each cell gives test VRMSE / R^2 (mean ± standard deviation across three encoder seeds); controls are seed-independent. Lower VRMSE is better; bold identifies the best objective, and † denotes a VRMSE lead exceeding the larger seed standard deviation of the two best objectives. Latent-Future has the lowest VRMSE on three of the four targets and Pixel-Future on one.

Target	LC	LF	PC	PF	Reg./time	Persist.
	0.330 ± 0.032	0.272 ± 0.015	0.315 ± 0.015	$\mathbf{0.259 \pm 0.047}$	0.533	0.665
Kinetic energy	0.890 ± 0.022	0.926 ± 0.008	0.901 ± 0.010	0.931 ± 0.024	0.716	0.558
	0.100 ± 0.019	$\mathbf{0.055 \pm 0.011}$	0.070 ± 0.008	0.056 ± 0.009	0.047	0.332
Enstrophy	0.990 ± 0.004	0.997 ± 0.001	0.995 ± 0.001	0.997 ± 0.001	0.998	0.890
	0.187 ± 0.043	$\mathbf{0.087 \pm 0.010}$	0.121 ± 0.017	0.100 ± 0.020	0.344	1.089
Nematic order	0.964 ± 0.015	0.992 ± 0.002	0.985 ± 0.004	0.990 ± 0.004	0.882	-0.185
	0.158 ± 0.012	$\mathbf{0.156 \pm 0.006}$	0.157 ± 0.001	0.175 ± 0.017	0.228	0.466
Nematic gradient energy	0.975 ± 0.004	0.976 ± 0.002	0.975 ± 0.000	0.969 ± 0.006	0.948	0.783

Table 26: Active matter pooled physical-target probes at target-start offset 16 frames. Each cell gives test VRMSE / R^2 (mean $\pm$ standard deviation across three encoder seeds); controls are seed-independent. Lower VRMSE is better; bold identifies the best objective, and † denotes a VRMSE lead exceeding the larger seed standard deviation of the two best objectives. Latent-Future has the lowest VRMSE on three of the four targets and Pixel-Current on one.

Target	LC	LF	PC	PF	Reg./time	Persist.
	0.638 ± 0.026	$\mathbf{0.403 \pm 0.006}^{\dagger}$	0.612 ± 0.036	0.432 ± 0.019	0.514	0.692
Kinetic energy	0.593 ± 0.033	0.838 ± 0.005	0.625 ± 0.044	0.813 ± 0.016	0.736	0.521
	0.180 ± 0.011	$\mathbf{0.072 \pm 0.013}$	0.124 ± 0.018	0.082 ± 0.016	0.038	0.353
Enstrophy	0.967 ± 0.004	0.995 ± 0.002	0.984 ± 0.004	0.993 ± 0.003	0.999	0.876
	0.644 ± 0.055	$\mathbf{0.141 \pm 0.006}^{\dagger}$	0.392 ± 0.054	0.241 ± 0.024	0.752	1.676
Nematic order	0.583 ± 0.069	0.980 ± 0.002	0.844 ± 0.044	0.941 ± 0.012	0.434	-1.810
	0.220 ± 0.015	0.197 ± 0.016	$\mathbf{0.175 \pm 0.011}^{\dagger}$	0.208 ± 0.024	0.186	0.518
Nematic gradient energy	0.951 ± 0.006	0.961 ± 0.006	0.969 ± 0.004	0.956 ± 0.010	0.965	0.732

Table 27: Active matter pooled physical-target probes at target-start offset 40 frames. Each cell gives test VRMSE / R^2 (mean $\pm$ standard deviation across three encoder seeds); controls are seed-independent. Lower VRMSE is better; bold identifies the best objective, and † denotes a VRMSE lead exceeding the larger seed standard deviation of the two best objectives. Latent-Future has the lowest VRMSE on three of the four targets and Pixel-Current on one.

Target	LC	LF	PC	PF	Reg./time	Persist.
	0.806 ± 0.053	$\mathbf{0.567 \pm 0.027}^{\dagger}$	0.762 ± 0.030	0.743 ± 0.144	0.453	0.784
Kinetic energy	0.348 ± 0.087	0.678 ± 0.031	0.418 ± 0.045	0.434 ± 0.203	0.795	0.385
	0.224 ± 0.008	$\mathbf{0.090 \pm 0.017}$	0.188 ± 0.008	0.107 ± 0.002	0.044	0.358
Enstrophy	0.950 ± 0.004	0.992 ± 0.003	0.965 ± 0.003	0.989 ± 0.000	0.998	0.872
	0.575 ± 0.023	$\mathbf{0.544 \pm 0.036}$	0.549 ± 0.029	0.573 ± 0.014	0.502	7.613
Nematic order	0.669 ± 0.026	0.704 ± 0.038	0.698 ± 0.032	0.671 ± 0.016	0.748	-56.957
	0.286 ± 0.024	0.288 ± 0.018	$\mathbf{0.245 \pm 0.016}^{\dagger}$	0.289 ± 0.040	0.170	0.635
Nematic gradient energy	0.918 ± 0.014	0.917 ± 0.010	0.940 ± 0.008	0.915 ± 0.023	0.971	0.596

Table 28: Rayleigh–Bénard pooled physical-target probes at target-start offset 0 frames. Each cell gives test VRMSE / R^2 (mean $\pm$ standard deviation across three encoder seeds); controls are seed-independent. Lower VRMSE is better; bold identifies the best objective, and † denotes a VRMSE lead exceeding the larger seed standard deviation of the two best objectives. Latent-Future has the lowest VRMSE on three of the five targets, Pixel-Current on one and Pixel-Future on one.

