EVOSTATE: PREDICTING SPATIOTEMPORAL PHYSICAL DYNAMICS WITH HIERARCHICAL LATENT STATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive prediction of physical dynamics raises a basic design question: which latent states should persist across steps, and which can be recomputed? We introduce EvoState, which separates hierarchical representation learning, latent evolution, and physical readout. Its transition model jointly advances Fine, Middle, and Coarse states, while a fixed readout reconstructs physical fields from Fine. On The Well's shear-flow benchmark, the 11.2M-parameter system achieves Test VRMSE of 0.349 over steps 6–12 and 0.657 over steps 13–30, below the published benchmark baselines. On a common validation panel, EvoState also achieves lower rollout error from step eight onward than three 140M-parameter Walrus-style models trained on shear flow alone, reaching 0.608 NRMSE at step 64 compared with 1.161–2.412. In a two-epoch, single-seed comparison under a matched training schedule, recomputing coarse states from Fine or omitting them increases mean paired per-window 64-step NRMSE by 20.9% and 23.7%, respectively, relative to carrying all three levels. Recomputing coarse states at inference brings them closer to true-frame encodings but worsens physical predictions. These results favor carrying jointly evolved states under the evaluated training recipe and show that same-frame latent accuracy alone is insufficient to assess their utility for an already-trained transition.

1 INTRODUCTION

Spatiotemporal physical prediction infers how high-dimensional physical fields evolve from a finite observation history. Long-horizon prediction is commonly autoregressive: each predicted state becomes an input to the next transition (Sanchez-Gonzalez et al., 2020). A useful state must preserve future-relevant information in a form that remains accessible after repeated updates. The organization of that persistent state therefore matters as much as its one-step predictive content.

Increasing token-grid density is an intuitive response. A finer grid provides more spatial locations at which local detail can be represented. We test this choice on shear flow with three Walrus-style predictors (Walrus-SF16/32/64) that have equal trainable parameter counts and matched training budget (McCabe et al., 2025) (Fig. 1). After 24,576 training windows, increasing the token grid from 16×16 to 64×64 reduces four-channel mean NRMSE by 28.6% at $h = 1$. At $h = 8$, however, the denser model has 9.4% higher error, while its ten-step rollout takes approximately 4.4 times as long in the evaluated implementation.

A hierarchy offers another way to organize the persistent state: broad spatial structure sits at coarse resolution next to local detail in Fine. It also raises the question this paper answers. Coarse levels encoded from one frame are functions of its Fine state, so an autoregressive model could keep only Fine and recompute the rest at every step, or drop the coarse levels altogether. Which scales must persist between transitions, and which may be computed within a step?

EvoState separates a hierarchical same-frame encoder (Stage A), a joint latent transition over Fine, Middle, and Coarse (Stage B), and a fixed Fine-only physical readout (Stage C). With the encoder and readout fixed, we compare three training formulations that carry, recompute, or omit coarse states in the latent history: EvoState carries all three predicted levels, EvoState w/ recomputed coarse recomputes Middle and Coarse from the predicted Fine state, and EvoState w/o coarse keeps Fine alone. They share architecture, initialization, data order, and schedule; in the latter two, coarse losses use the recomputed states, which also changes the supervision path. Our contributions are:

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.

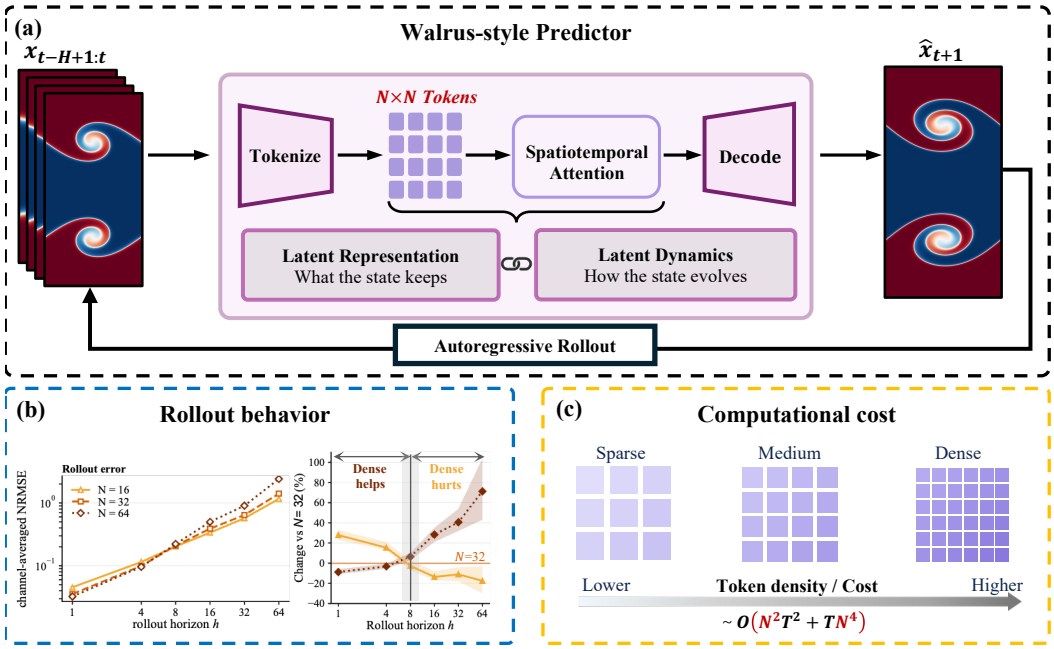

Figure 1: **Token-grid density, rollout error, and cost in shear flow.** (a) Walrus-style autoregressive predictor with an $N \times N$ token grid (schematic). (b) Four-channel mean NRMSE for $N = 16, 32, 64$ (left) and paired percentage change from $N = 32$ (right) on 336 validation origins; positive values mean higher error, and shading gives physical-file cluster bootstrap intervals. (c) Schematic cost trend; the annotation is the attention cost of one block over T frames of $N \times N$ tokens. In the evaluated single-GPU implementation, $N = 64$ takes approximately $6.1\times$ as long to train and $4.4\times$ as long for a ten-step rollout. Each model has 140,126,993 parameters and was trained on 24,576 windows with one seed.

- A fixed-parameter, fixed-budget density diagnostic: denser token grids improve the first prediction but increase later error and per-step cost.
- EvoState (11.2M parameters), a three-level persistent latent state updated by one transition and read by a fixed decoder, so that coarse-state feedback can be compared with the encoder and readout held fixed.
- Under a matched training schedule on shear flow, the carried formulation has lower long-horizon error than the recomputed and absent-coarse formulations, and lower same-frame coarse-state error does not translate into lower physical error.
- The complete system overtakes a 140M-parameter density family by the eighth step and stays below Persistence in every official Test condition.

2 Related Work

Autoregressive neural solvers. Graph-network simulators encode particles, pass messages, and decode a one-step acceleration (Sanchez-Gonzalez et al., 2020). Message-passing PDE solvers expose the transition to its own predictions through pushforward training and temporal bundling (Brandstetter et al., 2022). On grids, Walrus adapts the token layout in a transformer pretrained across continuum systems (McCabe et al., 2025), while Overtone cycles patch size during rollout to redistribute harmonic error and inference cost (Mukhopadhyay et al., 2026). Transolver pools mesh points into learnable slices before attention (Wu et al., 2024); the slices are recomputed in each layer, not carried as state.

Multiscale and hierarchical prediction. The closest multiscale methods refine one prediction. PDE-Refiner refines one next state over several denoising steps so that low-amplitude frequencies

receive their own updates (Lippe et al., 2023). Hierarchical implicit neural emulators carry coarse copies of future fields in the rolled state. The copies are fixed downsamplings, and the network predicts the next field together with later coarse copies (Jiang et al., 2025). Multiscale diffusion autoencoders couple prescribed scale vectors with a graph neural ODE to guide coarse-to-fine diffusion (Li et al., 2025). AROMA compresses a field into one set of spatially local tokens and unrolls the tokens with a diffusion transformer (Serrano et al., 2024). Our method instead keeps three nested latent levels in the persistent state and updates them jointly with one transition model. A separate frozen readout maps the predicted Fine state to the physical field.

Self-supervised state representations. The encoder's training target is a separate choice: masked autoencoders reconstruct hidden pixels (He et al., 2022), whereas I-JEPA predicts masked target features from the visible part of the same frame (Assran et al., 2023). On a controlled moving-dot task, that objective keeps slowly varying features, including a stable background that is not the causal variable (Sobal et al., 2022). ScaleAware-JEPA sets its masks from an external scale coordinate (Li, 2026). We use the masked feature loss to pretrain each resolution. What separates EvoState from these methods is which latent levels remain in the state after each transition.

3 METHOD

We study autoregressive prediction of spatiotemporal physical systems from a short history of observed fields. Let $x_t : \Omega \to \mathbb{R}^{n_{\mathrm{var}}}$ denote the field at time t over a spatial domain Ω with n_{var} observed variables. From four observations $(x_{t-3\delta}, \dots, x_t)$, the model repeatedly applies one transition to predict $\hat{x}_{t+k\delta}$ for $k = 1, \dots, K$. The sampling interval δ stays fixed within a rollout, and future observations are unavailable. We instantiate the method on a doubly periodic 256×512 shear-flow grid. The four channels are two velocity components, pressure, and a passive tracer. Each is standardized with training-set statistics. Appendix A.1 details the data and standardization; Appendix A.2 defines the physical metrics.

The method encodes four observations and keeps all three levels in its latent history (Figure 2). Stage A learns the output levels, Stage B evolves them, and Stage C reads Fine back to the field.

A tilde marks reconstruction of an observed frame and a hat marks a prediction. Modules carry parameter subscripts only while trained (A_θ and its moving average $A_{\bar{\theta}}$); frozen maps are written A, B, and C. Losses average over channels and space within each sample, then over samples.

3.1 HIERARCHICAL ENCODER (STAGE A)

A single latent grid places every spatial interaction at one resolution. We retain three output levels of a Swin-style encoder (Liu et al., 2021) (Figure 2, Stage A):

$$Z_t = A(x_t) = \big(z_t^{\mathrm{f}},\ z_t^{\mathrm{m}},\ z_t^{\mathrm{c}}\big). \tag{1}$$

In the experimental system, Fine, Middle, and Coarse have shapes $32 \times 128 \times 256$, $64 \times 64 \times 128$, and $128 \times 32 \times 64$, and all spatial operations are periodic. Patch merging gives $z_t^{\mathrm{m}} = f_{\mathrm{m}}(z_t^{\mathrm{f}})$ and $z_t^{\mathrm{c}} = f_{\mathrm{c}}(z_t^{\mathrm{m}})$ for an encoded frame. The coarser output levels make broad spatial structure directly available to the updater; Section 3.2 explains how they become independent state during rollout.

Stage A trains the three output levels with a same-frame masked feature objective, without future-frame supervision (Assran et al., 2023). An online encoder sees masked inputs; an exponential-moving-average encoder provides stopped full-frame targets at aligned hidden positions Ω_ℓ (Appendix B.3).

$$\mathcal{L}_A = \frac{1}{3} \sum_{\ell \in \{\mathrm{f},\mathrm{m},\mathrm{c}\}} \mathcal{L}_{\cos}^{\Omega_\ell}\Big(P^\ell\big(z_t^{\ell,\mathrm{mask}}\big),\ \mathrm{sg}\big[[A_{\bar{\theta}}(x_t)]_\ell\big]\Big), \tag{2}$$

Here $z_t^{\ell,\mathrm{mask}}$ is the online output level, P^ℓ is its predictor head, $\mathcal{L}_{\cos}^{\Omega_\ell}$ averages cosine distance over hidden positions, and sg stops gradients. The selected encoder checkpoint is frozen for readout and dynamics training.

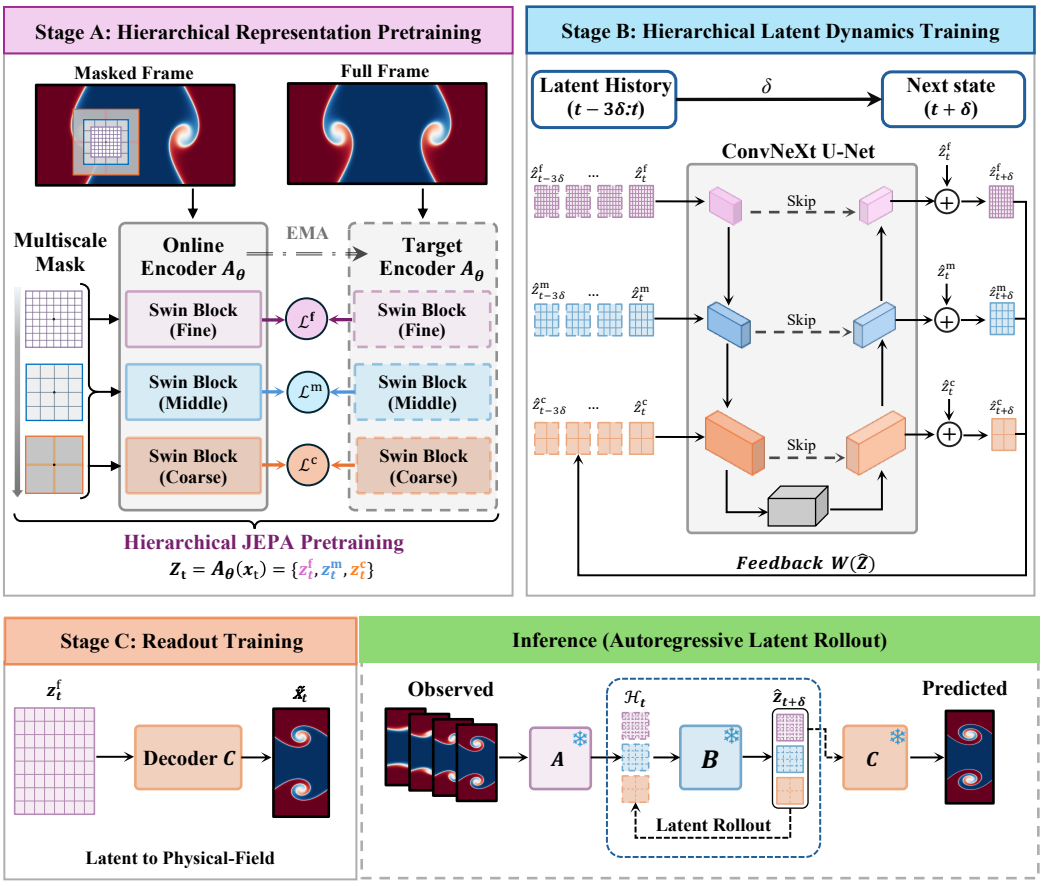

Figure 2: **Training and inference paths of EvoState.** *Stage A*: the encoder produces Fine, Middle, and Coarse representations and applies a same-frame masked JEPA objective at every output level. The masks cover the same physical cells across levels, and $\mathcal{L}^{\ell}$ denotes the hidden-position cosine loss $\mathcal{L}^{\Omega_\ell}_{\cos}$ at level ℓ. *Stage B*: a periodic ConvNeXt U-Net updates all three levels from a four-state latent history. Each predicted level is added residually, then enters the latent history as its newest state for the next transition. *Stage C*: a deterministic readout C is trained from the frozen Fine representation to the physical field. *Inference*: A initializes the latent history, B evolves it autoregressively, and C reads out each predicted Fine state.

3.2 Joint Latent Transition (Stage B)

We keep all three levels in the persistent state and evolve them jointly. The latent history is a sliding window of the four most recent hierarchical states, initialized with encoded observations ($\hat{Z}_{t-j\delta} = Z_{t-j\delta}$ for $j = 0, \ldots, 3$). One transition updates every level. The window drops its oldest state and appends the new state:

$$\hat{Z}_{t+(k+1)\delta} = \hat{Z}_{t+k\delta} + B\big(\hat{\mathcal{H}}_{t+k\delta},\, c_\delta\big), \qquad \hat{\mathcal{H}}_{t+k\delta} = \big(\hat{Z}_{t+(k-3)\delta}, \ldots, \hat{Z}_{t+k\delta}\big), \tag{3}$$

where $c_\delta = \log_2(\delta/\delta_{\mathrm{ref}})$ encodes the known interval relative to the median training interval (McCabe et al., 2025). Reynolds number and trajectory identity are not inputs.

B is one periodic ConvNeXt U-Net (Liu et al., 2022): each level's history enters the descending path at its own resolution, and skip connections return coarse context to three residual heads (Appendix B.1).