Target	LC	LF	PC	PF	Reg./time	Persist.
	0.097 ± 0.007	$\mathbf{0.056 \pm 0.006}$	0.063 ± 0.008	0.056 ± 0.001	0.222	
Enstrophy	0.991 ± 0.001	0.997 ± 0.001	0.996 ± 0.001	0.997 ± 0.000	0.951	–
	0.114 ± 0.018	$\mathbf{0.113 \pm 0.011}$	0.139 ± 0.020	0.155 ± 0.021	0.278	
Buoyancy gradient energy	0.987 ± 0.004	0.987 ± 0.003	0.980 ± 0.005	0.976 ± 0.007	0.923	–
	0.133 ± 0.012	0.117 ± 0.016	0.112 ± 0.002	$\mathbf{0.100 \pm 0.009}^{\dagger}$	0.480	
Convective flux	0.982 ± 0.003	0.986 ± 0.004	0.987 ± 0.001	0.990 ± 0.002	0.770	–
	0.048 ± 0.001	0.033 ± 0.003	$\mathbf{0.029 \pm 0.002}$	0.029 ± 0.000	0.415	
Pressure gradient magnitude	0.998 ± 0.000	0.999 ± 0.000	0.999 ± 0.000	0.999 ± 0.000	0.828	–
	0.096 ± 0.008	$\mathbf{0.089 \pm 0.002}$	0.098 ± 0.005	0.113 ± 0.009	0.227	
Buoyancy Laplacian magnitude	0.991 ± 0.002	0.992 ± 0.000	0.990 ± 0.001	0.987 ± 0.002	0.949	–

Table 29: Rayleigh–Bénard pooled physical-target probes at target-start offset 8 frames. Each cell gives test VRMSE / R^2 (mean ± standard deviation across three encoder seeds); controls are seed-independent. Lower VRMSE is better; bold identifies the best objective, and † denotes a VRMSE lead exceeding the larger seed standard deviation of the two best objectives. Pixel-Future has the lowest VRMSE on three of the five targets and Latent-Future on two.

Target	LC	LF	PC	PF	Reg./time	Persist.
	0.153 ± 0.008	0.122 ± 0.006	0.121 ± 0.004	$\mathbf{0.098 \pm 0.006}^{\dagger}$	0.216	0.138
Enstrophy	0.976 ± 0.002	0.985 ± 0.002	0.985 ± 0.001	0.990 ± 0.001	0.953	0.981
	0.139 ± 0.004	$\mathbf{0.111 \pm 0.008}^{\dagger}$	0.155 ± 0.007	0.156 ± 0.004	0.293	0.421
Buoyancy gradient energy	0.981 ± 0.001	0.988 ± 0.002	0.976 ± 0.002	0.976 ± 0.001	0.914	0.822
	0.188 ± 0.005	0.165 ± 0.012	0.168 ± 0.006	$\mathbf{0.129 \pm 0.005}^{\dagger}$	0.313	0.591
Convective flux	0.965 ± 0.002	0.973 ± 0.004	0.972 ± 0.002	0.983 ± 0.001	0.902	0.651
	0.063 ± 0.003	0.048 ± 0.003	0.047 ± 0.001	$\mathbf{0.045 \pm 0.001}^{\dagger}$	0.418	0.117
Pressure gradient magnitude	0.996 ± 0.000	0.998 ± 0.000	0.998 ± 0.000	0.998 ± 0.000	0.825	0.986
	0.118 ± 0.005	$\mathbf{0.092 \pm 0.007}^{\dagger}$	0.124 ± 0.009	0.127 ± 0.004	0.188	0.328
Buoyancy Laplacian magnitude	0.986 ± 0.001	0.992 ± 0.001	0.985 ± 0.002	0.984 ± 0.001	0.965	0.892

Table 30: Rayleigh–Bénard pooled physical-target probes at target-start offset 16 frames. Each cell gives test VRMSE / R^2 (mean ± standard deviation across three encoder seeds); controls are seed-independent. Lower VRMSE is better; bold identifies the best objective, and † denotes a VRMSE lead exceeding the larger seed standard deviation of the two best objectives. Latent-Future has the lowest VRMSE on four of the five targets and Pixel-Future on one.

Target	LC	LF	PC	PF	Reg./time	Persist.
	0.174 ± 0.011	$\mathbf{0.126 \pm 0.006}$	0.132 ± 0.001	0.134 ± 0.002	0.182	0.193
Enstrophy	0.970 ± 0.004	0.984 ± 0.002	0.983 ± 0.000	0.982 ± 0.000	0.967	0.963
	0.168 ± 0.014	$\mathbf{0.087 \pm 0.004}^{\dagger}$	0.170 ± 0.001	0.176 ± 0.005	0.419	0.775
Buoyancy gradient energy	0.972 ± 0.005	0.992 ± 0.001	0.971 ± 0.000	0.969 ± 0.002	0.825	0.400
	0.219 ± 0.004	0.210 ± 0.003	0.214 ± 0.001	$\mathbf{0.199 \pm 0.006}^{\dagger}$	0.260	0.598
Convective flux	0.952 ± 0.002	0.956 ± 0.001	0.954 ± 0.000	0.960 ± 0.002	0.932	0.642
	0.083 ± 0.003	$\mathbf{0.067 \pm 0.004}$	0.070 ± 0.001	0.068 ± 0.002	0.403	0.170
Pressure gradient magnitude	0.993 ± 0.000	0.995 ± 0.001	0.995 ± 0.000	0.995 ± 0.000	0.837	0.971
	0.156 ± 0.006	$\mathbf{0.084 \pm 0.005}^{\dagger}$	0.145 ± 0.003	0.150 ± 0.001	0.239	0.701
Buoyancy Laplacian magnitude	0.976 ± 0.002	0.993 ± 0.001	0.979 ± 0.001	0.977 ± 0.000	0.943	0.509

Table 31: Rayleigh–Bénard pooled physical-target probes at target-start offset 40 frames. Each cell gives test VRMSE / R^2 (mean ± standard deviation across three encoder seeds); controls are seed-independent. Lower VRMSE is better; bold identifies the best objective, and † denotes a VRMSE lead exceeding the larger seed standard deviation of the two best objectives. Latent-Future has the lowest VRMSE on all five targets.