The coarse-state feedback rule is the central design choice. At encoding time, $z_t^{\mathrm{m}} = f_{\mathrm{m}}(z_t^{\mathrm{f}})$ and $z_t^{\mathrm{c}} = f_{\mathrm{c}}(z_t^{\mathrm{m}})$. During rollout, $\hat{z}^{\mathrm{m}}$ and $\hat{z}^{\mathrm{c}}$ come from their own residual heads, and in general $\hat{z}^{\mathrm{m}} \neq f_{\mathrm{m}}(\hat{z}^{\mathrm{f}})$. To test what this buys, we keep B, its three heads, and all training targets and

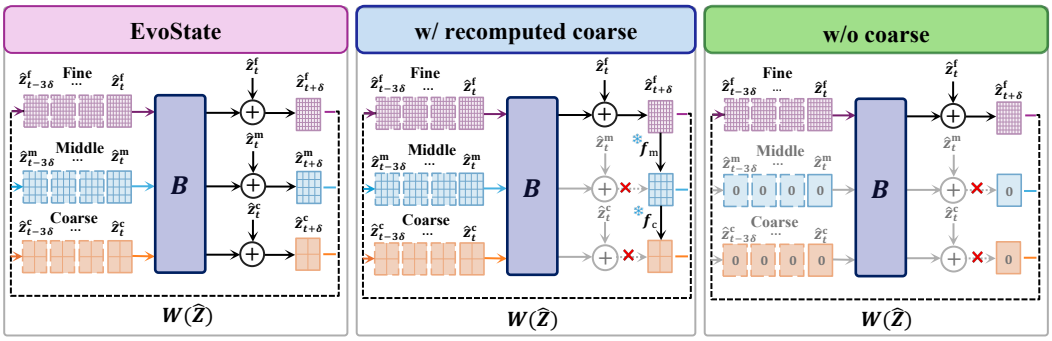

Figure 3: **Coarse-state feedback in one transition step.** All variants share B, its three heads, the encoder, the readout, and the loss weights. EvoState feeds back B's Middle and Coarse predictions; EvoState w/ recomputed coarse uses $f_{\mathrm{m}}(\hat{z}^{\mathrm{f}})$ and $f_{\mathrm{c}}(f_{\mathrm{m}}(\hat{z}^{\mathrm{f}}))$; EvoState w/o coarse zeroes both, history included. Gray heads are computed but not fed back.

compare three rules for the state $W(\hat{Z})$ that enters the window after each step:

$$W(\hat{Z}) = \begin{cases} (\hat{z}^{\mathrm{f}},\ \hat{z}^{\mathrm{m}},\ \hat{z}^{\mathrm{c}}), & \text{EvoState (carried),} \\ (\hat{z}^{\mathrm{f}},\ f_{\mathrm{m}}(\hat{z}^{\mathrm{f}}),\ f_{\mathrm{c}}(f_{\mathrm{m}}(\hat{z}^{\mathrm{f}}))), & \text{EvoState w/ recomputed coarse,} \\ (\hat{z}^{\mathrm{f}},\ 0,\ 0), & \text{EvoState w/o coarse.} \end{cases} \tag{4}$$

Here f_{m} and f_{c} are the frozen Stage A merge stages, deterministic for a given Fine state; W also applies to the encoded history. In EvoState w/ recomputed coarse and EvoState w/o coarse, the coarse loss terms are evaluated on the recomputed states, so the rule changes both the next input and the gradient path into Fine.

Each predicted endpoint is regressed onto the frozen encoding of its true field. With $d(u, v) = \|u - v\|_2^2/|u|$ in standardized latent coordinates, the Stage B objective is

$$\mathcal{L}_B = \sum_{\ell \in \{\mathrm{f},\mathrm{m},\mathrm{c}\}} w_\ell\, d\big(\hat{z}^{\ell}_{t+h\delta},\, \mathrm{sg}[z^{\ell}_{t+h\delta}]\big) + 0.1\,\mathcal{L}_{\mathrm{rank}} + 0.5\,\frac{d\big(\hat{z}^{\mathrm{f}}_{t+h\delta},\, z^{\mathrm{f}}_{t+h\delta}\big)}{\mathrm{sg}\big[d(z^{\mathrm{f}}_{t},\, z^{\mathrm{f}}_{t+h\delta})\big] + 0.05}, \tag{5}$$

averaged over sampled windows, with $(w_{\mathrm{f}}, w_{\mathrm{m}}, w_{\mathrm{c}}) = (1, 0.25, 0.25)$. Fine has unit weight because it feeds the readout. The ranking term contrasts the predicted Fine state with eight frozen encodings at other times on the same trajectory, and the last term scales Fine error by the latent Persistence distance (Appendix B.4). Training rolls out on predicted states: short horizons $h \le 8$ backpropagate through all steps, and for $h \in \{32, 64\}$ only the final eight steps carry gradients.

3.3 Physical readout (Stage C)

Stage C is fitted on the frozen Stage A representation independently of Stage B and is held fixed for the reported physical evaluations. The readout C, a periodic convolutional decoder, reconstructs the field from the Fine state alone:

$$\tilde{x}_t = C\big(z^{\mathrm{f}}_t\big), \qquad \mathcal{L}_{\mathrm{rec}} = \mathrm{MSE}\big(\tilde{x}_t, x_t\big), \tag{6}$$

in standardized field coordinates. Middle and Coarse affect the field only through their interaction with Fine inside the transition. Because C is frozen, physical-error differences between transition models cannot arise from readout adaptation.

At deployment, A encodes four observed frames once to initialize $\mathcal{H}_t$ (Figure 2, Inference). Each step applies B, drops the oldest history state, appends the predicted state, and reads its Fine output level through C: $\hat{x}_{t+k\delta} = C(\hat{z}^{\mathrm{f}}_{t+k\delta})$. Predicted fields never re-enter A. A step evaluates the roughly 9.5-million-parameter B once, plus C for each decoded frame.

4 EXPERIMENTAL SETUP

Dataset. We evaluate on The Well (Ohana et al., 2024) *shear_flow* benchmark with its official Train, Valid, and Test splits (896, 112, and 112 trajectories). Each trajectory has 200 frames of four physical channels on a doubly periodic 256×512 grid, across 28 $(\mathrm{Re}, \mathrm{Sc})$ conditions. Fields are standardized with Train statistics that stay fixed across splits (Appendix A.1). The official splits contain exact-content overlap with Train. We retain them for benchmark reporting; these scores therefore do not establish performance on an entirely unseen set of trajectories.

Baselines. The Well reports four trained baselines on this dataset, and Walrus reports four multi-physics foundation models fine-tuned on it (Table 1). For a controlled comparison on identical windows, we use three Walrus-style predictors with 16×16, 32×32, and 64×64 token grids (Walrus-SF16/32/64, 140M parameters each), trained on shear flow alone. We evaluate them with EvoState on the same 336 Valid windows (Appendix A.3). The frozen readout reproduces true same-frame fields at 0.00912 VRMSE (Appendix B.5).

Ablation variants. The three variants of Eq. (4) start from one parameter state and share a seed, Train window order, global batch 32, optimizer, learning-rate schedule, and the weighted latent regression of Eq. (5), without the ranking and Persistence-relative terms; the *full model* is the seven-epoch EvoState of Section 5.1. The first half epoch (54,208 windows) samples only $h \leq 8$; afterwards short, medium, and long horizons are drawn in proportions 75/20/5. All variants are evaluated at 0.5, 1, 1.5, and 2 epochs on the same 336 Valid windows from 28 physical files. Relative NRMSE changes are averaged per window; 95% intervals resample the 28 files as clusters. Two epochs and $h \in \{32, 64\}$ are the prespecified primary endpoints. When relative and absolute differences disagree in sign, the ordering counts as tied. A short-horizon control trains EvoState and EvoState w/ recomputed coarse from their half-epoch states to one epoch using only short horizons (Appendix C.4).

Evaluation. All model selection uses Valid. The evaluated full-model checkpoint was fixed before Test scoring. Physical errors use The Well's RMSE, NRMSE, and VRMSE implementation. The single-step Test score averages all 21,952 $H4 \rightarrow h1$ windows. Rollout scores start from each of the 112 Test trajectories, average four channels at each horizon, and then average horizons 6–12 and 13–30. A rollout receives four observed frames and no later field or true latent state. Persistence repeats the last observed frame at every horizon.

5 RESULTS

5.1 BENCHMARK RESULTS

On the same 336 Valid windows (Fig. 4), all three 140M-parameter Walrus-SF models predict $h = 1$ better, but EvoState leads from $h = 8$ onward. At $h = 64$, its NRMSE is 0.608 versus 1.161–2.412 for the Walrus-SF models and 1.712 for Persistence. Its measured step time is about 10 ms versus 84–368 ms on different hardware. The lead from $h = 16$ also holds against their 8,192- and 16,384-window checkpoints (Appendix D.3).

On official Test, EvoState reduces VRMSE relative to Persistence by 19.8% at one step and by 60.2% and 68.1% over steps 6–12 and 13–30 (Table 1). It is lower than the published The Well baselines but above the fine-tuned foundation models in the separate Walrus evaluation setting. The EvoState-to-Persistence ratio is below one in all 28 Test conditions at $h = 8$, 32, and 64, with medians 0.50, 0.34, and 0.41 (Fig. 5).