Target	LC	LF	PC	PF	Reg./time	Persist.
	0.239 ± 0.012	$\mathbf{0.141 \pm 0.011}^{\dagger}$	0.224 ± 0.002	0.226 ± 0.003	0.213	0.355
Enstrophy	0.943 ± 0.006	0.980 ± 0.003	0.950 ± 0.001	0.949 ± 0.001	0.955	0.874
	0.475 ± 0.129	$\mathbf{0.271 \pm 0.001}^{\dagger}$	0.609 ± 0.000	0.609 ± 0.001	0.511	1.248
Buoyancy gradient energy	0.764 ± 0.125	0.927 ± 0.000	0.630 ± 0.000	0.630 ± 0.001	0.739	-0.558
	0.337 ± 0.041	$\mathbf{0.203 \pm 0.004}^{\dagger}$	0.375 ± 0.001	0.374 ± 0.002	0.411	0.779
Convective flux	0.885 ± 0.027	0.959 ± 0.002	0.860 ± 0.001	0.860 ± 0.001	0.831	0.394
	0.116 ± 0.004	$\mathbf{0.081 \pm 0.004}^{\dagger}$	0.103 ± 0.001	0.106 ± 0.001	0.354	0.326
Pressure gradient magnitude	0.986 ± 0.001	0.993 ± 0.001	0.989 ± 0.000	0.989 ± 0.000	0.875	0.894
	0.515 ± 0.144	$\mathbf{0.229 \pm 0.008}^{\dagger}$	0.664 ± 0.000	0.664 ± 0.000	0.481	1.353
Buoyancy Laplacian magnitude	0.721 ± 0.151	0.948 ± 0.004	0.559 ± 0.000	0.559 ± 0.001	0.769	-0.829

Table 32: Shear flow pooled physical-target probes at target-start offset 0 frames. Each cell gives test VRMSE / R^2 (mean ± standard deviation across three encoder seeds); controls are seed-independent. Lower VRMSE is better; bold identifies the best objective, and † denotes a VRMSE lead exceeding the larger seed standard deviation of the two best objectives. Pixel-Current has the lowest VRMSE on three of the four targets and Pixel-Future on one.

Target	LC	LF	PC	PF	Reg./time	Persist.
	0.009 ± 0.002	0.009 ± 0.001	$\mathbf{0.005 \pm 0.000}^{\dagger}$	0.006 ± 0.000	0.940	
Cross-stream kinetic energy	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.116	–
	0.027 ± 0.003	0.025 ± 0.002	0.023 ± 0.003	$\mathbf{0.018 \pm 0.001}^{\dagger}$	0.982	
Enstrophy	0.999 ± 0.000	0.999 ± 0.000	0.999 ± 0.000	1.000 ± 0.000	0.036	–
	0.019 ± 0.002	0.019 ± 0.001	$\mathbf{0.011 \pm 0.001}^{\dagger}$	0.013 ± 0.001	0.881	
Tracer variance	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.223	–
	0.160 ± 0.011	0.090 ± 0.010	$\mathbf{0.048 \pm 0.011}^{\dagger}$	0.071 ± 0.011	0.930	
Tracer gradient energy	0.974 ± 0.004	0.992 ± 0.002	0.998 ± 0.001	0.995 ± 0.002	0.136	–

Table 33: Shear flow pooled physical-target probes at target-start offset 8 frames. Each cell gives test VRMSE / R^2 (mean ± standard deviation across three encoder seeds); controls are seed-independent. Lower VRMSE is better; bold identifies the best objective, and † denotes a VRMSE lead exceeding the larger seed standard deviation of the two best objectives. Pixel-Current has the lowest VRMSE on two of the four targets, Latent-Future on one and Pixel-Future on one.

Target	LC	LF	PC	PF	Reg./time	Persist.
	0.025 ± 0.002	0.018 ± 0.002	0.021 ± 0.000	$\mathbf{0.014 \pm 0.001}^{\dagger}$	0.985	0.388
Cross-stream kinetic energy	0.999 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.030	0.850
	0.043 ± 0.004	$\mathbf{0.032 \pm 0.004}$	0.040 ± 0.000	0.034 ± 0.002	0.988	0.204
Enstrophy	0.998 ± 0.000	0.999 ± 0.000	0.998 ± 0.000	0.999 ± 0.000	0.024	0.958
	0.024 ± 0.002	0.019 ± 0.001	$\mathbf{0.016 \pm 0.001}^{\dagger}$	0.017 ± 0.001	0.880	0.069
Tracer variance	0.999 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000	0.226	0.995
	0.147 ± 0.017	0.098 ± 0.024	$\mathbf{0.076 \pm 0.004}$	0.082 ± 0.011	0.929	0.256
Tracer gradient energy	0.978 ± 0.005	0.990 ± 0.005	0.994 ± 0.001	0.993 ± 0.002	0.138	0.935

Table 34: Shear flow pooled physical-target probes at target-start offset 16 frames. Each cell gives test VRMSE / R^2 (mean ± standard deviation across three encoder seeds); controls are seed-independent. Lower VRMSE is better; bold identifies the best objective, and † denotes a VRMSE lead exceeding the larger seed standard deviation of the two best objectives. Latent-Future has the lowest VRMSE on three of the four targets and Pixel-Future on one.

Target	LC	LF	PC	PF	Reg./time	Persist.
	0.031 ± 0.003	0.025 ± 0.001	0.030 ± 0.003	$\mathbf{0.023 \pm 0.003}$	1.007	0.602
Cross-stream kinetic energy	0.999 ± 0.000	0.999 ± 0.000	0.999 ± 0.000	0.999 ± 0.000	-0.014	0.637
	0.058 ± 0.001	$\mathbf{0.040 \pm 0.005}$	0.055 ± 0.001	0.044 ± 0.002	0.993	0.378
Enstrophy	0.997 ± 0.000	0.998 ± 0.000	0.997 ± 0.000	0.998 ± 0.000	0.013	0.857
	0.041 ± 0.005	$\mathbf{0.023 \pm 0.001}^{\dagger}$	0.038 ± 0.001	0.028 ± 0.001	0.884	0.178
Tracer variance	0.998 ± 0.000	0.999 ± 0.000	0.999 ± 0.000	0.999 ± 0.000	0.218	0.968
	0.235 ± 0.021	$\mathbf{0.116 \pm 0.015}^{\dagger}$	0.190 ± 0.011	0.166 ± 0.005	0.952	0.582
Tracer gradient energy	0.945 ± 0.009	0.986 ± 0.003	0.964 ± 0.004	0.972 ± 0.002	0.094	0.661

Table 35: Shear flow pooled physical-target probes at target-start offset 40 frames. Each cell gives test VRMSE / R^2 (mean ± standard deviation across three encoder seeds); controls are seed-independent. Lower VRMSE is better; bold identifies the best objective, and † denotes a VRMSE lead exceeding the larger seed standard deviation of the two best objectives. Latent-Future has the lowest VRMSE on all four targets.