5.2 COARSE-STATE ABLATION

At the prespecified two-epoch ablation training budget, EvoState has the lowest mean NRMSE from $h = 8$ onward (Fig. 6). Relative to it, EvoState w/ recomputed coarse raises paired per-window error by 27.4% at $h = 32$ and 20.9% at $h = 64$. EvoState w/o coarse raises it by 29.8% and 23.7% (Table 2); all four intervals exclude zero. The latter matches EvoState at $h = 1$ but loses ground

Table 1: Shear-flow VRMSE under two published evaluation settings (lower is better). The Well block reports mean VRMSE from $t = 0$; each published baseline used 12 H100-hours. The foundation-model block reports median VRMSE from $T = 17$. EvoState starts at $t = 0$ in the lower block. Sources appear in Appendix D.3.

The Well benchmark

Model	Training data	One step	6–12	13–30
FNO	shear flow	1.189	> 10	> 10
TFNO	shear flow	1.472	> 10	> 10
U-net	shear flow	3.447	> 10	> 10
CNextU-net	shear flow	0.8080	2.33	> 10
EvoState	**shear flow**	**0.04155**	**0.3491**	**0.6573**

Published results under a different evaluation protocol

Model	Params	Training data	One step	1–20	21–60
MPP-L	409M	multi-physics+FT	0.0071	0.0377	0.1814
Poseidon-L	629M	multi-physics+FT	0.0090	0.0657	0.2408
DPOT-H	1.03B	multi-physics+FT	0.0162	0.0772	0.2880
Walrus-FT	1.29B	multi-physics+FT	0.0012	0.0146	0.0810
EvoState	**11.2M**	**shear flow**	**0.0204**	**0.264**	**0.389**

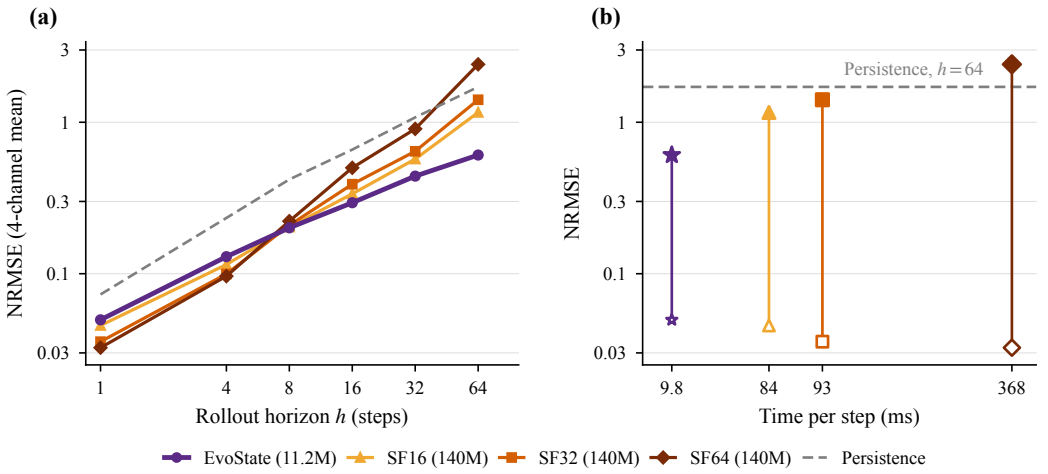

Figure 4: **Horizon and per-step cost on the Valid evaluation set.** (a) NRMSE of the full EvoState, three shear-flow-only Walrus-SF grids (24,576 training windows; orange shades and markers by grid density), and Persistence on the same 336 Valid windows. (b) The same models against inference time per rollout step; each segment runs from $h = 1$ (open) to $h = 64$ (filled); the dashed line marks Persistence at $h = 64$. EvoState was timed on an RTX 4090 and the SF grids on an RTX 5090.

with rollout, whereas the recomputed variant is already worse at $h = 1$. The two alternatives differ at $h = 32$; their relative $h = 64$ interval crosses zero.

5.3 Curriculum control

At half an epoch, EvoState w/ recomputed coarse has lower $h = 64$ error. A short-horizon control trains both variants from that common start to one epoch without adding long horizons. EvoState then leads by 9.3% at $h = 32$, while the $h = 64$ relative comparison is tied under the prespecified sign rule (Fig. 6c; Table 2). With short-horizon training alone, the carried formulation thus overtakes the recomputed one through $h = 32$; the $h = 64$ lead appears only once long horizons enter training (Appendix C.4).

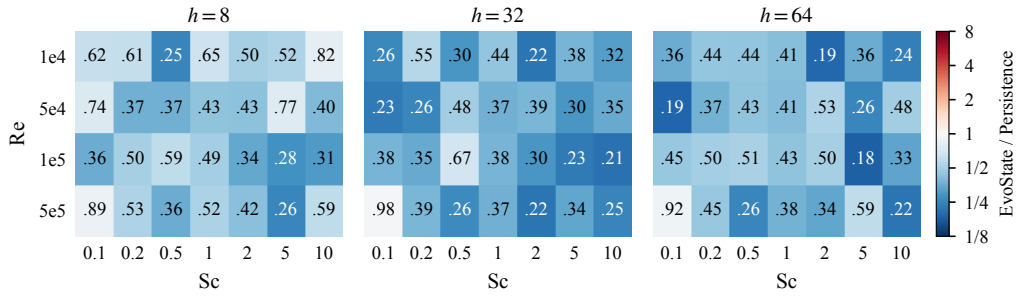

Figure 5: **EvoState relative to Persistence in every Test condition.** Each cell is the ratio of median four-channel VRMSE over the four official Test trajectories of one Reynolds–Schmidt condition; values below one favor EvoState.

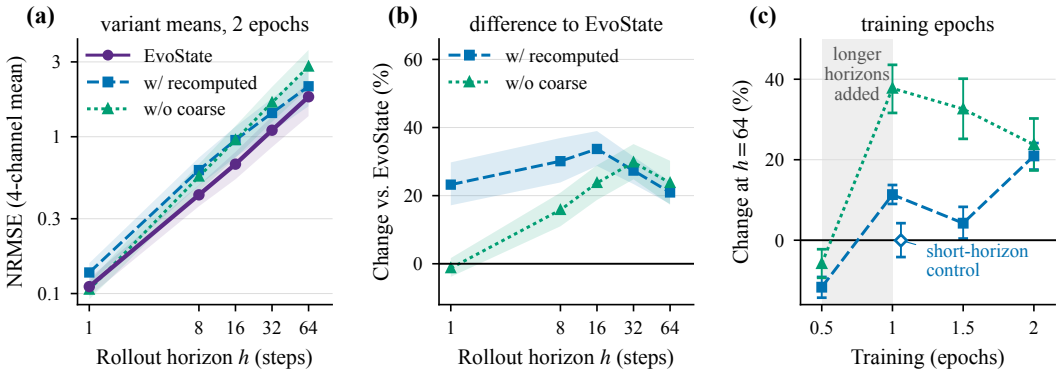

Figure 6: **Carried, recomputed, and absent coarse levels on the matched Valid evaluation set.** (a) Mean four-channel NRMSE at two epochs with 28-file cluster-bootstrap 95% intervals. (b) Mean paired per-window change against EvoState at two epochs; positive values favor EvoState. (c) The same change at $h = 64$ across training epochs. Medium and long horizons enter training after 0.5 epoch (gray band); the open marker is the short-horizon control. All variants share the ablation recipe and one seed.

5.4 Coarse-state replacement

We replace the full model's carried Middle and Coarse states at each inference step with quantities derived from predicted Fine. Although their $h = 16$ latent errors fall by 43% and 75%, respectively, rollout NRMSE rises 113% at $h = 64$ (95% interval 77–154%; Fig. 7). For the trained transition, lower coarse-state error relative to true-frame encodings does not translate into lower physical prediction error (Appendix D.4).

5.5 Predicted fields and spectra

On the pre-specified medium-Reynolds Test trajectory (Fig. 8), the forecast follows the rolled-up tracer regions through $h = 32$. At $h = 64$ the vortex layout remains while filaments lose sharpness and shift in phase (other regimes in Appendix E). At $h = 64$, the full model retains tracer coherence of 0.90 in the lowest wavenumber band but only 0.22 in the highest; high-band energy falls to 0.29 of truth. The scale-resolved and per-channel diagnostics are in Appendices D.5 and C.5.