Target	LC	LF	PC	PF	Reg./time	Persist.
	0.049 ± 0.005	$\mathbf{0.033 \pm 0.004}^{\dagger}$	0.050 ± 0.003	0.038 ± 0.003	1.010	0.751
Cross-stream kinetic energy	0.998 ± 0.001	0.999 ± 0.000	0.998 ± 0.000	0.999 ± 0.000	-0.019	0.436
	0.072 ± 0.002	$\mathbf{0.044 \pm 0.003}^{\dagger}$	0.065 ± 0.002	0.050 ± 0.003	1.001	0.650
Enstrophy	0.995 ± 0.000	0.998 ± 0.000	0.996 ± 0.000	0.998 ± 0.000	-0.003	0.577
	0.078 ± 0.009	$\mathbf{0.036 \pm 0.003}^{\dagger}$	0.073 ± 0.006	0.057 ± 0.005	0.902	0.483
Tracer variance	0.994 ± 0.001	0.999 ± 0.000	0.995 ± 0.001	0.997 ± 0.001	0.186	0.766
	0.455 ± 0.037	$\mathbf{0.214 \pm 0.028}^{\dagger}$	0.406 ± 0.022	0.341 ± 0.033	0.973	0.908
Tracer gradient energy	0.792 ± 0.034	0.954 ± 0.012	0.835 ± 0.018	0.883 ± 0.023	0.053	0.176

Table 36: Active matter token physical-target probes at target-start offset 0 frames. Each cell gives test VRMSE / R^2 (mean $\pm$ standard deviation across three encoder seeds); controls are seed-independent. Lower VRMSE is better; bold identifies the best objective, and † denotes a VRMSE lead exceeding the larger seed standard deviation of the two best objectives. Token-level scores depend on the physical target and system.

Target	LC	LF	PC	PF	Reg./time + position	Persist.
	$\mathbf{0.098 \pm 0.003}^{\dagger}$	0.111 ± 0.003	0.124 ± 0.006	0.122 ± 0.009	0.890	
Kinetic energy	0.990 ± 0.001	0.988 ± 0.001	0.985 ± 0.001	0.985 ± 0.002	0.208	–
	$\mathbf{0.117 \pm 0.003}$	0.124 ± 0.010	0.150 ± 0.003	0.143 ± 0.012	0.786	
Enstrophy	0.986 ± 0.001	0.985 ± 0.002	0.977 ± 0.001	0.979 ± 0.003	0.382	–
	0.117 ± 0.006	0.120 ± 0.008	$\mathbf{0.108 \pm 0.004}^{\dagger}$	0.125 ± 0.007	0.321	
Nematic order	0.986 ± 0.001	0.986 ± 0.002	0.988 ± 0.001	0.984 ± 0.002	0.897	–
	$\mathbf{0.165 \pm 0.004}^{\dagger}$	0.184 ± 0.005	0.196 ± 0.005	0.182 ± 0.005	0.807	
Nematic gradient energy	0.973 ± 0.001	0.966 ± 0.002	0.962 ± 0.002	0.967 ± 0.002	0.348	–

Table 37: Active matter token physical-target probes at target-start offset 8 frames. Each cell gives test VRMSE / R^2 (mean $\pm$ standard deviation across three encoder seeds); controls are seed-independent. Lower VRMSE is better; bold identifies the best objective, and † denotes a VRMSE lead exceeding the larger seed standard deviation of the two best objectives. Token-level scores depend on the physical target and system.

Target	LC	LF	PC	PF	Reg./time + position	Persist.
	0.895 ± 0.025	$\mathbf{0.872 \pm 0.005}$	0.917 ± 0.008	0.876 ± 0.041	0.891	1.099
Kinetic energy	0.199 ± 0.045	0.240 ± 0.010	0.160 ± 0.014	0.232 ± 0.072	0.206	-0.207
	0.843 ± 0.007	0.848 ± 0.004	0.845 ± 0.003	$\mathbf{0.842 \pm 0.005}$	0.829	1.061
Enstrophy	0.290 ± 0.011	0.281 ± 0.006	0.286 ± 0.005	0.291 ± 0.008	0.313	-0.125
	0.510 ± 0.061	$\mathbf{0.289 \pm 0.013}$	0.432 ± 0.029	0.308 ± 0.019	0.369	1.020
Nematic order	0.737 ± 0.061	0.916 ± 0.008	0.813 ± 0.025	0.905 ± 0.012	0.864	-0.040
	0.853 ± 0.011	$\mathbf{0.846 \pm 0.001}$	0.853 ± 0.008	0.848 ± 0.009	0.835	1.065
Nematic gradient energy	0.272 ± 0.019	0.284 ± 0.001	0.272 ± 0.014	0.280 ± 0.015	0.303	-0.135

Table 38: Active matter token physical-target probes at target-start offset 16 frames. Each cell gives test VRMSE / R^2 (mean $\pm$ standard deviation across three encoder seeds); controls are seed-independent. Lower VRMSE is better; bold identifies the best objective, and † denotes a VRMSE lead exceeding the larger seed standard deviation of the two best objectives. Token-level scores depend on the physical target and system.

Target	LC	LF	PC	PF	Reg./time + position	Persist.
	0.925 ± 0.016	$\mathbf{0.890 \pm 0.013}$	0.952 ± 0.010	0.910 ± 0.035	0.863	1.160
Kinetic energy	0.144 ± 0.030	0.208 ± 0.023	0.094 ± 0.019	0.172 ± 0.064	0.256	-0.346
	0.856 ± 0.005	$\mathbf{0.853 \pm 0.003}$	0.868 ± 0.014	0.855 ± 0.003	0.837	1.162
Enstrophy	0.267 ± 0.008	0.272 ± 0.005	0.246 ± 0.024	0.268 ± 0.005	0.300	-0.351
	0.907 ± 0.034	$\mathbf{0.400 \pm 0.011}^{\dagger}$	0.877 ± 0.003	0.777 ± 0.019	0.902	1.417
Nematic order	0.176 ± 0.061	0.840 ± 0.009	0.231 ± 0.005	0.396 ± 0.029	0.186	-1.007
	0.882 ± 0.008	0.879 ± 0.005	$\mathbf{0.871 \pm 0.008}^{\dagger}$	0.881 ± 0.007	0.846	1.146
Nematic gradient energy	0.222 ± 0.015	0.227 ± 0.009	0.241 ± 0.013	0.224 ± 0.012	0.285	-0.314

Table 39: Active matter token physical-target probes at target-start offset 40 frames. Each cell gives test VRMSE / R^2 (mean $\pm$ standard deviation across three encoder seeds); controls are seed-independent. Lower VRMSE is better; bold identifies the best objective, and † denotes a VRMSE lead exceeding the larger seed standard deviation of the two best objectives. Token-level scores depend on the physical target and system.

Target	LC	LF	PC	PF	Reg./time + position	Persist.
	0.902 ± 0.007	$\mathbf{0.879 \pm 0.013}^{\dagger}$	0.931 ± 0.004	0.921 ± 0.006	0.815	1.021
Kinetic energy	0.186 ± 0.012	0.227 ± 0.024	0.134 ± 0.007	0.151 ± 0.011	0.336	-0.043
	$\mathbf{0.851 \pm 0.001}$	0.854 ± 0.011	0.859 ± 0.003	0.857 ± 0.008	0.832	1.100
Enstrophy	0.275 ± 0.001	0.271 ± 0.018	0.262 ± 0.006	0.266 ± 0.013	0.308	-0.209
	0.929 ± 0.015	$\mathbf{0.782 \pm 0.009}^{\dagger}$	0.914 ± 0.005	0.876 ± 0.017	0.878	4.017
Nematic order	0.137 ± 0.028	0.389 ± 0.014	0.164 ± 0.008	0.232 ± 0.030	0.229	-15.138
	$\mathbf{0.856 \pm 0.013}$	0.866 ± 0.001	0.861 ± 0.008	0.869 ± 0.006	0.816	1.148
Nematic gradient energy	0.268 ± 0.022	0.249 ± 0.002	0.259 ± 0.013	0.245 ± 0.011	0.335	-0.317

Table 40: Rayleigh–Bénard token physical-target probes at target-start offset 0 frames. Each cell gives test VRMSE / R^2 (mean $\pm$ standard deviation across three encoder seeds); controls are seed-independent. Lower VRMSE is better; bold identifies the best objective, and † denotes a VRMSE lead exceeding the larger seed standard deviation of the two best objectives. Token-level scores depend on the physical target and system.

Target	LC	LF	PC	PF	Reg./time + position	Persist.
	0.241 ± 0.019	0.251 ± 0.028	$\mathbf{0.231 \pm 0.006}$	0.261 ± 0.009	0.754	
Enstrophy	0.942 ± 0.009	0.936 ± 0.014	0.947 ± 0.003	0.932 ± 0.005	0.431	–
	0.178 ± 0.009	0.169 ± 0.014	0.169 ± 0.006	$\mathbf{0.151 \pm 0.009}^{\dagger}$	0.512	
Buoyancy gradient energy	0.968 ± 0.003	0.971 ± 0.005	0.971 ± 0.002	0.977 ± 0.003	0.738	–
	0.050 ± 0.001	0.045 ± 0.004	$\mathbf{0.045 \pm 0.000}$	0.046 ± 0.001	0.998	
Convective flux	0.998 ± 0.000	0.998 ± 0.000	0.998 ± 0.000	0.998 ± 0.000	0.003	–
	0.188 ± 0.002	0.204 ± 0.007	$\mathbf{0.166 \pm 0.006}^{\dagger}$	0.172 ± 0.003	0.678	
Pressure gradient magnitude	0.965 ± 0.001	0.958 ± 0.003	0.972 ± 0.002	0.970 ± 0.001	0.541	–
	0.317 ± 0.034	$\mathbf{0.254 \pm 0.015}^{\dagger}$	0.280 ± 0.014	0.282 ± 0.021	0.408	
Buoyancy Laplacian magnitude	0.898 ± 0.022	0.935 ± 0.008	0.921 ± 0.008	0.920 ± 0.012	0.834	–

Table 41: Rayleigh–Bénard token physical-target probes at target-start offset 8 frames. Each cell gives test VRMSE / R^2 (mean $\pm$ standard deviation across three encoder seeds); controls are seed-independent. Lower VRMSE is better; bold identifies the best objective, and † denotes a VRMSE lead exceeding the larger seed standard deviation of the two best objectives. Token-level scores depend on the physical target and system.

Target	LC	LF	PC	PF	Reg./time + position	Persist.
	0.713 ± 0.018	0.688 ± 0.005	0.692 ± 0.006	$\mathbf{0.667 \pm 0.004}^{\dagger}$	0.749	0.990
Enstrophy	0.492 ± 0.025	0.527 ± 0.007	0.521 ± 0.008	0.555 ± 0.005	0.438	0.019
	0.384 ± 0.014	$\mathbf{0.316 \pm 0.004}^{\dagger}$	0.375 ± 0.006	0.373 ± 0.002	0.558	0.489
Buoyancy gradient energy	0.852 ± 0.010	0.900 ± 0.003	0.859 ± 0.004	0.861 ± 0.001	0.689	0.761
	0.455 ± 0.003	0.419 ± 0.033	0.400 ± 0.003	$\mathbf{0.364 \pm 0.006}^{\dagger}$	0.998	0.582
Convective flux	0.793 ± 0.003	0.824 ± 0.027	0.840 ± 0.002	0.868 ± 0.004	0.004	0.661
	0.627 ± 0.003	0.609 ± 0.001	0.607 ± 0.002	$\mathbf{0.604 \pm 0.001}^{\dagger}$	0.697	0.825
Pressure gradient magnitude	0.607 ± 0.003	0.629 ± 0.001	0.632 ± 0.003	0.635 ± 0.001	0.514	0.320
	0.394 ± 0.014	$\mathbf{0.314 \pm 0.002}^{\dagger}$	0.376 ± 0.006	0.373 ± 0.007	0.490	0.480
Buoyancy Laplacian magnitude	0.844 ± 0.011	0.901 ± 0.001	0.859 ± 0.004	0.861 ± 0.005	0.760	0.770

Table 42: Rayleigh–Bénard token physical-target probes at target-start offset 16 frames. Each cell gives test VRMSE / R^2 (mean $\pm$ standard deviation across three encoder seeds); controls are seed-independent. Lower VRMSE is better; bold identifies the best objective, and † denotes a VRMSE lead exceeding the larger seed standard deviation of the two best objectives. Token-level scores depend on the physical target and system.