6 Discussion and Limitations

Under the matched schedule, carrying B's own Middle and Coarse predictions gives the lowest long-horizon error. In the full model, coarse states closer to their true encodings worsen rollout;

Table 2: Paired per-window relative NRMSE change (%) against EvoState on 336 Valid windows, with 28-file cluster-bootstrap 95% intervals. Positive values favor EvoState. Two epochs is the prespecified primary training budget; all rows use the ablation recipe.

	w/ recomputed − EvoState		w/o coarse − EvoState	
Training	$h = 32$	$h = 64$	$h = 32$	$h = 64$
0.5 epoch (common start)	-6.9 $[-9.9, -4.2]$	-11.7 $[-14.3, -9.1]$	$+3.4$ $[0.6, 6.3]^*$	-5.8 $[-9.3, -2.2]$
1 epoch	$+16.9$ $[14.1, 19.8]$	$+11.3$ $[9.0, 13.7]$	$+43.6$ $[37.1, 50.3]$	$+37.7$ $[31.6, 43.6]$
1 epoch, short-horizon control	$+9.3$ $[4.6, 14.1]$	-0.0 $[-4.2, 4.2]^*$	–	–
1.5 epochs	$+10.2$ $[6.3, 14.4]$	$+4.2$ $[0.4, 8.3]$	$+31.1$ $[24.1, 38.3]$	$+32.6$ $[25.2, 40.1]$
2 epochs	$+27.4$ $[23.4, 31.5]$	$+20.9$ $[17.6, 24.1]$	$+29.8$ $[24.8, 35.1]$	$+23.7$ $[17.3, 30.2]$

*Relative and absolute differences have opposite signs; the prespecified rule counts the ordering as tied.

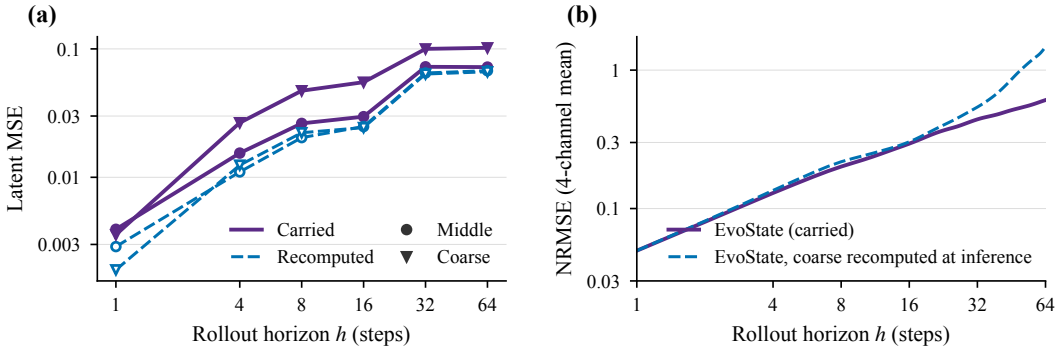

Figure 7: **Coarse-state replacement at inference on Valid.** (a) Latent MSE of Middle (circles) and Coarse (triangles) against true-frame encodings, carried by EvoState (solid) or recomputed from the predicted Fine state at every step (dashed). (b) NRMSE of the two rollouts. Both use the trained full model; neither is the separately trained EvoState w/ recomputed coarse.

the replacement also changes the joint states the transition encounters, leaving adaptation to the carried-state distribution as a possible explanation. As limitations, EvoState's lead over the denser Walrus-SF grids comes with a different architecture and training budget, so it is not attributed to token count. The one-seed ablations use a shorter recipe. Walrus-SF timing uses different hardware, and foundation-model scores start rollouts at another time. Pressure is worse than Persistence at one step (0.0530 versus 0.0340 VRMSE), and the smallest scales lose energy during rollout.

7 Conclusion

On shear flow and under a matched two-epoch schedule, carrying three levels yields lower long-rollout error than recomputing or omitting the coarse levels. The full 11.2M-parameter EvoState reaches 0.61 NRMSE at $h = 64$ versus 1.16–2.41 for the 140M-parameter Walrus-SF models, and reduces official Test rollout VRMSE by 60.2–68.1% relative to Persistence.

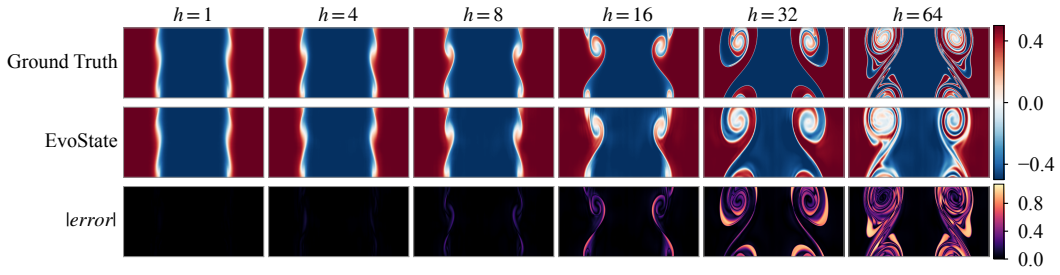

Figure 8: **Test tracer rollout at** $\mathrm{Re} = 5 \times 10^4$**,** $\mathrm{Sc} = 10$**.** Rows show truth, EvoState, and absolute error at $h = 1, 4, 8, 16, 32, 64$; truth and prediction share a color scale.

AI USE STATEMENT

Generative AI tools were used in three ways. For writing, they helped polish the wording, grammar, and concision of author-written text. For retrieval, they helped search for and summarize related work; the authors checked each cited source against the original publication. For research execution, they assisted in implementing and debugging training and evaluation code and in organizing experiment records. The authors checked AI-assisted output against the source code and result records and take full responsibility for the content of this paper.

ETHICS STATEMENT

This work studies publicly available simulated physical fields and does not involve human participants or personal data.

REPRODUCIBILITY STATEMENT

To support reproduction, we provide the data split and preprocessing configuration, model and training specifications, checkpoint rule, and evaluation procedure in the main text and Appendices A– E. The underlying *shear_flow* dataset is publicly available through The Well. We will publicly release the code and trained model weights.

REFERENCES

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

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

Zhongkai Hao, Chang Su, Songming Liu, Julius Berner, Chengyang Ying, Hang Su, Anima Anandkumar, Jian Song, and Jun Zhu. DPOT: Auto-regressive denoising operator transformer for large-scale PDE pre-training. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 17616–17635, 2024. URL https://arxiv.org/abs/2403.03542.

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*, pp. 16000–16009, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/He_Masked_Autoencoders_Are_Scalable_Vision_Learners_CVPR_2022_paper.html.

Maximilian Herde, Bogdan Raonic, Tobias Rohner, Roger Käppeli, Roberto Molinaro, Emmanuel De Bézenac, and Siddhartha Mishra. POSEIDON: Efficient foundation models for PDEs. In *Advances in Neural Information Processing Systems*, 2024. URL https://papers.neurips.cc/paper_files/paper/2024/file/84e1b1ec17bb11c57234e96433022a9a-Paper-Conference.pdf.

Ruoxi Jiang, Xiao Zhang, Karan Jakhar, Peter Y. Lu, Pedram Hassanzadeh, Michael Maire, and Rebecca Willett. Hierarchical implicit neural emulators. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2506.04528.

Guang-Xing Li. ScaleAware-JEPA: Latent representation for discovery in multiscale physical fields. *arXiv preprint arXiv:2606.29723*, 2026. URL https://arxiv.org/abs/2606.29723.

Ruikun Li, Jingwen Cheng, Huandong Wang, Qingmin Liao, and Yong Li. Predicting the dynamics of complex system via multiscale diffusion autoencoder. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2025. URL https://arxiv.org/abs/2505.02450.

Phillip Lippe, Bastiaan S. Veeling, Paris Perdikaris, Richard E. Turner, and Johannes Brandstetter. PDE-Refiner: Achieving accurate long rollouts with neural PDE solvers. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2308.05732.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021. URL https://arxiv.org/abs/2103.14030.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022. URL https://arxiv.org/abs/2201.03545.

Michael McCabe, Bruno Regaldo-Saint Blancard, Liam 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 spatiotemporal surrogate models. In *Advances in Neural Information Processing Systems*, 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. *arXiv preprint arXiv:2511.15684*, 2025. URL https://arxiv.org/abs/2511.15684.

Payel Mukhopadhyay, Michael McCabe, Ruben Ohana, and Miles Cranmer. Overtone: Cyclic patch modulation for clean, efficient, and flexible physics emulators. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2507.09264.

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. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 37, pp. 44989–45037, 2024.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter W. Battaglia. Learning to simulate complex physics with graph networks. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. URL https://arxiv.org/abs/2002.09405.

Louis Serrano, Thomas X. Wang, Etienne Le Naour, Jean-Noël Vittaut, and Patrick Gallinari. AROMA: Preserving spatial structure for latent PDE modeling with local neural fields. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.02176.

Vlad Sobal, Jyothir S V, Siddhartha Jalagam, Nicolas Carion, Kyunghyun Cho, and Yann LeCun. Joint embedding predictive architectures focus on slow features. In *NeurIPS 2022 Workshop on Self-Supervised Learning: Theory and Practice*, 2022. URL https://arxiv.org/abs/2211.10831.

Haixu Wu, Huakun Luo, Haowen Wang, Jianmin Wang, and Mingsheng Long. Transolver: A fast transformer solver for PDEs on general geometries. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2402.02366.

A Data and evaluation

A.1 Dataset and normalization

We use The Well *shear_flow* dataset: four physical channels on a doubly periodic 256×512 grid under 28 Reynolds–Schmidt conditions, $\mathrm{Re} \in \{10^4, 5 \times 10^4, 10^5, 5 \times 10^5\}$ and $\mathrm{Sc} \in \{0.1, 0.2, 0.5, 1, 2, 5, 10\}$. Each condition is one HDF5 file per split. The official-v1 Train, Valid, and Test files hold 32, 4, and 4 trajectories per condition (896, 112, and 112 in total), each with 200

frames. All results use this official split unchanged. An exact-content check of the released records finds that some Test trajectories also occur in Train.

Fields are standardized per channel with Train statistics that stay fixed on Valid and Test: pressure $\mu = -1.23 \times 10^{-9}$, $\sigma = 0.0820523$; tracer $\mu = 0.00238452$, $\sigma = 0.358282$; the two velocity components share $\mu = 0.00119217$, $\sigma = 0.293614$.

A.2 Metrics

Physical errors use the RMSE, NRMSE, and VRMSE implementations of The Well (commit ad50de0e), in physical units over the full grid. For target y and prediction $\hat{y}$ over n grid points,

$$\mathrm{NRMSE} = \sqrt{\frac{\frac{1}{n}\sum_q (\hat{y}_q - y_q)^2}{\frac{1}{n}\sum_q y_q^2 + 10^{-7}}}, \qquad \mathrm{VRMSE} = \sqrt{\frac{\frac{1}{n}\sum_q (\hat{y}_q - y_q)^2}{s_y^2 + 10^{-7}}},$$

where s_y^2 is the spatial sample variance. Each score is computed per sample, horizon, and channel. The official single-step score averages all 21,952 Test $H4 \to h1$ windows. The rollout score averages 112 Test trajectories and four channels at each horizon, then averages horizons 6–12 and 13–30 with both endpoints included. VRMSE divides by the spatial variance of the target, so frames with nearly uniform pressure or u_y produce a heavy upper tail; medians are reported alongside means where this matters.

Spectral diagnostics remove each field's spatial mean and apply an orthonormal 2D Fourier transform. Mode $q = (k_y, k_x)$ belongs to shell $\kappa(q) = \mathrm{round}\sqrt{k_y^2 + (k_x/2)^2}$; the bands are shells 1–4, 5–16, and 17–128. With predicted and true coefficients $\hat{Y}$, Y and band sums $P_K = \sum |\hat{Y}|^2$, $T_K = \sum |Y|^2$, $X_K = \sum \mathrm{Re}[\hat{Y}\overline{Y}]$, the energy ratio is $E_K = P_K/T_K$ and the coherence is $C_K = X_K/\sqrt{P_K T_K}$.

A.3 Evaluation sets

The Valid evaluation set contains 336 windows: all 112 official Valid trajectories, each started at frames 0, 65, and 130, from 28 physical files. Each model receives its own history length (four frames for EvoState, six for Walrus-SF, ending at the same frame) and predicts the same 64 target frames without further observations. Intervals resample the 28 files as clusters with all their windows. The Walrus-SF family uses 140,126,993 trainable parameters per model with 16×16, 32×32, and 64×64 token grids, trained on shear flow alone with equal training windows. The pre-specified Test field examples in Figs. 8 and 17 were fixed by Reynolds regime before their predictions were viewed.

B Implementation details

B.1 Architecture

Stage A is a three-stage Swin V2 encoder with channels $32/64/128$ and depths $2/4/2$ (670,948 parameters). Stage C is a Fine-only periodic convolutional readout (1,005,328 parameters). Stage B (9,493,216 parameters) is a periodic ConvNeXt U-Net with encoder widths $64/128/256/384$, block depths $2/2/4/4$, and decoder widths $256/128/64$ with two blocks per stage. Downsampling applies channel normalization, a separable periodic $[1,3,3,1]/8$ low-pass filter, stride-two sampling, and projection; upsampling is periodic bilinear interpolation at phase-preserving coordinates. Write D and U for these operators, $[\cdot;\cdot]$ for channel concatenation, $F_\ell, \tilde{F}_\ell$ for 1×1 fusion convolutions, and $S_\ell, \tilde{S}_\ell$ for the encoder and decoder ConvNeXt stages. With the flattened history $\hat{\mathcal{H}}^\ell$ of level ℓ,

$$\begin{aligned}
e^{\mathrm{f}} &= S_{\mathrm{f}}\big(\mathrm{Stem}([\hat{\mathcal{H}}^{\mathrm{f}}; c_\delta])\big), & u^{\mathrm{c}} &= \tilde{S}_{\mathrm{c}}\big(\tilde{F}_{\mathrm{c}}([U(e^{\mathrm{b}}); e^{\mathrm{c}}])\big),\\
e^{\mathrm{m}} &= S_{\mathrm{m}}\big(F_{\mathrm{m}}([D(e^{\mathrm{f}}); \hat{\mathcal{H}}^{\mathrm{m}}])\big), & u^{\mathrm{m}} &= \tilde{S}_{\mathrm{m}}\big(\tilde{F}_{\mathrm{m}}([U(u^{\mathrm{c}}); e^{\mathrm{m}}])\big),\\
e^{\mathrm{c}} &= S_{\mathrm{c}}\big(F_{\mathrm{c}}([D(e^{\mathrm{m}}); \hat{\mathcal{H}}^{\mathrm{c}}])\big), & u^{\mathrm{f}} &= \tilde{S}_{\mathrm{f}}\big(\tilde{F}_{\mathrm{f}}([U(u^{\mathrm{m}}); e^{\mathrm{f}}])\big),\\
e^{\mathrm{b}} &= S_{\mathrm{b}}\big(D(e^{\mathrm{c}})\big), & \Delta^\ell &= \mathrm{Head}_\ell(u^\ell),
\end{aligned}$$

and $B(\hat{\mathcal{H}}, c_\delta) = (\Delta^{\mathrm{f}}, \Delta^{\mathrm{m}}, \Delta^{\mathrm{c}})$.

B.2 Algorithm

Table 3 writes out one autonomous rollout and one Stage B training step. The feedback map W of Eq. (4) and the loss states Λ are the only lines that differ between EvoState, EvoState w/ recomputed coarse, and EvoState w/o coarse. W also acts on the encoded history, so EvoState w/o coarse starts from zero Middle and Coarse history; for the other two rules it leaves encoded states unchanged. Λ returns the raw updated state $\hat{Z}$ for EvoState and $(\hat{z}^{\mathrm{f}}, f_{\mathrm{m}}(\hat{z}^{\mathrm{f}}), f_{\mathrm{c}}(f_{\mathrm{m}}(\hat{z}^{\mathrm{f}})))$ for the other two. At inference the full model always uses the carried rule. During training, the latent targets come from the frozen encoder applied to the true future frames, and for long horizons the self-generated prefix is detached so that gradients flow only through the final steps.

Table 3: Autonomous rollout (left) and one Stage B training step (right). sg stops gradients; W is the feedback rule of Eq. (4) and Λ gives the states used in the loss.