Target	LC	LF	PC	PF	Reg./time + position	Persist.
	0.734 ± 0.017	$\mathbf{0.708 \pm 0.011}$	0.710 ± 0.016	0.716 ± 0.023	0.696	1.085
Enstrophy	0.461 ± 0.025	0.498 ± 0.015	0.495 ± 0.022	0.487 ± 0.033	0.516	-0.178
	0.383 ± 0.008	$\mathbf{0.316 \pm 0.014}^{\dagger}$	0.370 ± 0.005	0.372 ± 0.002	0.563	0.502
Buoyancy gradient energy	0.854 ± 0.006	0.900 ± 0.009	0.863 ± 0.004	0.861 ± 0.001	0.683	0.748
	0.554 ± 0.003	0.530 ± 0.022	0.517 ± 0.004	$\mathbf{0.502 \pm 0.007}^{\dagger}$	0.998	0.673
Convective flux	0.693 ± 0.004	0.718 ± 0.023	0.732 ± 0.004	0.748 ± 0.007	0.005	0.547
	0.664 ± 0.002	0.651 ± 0.002	0.652 ± 0.001	$\mathbf{0.650 \pm 0.005}$	0.710	0.848
Pressure gradient magnitude	0.560 ± 0.002	0.577 ± 0.003	0.574 ± 0.001	0.577 ± 0.007	0.496	0.282
	0.405 ± 0.009	$\mathbf{0.315 \pm 0.013}^{\dagger}$	0.400 ± 0.009	0.392 ± 0.002	0.498	0.655
Buoyancy Laplacian magnitude	0.836 ± 0.007	0.900 ± 0.008	0.840 ± 0.007	0.846 ± 0.001	0.752	0.571

Table 43: Rayleigh–Bénard token physical-target probes at target-start offset 40 frames. Each cell gives test VRMSE / R^2 (mean $\pm$ standard deviation across three encoder seeds); controls are seed-independent. Lower VRMSE is better; bold identifies the best objective, and † denotes a VRMSE lead exceeding the larger seed standard deviation of the two best objectives. Token-level scores depend on the physical target and system.

Target	LC	LF	PC	PF	Reg./time + position	Persist.
	0.764 ± 0.009	$\mathbf{0.751 \pm 0.005}$	0.754 ± 0.004	0.758 ± 0.004	0.736	1.027
Enstrophy	0.416 ± 0.014	0.435 ± 0.008	0.431 ± 0.007	0.426 ± 0.006	0.458	-0.055
	0.353 ± 0.008	$\mathbf{0.320 \pm 0.001}^{\dagger}$	0.343 ± 0.009	0.334 ± 0.002	0.484	0.522
Buoyancy gradient energy	0.875 ± 0.006	0.897 ± 0.001	0.882 ± 0.006	0.888 ± 0.001	0.766	0.727
	0.649 ± 0.003	0.634 ± 0.012	0.628 ± 0.001	$\mathbf{0.620 \pm 0.009}$	0.998	0.737
Convective flux	0.578 ± 0.004	0.597 ± 0.015	0.606 ± 0.002	0.615 ± 0.011	0.003	0.457
	0.699 ± 0.004	$\mathbf{0.688 \pm 0.003}^{\dagger}$	0.694 ± 0.004	0.693 ± 0.003	0.727	0.885
Pressure gradient magnitude	0.511 ± 0.005	0.527 ± 0.004	0.519 ± 0.006	0.519 ± 0.004	0.471	0.217
	0.342 ± 0.004	$\mathbf{0.283 \pm 0.012}^{\dagger}$	0.313 ± 0.005	0.314 ± 0.005	0.368	0.770
Buoyancy Laplacian magnitude	0.883 ± 0.003	0.920 ± 0.007	0.902 ± 0.003	0.901 ± 0.003	0.864	0.406

Table 44: Shear flow token physical-target probes at target-start offset 0 frames. Each cell gives test VRMSE / R^2 (mean $\pm$ standard deviation across three encoder seeds); controls are seed-independent. Lower VRMSE is better; bold identifies the best objective, and † denotes a VRMSE lead exceeding the larger seed standard deviation of the two best objectives. Token-level scores depend on the physical target and system.

Target	LC	LF	PC	PF	Reg./time + position	Persist.
	0.053 ± 0.008	$\mathbf{0.049 \pm 0.002}$	0.057 ± 0.004	0.055 ± 0.006	0.996	
Cross-stream kinetic energy	0.997 ± 0.001	0.998 ± 0.000	0.997 ± 0.000	0.997 ± 0.001	0.008	–
	0.121 ± 0.007	0.120 ± 0.003	$\mathbf{0.082 \pm 0.007}^{\dagger}$	0.092 ± 0.009	0.989	
Enstrophy	0.985 ± 0.002	0.985 ± 0.001	0.993 ± 0.001	0.992 ± 0.002	0.022	–
	0.066 ± 0.002	0.059 ± 0.001	$\mathbf{0.044 \pm 0.002}^{\dagger}$	0.051 ± 0.001	0.854	
Tracer variance	0.996 ± 0.000	0.997 ± 0.000	0.998 ± 0.000	0.997 ± 0.000	0.271	–
	0.617 ± 0.034	$\mathbf{0.476 \pm 0.029}$	0.481 ± 0.014	0.543 ± 0.033	1.057	
Tracer gradient energy	0.618 ± 0.041	0.773 ± 0.027	0.768 ± 0.014	0.704 ± 0.036	-0.117	–

Table 45: Shear flow token physical-target probes at target-start offset 8 frames. Each cell gives test VRMSE / R^2 (mean $\pm$ standard deviation across three encoder seeds); controls are seed-independent. Lower VRMSE is better; bold identifies the best objective, and † denotes a VRMSE lead exceeding the larger seed standard deviation of the two best objectives. Token-level scores depend on the physical target and system.