Rollout$(x_{t-3\delta:t}, K)$

$\hat{Z}_{t-j\delta} \leftarrow W(A(x_{t-j\delta}))$, $j = 0, \dots, 3$
$\hat{\mathcal{H}} \leftarrow (\hat{Z}_{t-3\delta}, \dots, \hat{Z}_t)$
for $k = 0, \dots, K-1$:
  $\hat{Z} \leftarrow \hat{Z}_{t+k\delta} + B(\hat{\mathcal{H}}, c_\delta)$
  $\hat{Z}_{t+(k+1)\delta} \leftarrow W(\hat{Z})$
  shift $\hat{Z}_{t+(k+1)\delta}$ into $\hat{\mathcal{H}}$, drop oldest
  $\hat{x}_{t+(k+1)\delta} \leftarrow C(\hat{z}^{\mathrm{f}}_{t+(k+1)\delta})$
return $\hat{x}_{t+\delta:t+K\delta}$

Train step$(x_{t-3\delta:t+h\delta}, h)$

$\hat{\mathcal{H}} \leftarrow \big(W(A(x_{t-j\delta}))\big)_{j=3,\dots,0}$; $Z^+ \leftarrow \mathrm{sg}[A(x_{t+h\delta})]$
for $k = 0, \dots, h-1$:
  if $h \in \{32, 64\}$ and $k = h - 8$: $\hat{\mathcal{H}} \leftarrow \mathrm{sg}[\hat{\mathcal{H}}]$
  $\hat{Z} \leftarrow \hat{Z}_{t+k\delta} + B(\hat{\mathcal{H}}, c_\delta)$
  $\hat{Z}_{t+(k+1)\delta} \leftarrow W(\hat{Z})$; shift into $\hat{\mathcal{H}}$
$\mathcal{L} \leftarrow \sum_\ell w_\ell\, d([\Lambda(\hat{Z})]_\ell, z^{\ell+})$, $\hat{Z}$ from step h
  (full model adds the ranking and Persistence terms)
update B by AdamW on $\nabla\mathcal{L}$

B.3 Stage A training

Stage A learns these output levels by predicting hidden same-frame features, with no future-frame supervision (Assran et al., 2023). One mask m is sampled on the field grid and mapped to all three levels by conservative cell aggregation. The online encoder replaces hidden Fine tokens after its stem; its coarser output levels follow from that masked input. Here A_θ denotes the online encoder applied to the masked input, and $A_{\bar{\theta}}$ its exponential-moving-average target encoder applied to the complete frame. At level $\ell \in \{\mathrm{f}, \mathrm{m}, \mathrm{c}\}$, we write $z_t^{\ell,\mathrm{mask}} = [A_\theta(x_t; m)]_\ell$ for the masked online output level and Ω_ℓ for its hidden positions, aligned in physical space across the hierarchy. A predictor head P^ℓ maps the masked output level toward the stopped target at those positions. The target follows $\bar{\theta} \leftarrow \tau\bar{\theta} + (1-\tau)\theta$; its coefficient is recorded below.

Stage A hides 25% of the 512 16×16 field tiles per frame with one mask shared by all output levels; each output level predicts the EMA target at the hidden cells with an equally weighted cosine loss. The selected seed-17 run used FP32, global batch 32 on two RTX 4090 GPUs, AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$, no weight decay, and EMA momentum 0.996. The learning rate warms up for 1,400 updates to 4×10^{-5}, decays to 10^{-6} by update 67,200, and then follows two short cosine cycles ($8 \times 10^{-6} \to 10^{-6}$ by update 78,400 and $4 \times 10^{-6} \to 10^{-6}$ by update 112,000), for 20 epochs or 3,584,000 frames in total. The update-112,000 EMA checkpoint, with masked-cosine Valid loss 0.00140534, was fixed before any Stage B evaluation.

B.4 Stage B objective and schedule

With per-window mean squared error d_i, the ranking term averages eight Fine-state hinges,

$$\mathcal{L}_{\mathrm{rank}} = \frac{1}{8B} \sum_{i=1}^{B} \sum_{j=1}^{8} \max\big(d_i(\hat{z}_i^{\mathrm{f}}, z_i^{\mathrm{f}+}) - d_i(\hat{z}_i^{\mathrm{f}}, z_{ij}^{\mathrm{f}-}) + 0.5\, d_i(z_i^{\mathrm{f}+}, z_{ij}^{\mathrm{f}-}),\, 0\big).$$

For $h \le 8$ the negatives are the current Fine state and the teacher frames $t + j\delta$, $1 \le j \le 8$, $j \ne h$; for $h \in \{32, 64\}$ they are the current state and frames $t + (h-j)\delta$, $1 \le j \le 7$. All negatives are frozen and detached. The Persistence-relative term in Eq. (5) uses the last observed Fine state of the same window.

Training used AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 10^{-4} on decay parameters, base learning rate 10^{-4}, and gradient clipping at 1.0. The schedule has two phases: the first 162,624 training windows use global batch 4 with a 0.1-epoch linear warm-up and cosine decay; the remaining windows use global batch 32 on four GPUs with cosine decay from 3×10^{-5} to 10^{-6}. The evaluated checkpoint after seven epochs was fixed on Valid before the Test scores were computed.

B.5 Readout accuracy

Stage C is trained on frozen Stage A Fine states and then held fixed. Table 4 evaluates it on true same-frame Fine states from all 22,400 Valid frames. Tracer and u_x have VRMSE below 0.002, whereas pressure and u_y have larger relative errors despite small RMSE. These on-distribution scores isolate the readout's same-frame error from the additional error introduced by latent rollout.

Table 4: Reconstruction through the frozen readout on 22,400 Valid Fine states.

Field	RMSE	VRMSE
Pressure	6.30×10^{-5}	0.01539
Tracer	6.62×10^{-4}	0.00193
u_x	2.46×10^{-4}	0.00060
u_y	1.80×10^{-4}	0.01857

B.6 Computational cost

Table 5 lists the measured cost of each stage. Stage A encodes the four observed frames once per rollout; every subsequent step evaluates B once and C for each decoded frame. Selected training runs used about 17, 43, and 10 RTX 4090 GPU-hours for Stages A, B, and C, excluding earlier screening and the final Test evaluation. B takes 7.9 ms per step and C 1.6 ms per decoded frame on the measured RTX 4090 setup. The larger FLOP count of C does not translate into greater latency in this implementation. Walrus-SF timings use an RTX 5090, so the rows document measured system cost rather than a hardware-controlled speed comparison.

Table 5: Parameters and inference cost at batch size one. EvoState was timed on an RTX 4090; Walrus-SF was timed on an RTX 5090 in its training framework.

Component	Params	GFLOPs	ms
A (per encoded frame)	0.67M	9.4	8.2
B (per step)	9.49M	41.4	7.9
C (per decoded frame)	1.01M	65.8	1.6
EvoState, 64 steps	11.2M	–	630
Walrus-SF16 (per step)	140.1M	–	84
Walrus-SF32 (per step)	140.1M	–	93
Walrus-SF64 (per step)	140.1M	–	368

C Coarse-state ablation details

C.1 Training setup

All three variants use the Stage B architecture of Appendix B.1 with its three residual heads and the frozen Stage A and Stage C checkpoints of the full model. They start from one parameter state and share seed 17, the ordered Train windows, global batch 32 on four RTX 4090 GPUs, AdamW, and the learning-rate clock. The objective is the weighted latent regression of Eq. (5) with weights $1/0.25/0.25$ and no ranking or Persistence-relative term. The only intervention is the coarse-state feedback rule of Eq. (4). EvoState w/o coarse sets the Middle and Coarse history inputs to zero; its coarse loss terms, like those of EvoState w/ recomputed coarse, are evaluated on the states recomputed from the predicted Fine state, so the rule also changes the gradient path. The first 54,208 windows (half an epoch) sample only short horizons $h \leq 8$. Later draws sample short,

medium, and long horizons in proportions 75/20/5. Matched checkpoints are taken after 54,208, 108,416, 162,624, and 216,832 windows (0.5 to 2 epochs).

C.2 Pairing and uncertainty

The pairing procedure admits a pair only if the two variants share origin identifiers, file membership, target horizons, metric code, physical units, readout checkpoint, and evaluator version. It reports each variant's mean NRMSE, the mean paired absolute difference, and the mean paired relative difference, which divides the difference by the reference variant's NRMSE within each window before averaging and is therefore not the ratio of two variant means. A cluster bootstrap resamples the 28 files with replacement, keeping each file's windows together (2,000 draws), and gives percentile 95% intervals. The rule fixed before the two-epoch results were read treats an ordering as tied when the relative and absolute differences disagree in sign, or when the one- and two-epoch directions disagree.

C.3 Complete paired results

Tables 6–8 give all three contrasts at every matched training budget and horizon, in relative and absolute units. Fig. 9 shows the corresponding variant means. At half an epoch the three variants are nearly indistinguishable up to $h = 32$, and all exceed NRMSE 4 at $h = 64$. From one epoch on, EvoState has the lowest mean from $h = 8$ at every matched training budget. Between the alternatives, EvoState w/o coarse is better than EvoState w/ recomputed coarse at $h = 1$ and $h = 8$ at two epochs (-17.4% and -6.2%), close to it at $h = 16$, and worse at $h = 32$ and $h = 64$. The $h = 64$ difference has a relative interval from -0.1% to $+11.4\%$ and an absolute difference of $+0.705$ NRMSE, so the alternatives are not reliably ordered there.