Target	LC	LF	PC	PF	Reg./time + position	Persist.
	0.294 ± 0.008	0.242 ± 0.016	0.240 ± 0.021	$\mathbf{0.237 \pm 0.001}$	0.999	0.398
Cross-stream kinetic energy	0.914 ± 0.005	0.941 ± 0.008	0.942 ± 0.010	0.944 ± 0.000	0.002	0.842
	0.385 ± 0.007	0.377 ± 0.015	0.351 ± 0.001	$\mathbf{0.321 \pm 0.012}^{\dagger}$	0.991	0.516
Enstrophy	0.852 ± 0.006	0.858 ± 0.011	0.877 ± 0.001	0.897 ± 0.008	0.018	0.734
	0.174 ± 0.007	0.153 ± 0.003	0.136 ± 0.002	$\mathbf{0.135 \pm 0.003}$	0.856	0.237
Tracer variance	0.970 ± 0.002	0.977 ± 0.001	0.982 ± 0.000	0.982 ± 0.001	0.267	0.944
	0.648 ± 0.017	$\mathbf{0.531 \pm 0.013}^{\dagger}$	0.630 ± 0.007	0.614 ± 0.025	1.027	0.804
Tracer gradient energy	0.580 ± 0.022	0.718 ± 0.014	0.604 ± 0.009	0.623 ± 0.031	-0.055	0.354

Table 46: Shear flow token physical-target probes at target-start offset 16 frames. Each cell gives test VRMSE / R^2 (mean $\pm$ standard deviation across three encoder seeds); controls are seed-independent. Lower VRMSE is better; bold identifies the best objective, and † denotes a VRMSE lead exceeding the larger seed standard deviation of the two best objectives. Token-level scores depend on the physical target and system.

Target	LC	LF	PC	PF	Reg./time + position	Persist.
	0.380 ± 0.013	0.357 ± 0.005	0.348 ± 0.004	$\mathbf{0.338 \pm 0.015}$	0.998	0.602
Cross-stream kinetic energy	0.856 ± 0.010	0.873 ± 0.003	0.879 ± 0.003	0.885 ± 0.011	0.005	0.637
	0.511 ± 0.004	0.478 ± 0.008	0.462 ± 0.026	$\mathbf{0.456 \pm 0.016}$	0.991	0.659
Enstrophy	0.739 ± 0.004	0.772 ± 0.008	0.786 ± 0.024	0.792 ± 0.015	0.017	0.565
	0.210 ± 0.012	0.181 ± 0.002	$\mathbf{0.172 \pm 0.002}$	0.172 ± 0.006	0.859	0.307
Tracer variance	0.956 ± 0.005	0.967 ± 0.001	0.971 ± 0.001	0.970 ± 0.002	0.262	0.906
	0.754 ± 0.025	$\mathbf{0.665 \pm 0.015}^{\dagger}$	0.687 ± 0.012	0.704 ± 0.005	1.004	0.945
Tracer gradient energy	0.431 ± 0.038	0.557 ± 0.020	0.528 ± 0.016	0.505 ± 0.007	-0.008	0.107

Table 47: Shear flow token physical-target probes at target-start offset 40 frames. Each cell gives test VRMSE / R^2 (mean $\pm$ standard deviation across three encoder seeds); controls are seed-independent. Lower VRMSE is better; bold identifies the best objective, and † denotes a VRMSE lead exceeding the larger seed standard deviation of the two best objectives. Token-level scores depend on the physical target and system.

Target	LC	LF	PC	PF	Reg./time + position	Persist.
	0.420 ± 0.022	0.385 ± 0.010	0.378 ± 0.009	$\mathbf{0.369 \pm 0.001}^{\dagger}$	0.994	0.687
Cross-stream kinetic energy	0.824 ± 0.018	0.852 ± 0.008	0.857 ± 0.006	0.864 ± 0.001	0.013	0.528
	0.638 ± 0.007	0.625 ± 0.013	0.615 ± 0.015	$\mathbf{0.557 \pm 0.013}^{\dagger}$	0.995	0.791
Enstrophy	0.593 ± 0.009	0.609 ± 0.017	0.621 ± 0.019	0.690 ± 0.014	0.011	0.374
	0.264 ± 0.009	0.226 ± 0.005	0.229 ± 0.003	$\mathbf{0.221 \pm 0.006}$	0.875	0.476
Tracer variance	0.930 ± 0.005	0.949 ± 0.002	0.948 ± 0.001	0.951 ± 0.003	0.235	0.773
	0.871 ± 0.025	$\mathbf{0.756 \pm 0.012}^{\dagger}$	0.829 ± 0.013	0.827 ± 0.018	0.943	1.048
Tracer gradient energy	0.242 ± 0.043	0.428 ± 0.019	0.313 ± 0.021	0.315 ± 0.030	0.110	-0.099

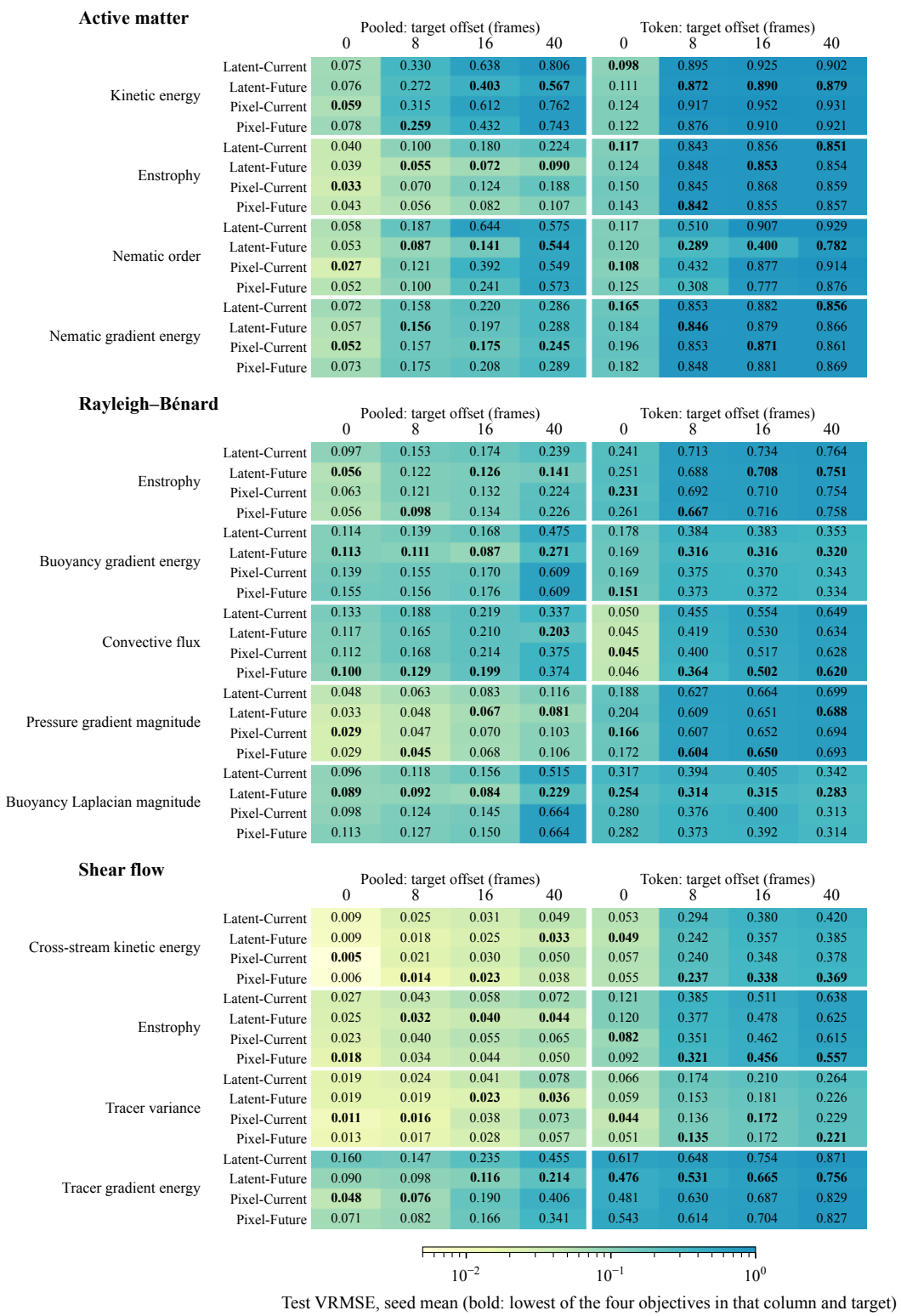

Figure 5: All pooled and token probe results by system, target and offset, for the four objectives (test VRMSE, mean over three encoder seeds). Differences are small at offset 0 and grow with offset, most in the pooled Rayleigh–Bénard and shear-flow targets.

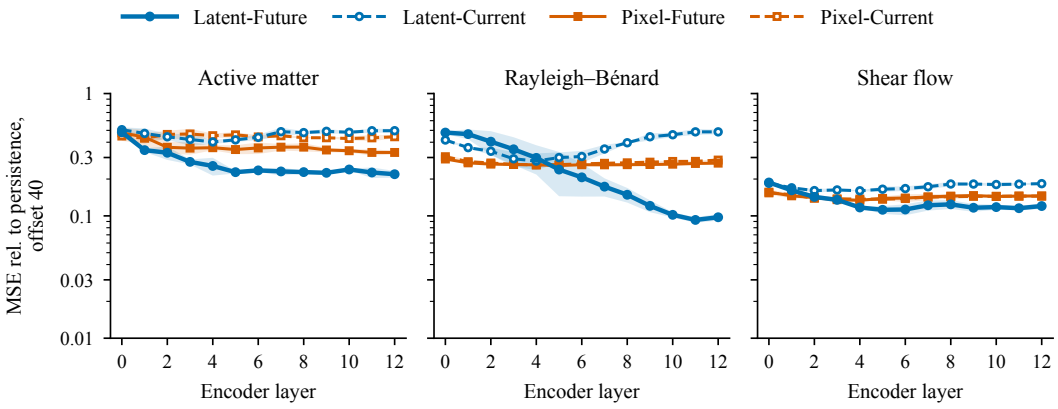

Figure 6: As Figure 3, with Ridge probes. Latent-Future reaches its lowest offset-40 error relative to persistence at layer 11 in Rayleigh–Bénard, as with MLP probes, but at layer 5 in shear flow, so the late-layer pattern of Figure 3 holds for linear probes only in Rayleigh–Bénard.

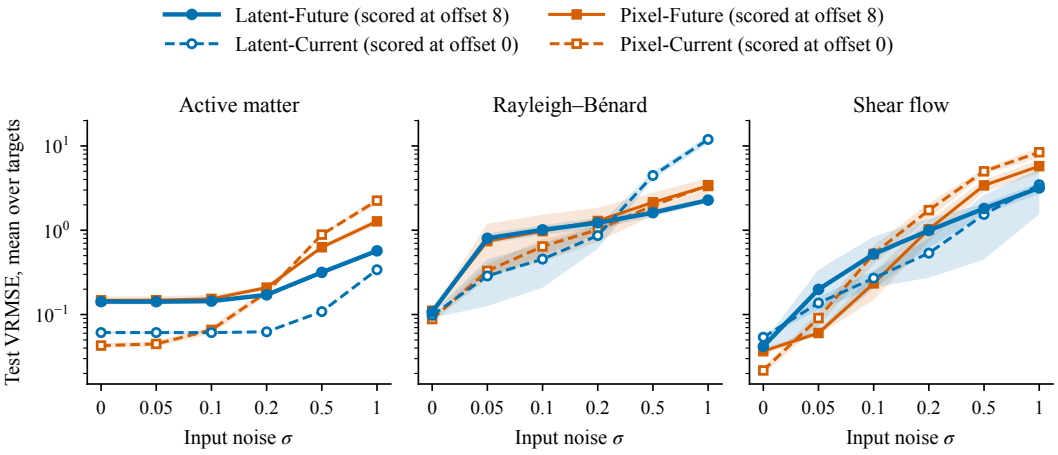

Figure 7: Pooled test VRMSE (log scale) under Gaussian input noise of standard deviation σ, with probes fitted on clean inputs and scored against clean targets; mean over targets, three encoder seeds and three corruption seeds, with ± 1 seed standard deviation bands. Current objectives are scored at offset 0 and future objectives at offset 8. Latent objectives tolerate noise far better in active matter, but not in Rayleigh–Bénard or shear flow, so latent prediction does not give uniform noise tolerance.