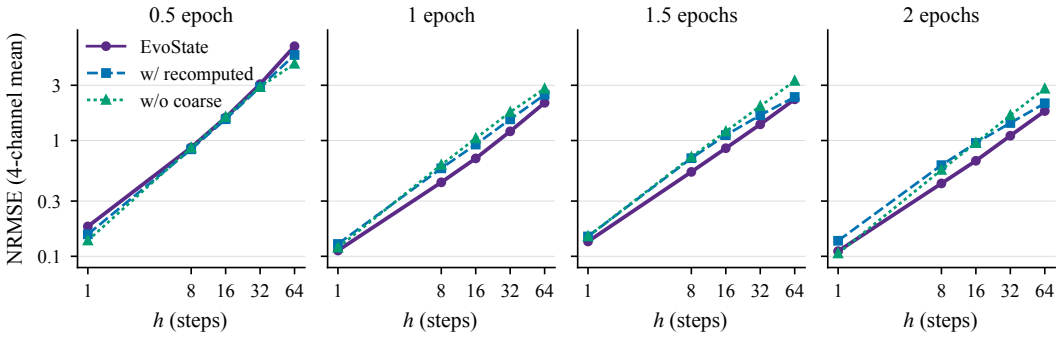

Figure 9: **Variant means at every matched training budget.** Four-channel NRMSE of the three coarse-state feedback rules on the 336-window Valid evaluation set after 0.5, 1, 1.5, and 2 epochs.

C.4 Curriculum control

The half-epoch checkpoints of EvoState and EvoState w/ recomputed coarse are trained further to one epoch with the short-horizon sampling of the first phase and every other setting unchanged. Fig. 10 and Table 9 compare them on the same Valid evaluation set. Without medium or long horizons, both variants reach higher long-horizon error than their mixed-curriculum counterparts (EvoState 3.37 against 2.12 at $h = 64$), but EvoState overtakes EvoState w/ recomputed coarse at every horizon up to $h = 32$. Through $h = 32$, the reversal of Fig. 6c therefore appears with further short-horizon training alone; at $h = 64$ the paired relative comparison remains tied under the prespecified rule.

C.5 Per-channel and spectral results

Fig. 11 splits the two-epoch variant means by channel. The coarse-state feedback rule matters most for u_y and tracer. EvoState has the lowest tracer error at every horizon and the lowest u_y error from $h = 8$; at $h = 64$, its u_y NRMSE is 3.26 against 4.11 for EvoState w/ recomputed coarse and 7.24

Table 6: **EvoState w/ recomputed coarse minus EvoState.** Matched four-channel NRMSE on 336 Valid origins; each value is the paired mean followed by a 95% interval from resampling 28 physical files. Relative values average window-wise ratios; absolute values are in NRMSE units. Positive values favor the reference variant.

Epoch	h	Relative change (%) [95% CI]	Absolute change [95% CI]
0.5	1	-9.6 $[-11.4, -7.9]$	-0.0263 $[-0.0302, -0.0221]$
	8	-5.0 $[-6.6, -3.4]$	-0.0318 $[-0.0543, -0.0105]$
	16	-4.1 $[-6.2, -2.1]$	-0.0517 $[-0.1245, +0.0096]$
	32	-6.9 $[-9.9, -4.2]$	-0.1731 $[-0.3760, +0.0027]$
	64	-11.7 $[-14.3, -9.1]$	-1.0419 $[-1.5144, -0.6293]$
1	1	$+15.9$ $[+11.5, +20.4]$	$+0.0156$ $[+0.0084, +0.0233]$
	8	$+23.4$ $[+19.5, +27.7]$	$+0.1401$ $[+0.0969, +0.1874]$
	16	$+21.1$ $[+17.6, +24.7]$	$+0.2224$ $[+0.1512, +0.2987]$
	32	$+16.9$ $[+14.1, +19.8]$	$+0.3243$ $[+0.2186, +0.4387]$
	64	$+11.3$ $[+9.0, +13.7]$	$+0.3669$ $[+0.2536, +0.4836]$
1.5	1	$+8.6$ $[+5.8, +11.5]$	$+0.0141$ $[+0.0082, +0.0199]$
	8	$+22.8$ $[+19.6, +26.1]$	$+0.1679$ $[+0.1260, +0.2110]$
	16	$+19.5$ $[+16.0, +23.3]$	$+0.2550$ $[+0.1859, +0.3242]$
	32	$+10.2$ $[+6.3, +14.4]$	$+0.2743$ $[+0.1830, +0.3675]$
	64	$+4.2$ $[+0.4, +8.3]$	$+0.1056$ $[+0.0360, +0.1781]$
2	1	$+23.2$ $[+17.2, +29.8]$	$+0.0255$ $[+0.0166, +0.0350]$
	8	$+30.1$ $[+23.9, +36.9]$	$+0.1868$ $[+0.1252, +0.2504]$
	16	$+33.7$ $[+28.7, +39.0]$	$+0.2838$ $[+0.2010, +0.3757]$
	32	$+27.4$ $[+23.4, +31.5]$	$+0.3171$ $[+0.2196, +0.4263]$
	64	$+20.9$ $[+17.6, +24.1]$	$+0.2997$ $[+0.1945, +0.4147]$

Table 7: **EvoState w/o coarse minus EvoState.** Matched four-channel NRMSE on 336 Valid origins; each value is the paired mean followed by a 95% interval from resampling 28 physical files. Relative values average window-wise ratios; absolute values are in NRMSE units. Positive values favor the reference variant.

Epoch	h	Relative change (%) [95% CI]	Absolute change [95% CI]
0.5	1	-19.5 $[-21.2, -17.7]$	-0.0442 $[-0.0494, -0.0392]$
	8	-3.7 $[-6.3, -0.8]$	-0.0058 $[-0.0456, +0.0354]$
	16	$+1.4$ $[-1.1, +4.0]$	$+0.0184$ $[-0.0694, +0.1077]$
	32	$+3.4$ $[+0.6, +6.3]$	-0.1663 $[-0.4073, +0.0640]$
	64	-5.8 $[-9.3, -2.2]$	-1.9190 $[-2.6525, -1.2289]$
1	1	$+11.7$ $[+7.5, +16.1]$	$+0.0066$ $[+0.0023, +0.0111]$
	8	$+33.6$ $[+27.8, +39.5]$	$+0.1809$ $[+0.1285, +0.2347]$
	16	$+40.3$ $[+34.1, +46.2]$	$+0.3430$ $[+0.2375, +0.4515]$
	32	$+43.6$ $[+37.1, +50.3]$	$+0.5608$ $[+0.3770, +0.7589]$
	64	$+37.7$ $[+31.6, +43.6]$	$+0.6828$ $[+0.3922, +0.9939]$
1.5	1	$+17.3$ $[+13.5, +21.1]$	$+0.0145$ $[+0.0102, +0.0195]$
	8	$+43.4$ $[+35.1, +52.1]$	$+0.1835$ $[+0.1473, +0.2224]$
	16	$+38.6$ $[+31.2, +46.4]$	$+0.3386$ $[+0.2375, +0.4427]$
	32	$+31.1$ $[+24.1, +38.3]$	$+0.6006$ $[+0.3750, +0.8439]$
	64	$+32.6$ $[+25.2, +40.1]$	$+1.0049$ $[+0.5979, +1.4328]$
2	1	-1.3 $[-3.9, +1.7]$	-0.0048 $[-0.0082, -0.0014]$
	8	$+15.8$ $[+11.0, +20.9]$	$+0.1317$ $[+0.0837, +0.1789]$
	16	$+23.7$ $[+18.7, +28.9]$	$+0.2846$ $[+0.1947, +0.3755]$
	32	$+29.8$ $[+24.8, +35.1]$	$+0.5518$ $[+0.3811, +0.7240]$
	64	$+23.7$ $[+17.3, +30.2]$	$+1.0051$ $[+0.6748, +1.3287]$

Table 8: **EvoState w/o coarse minus EvoState w/ recomputed coarse.** Matched four-channel NRMSE on 336 Valid origins; each value is the paired mean followed by a 95% interval from resampling 28 physical files. Relative values average window-wise ratios; absolute values are in NRMSE units. Positive values favor the reference variant.

Epoch	h	Relative change (%) [95% CI]	Absolute change [95% CI]
0.5	1	-13.3 $[-14.8, -11.7]$	-0.0179 $[-0.0223, -0.0141]$
	8	$+1.1$ $[-1.2, +3.5]$	$+0.0260$ $[-0.0034, +0.0578]$
	16	$+7.3$ $[+5.2, +9.6]$	$+0.0701$ $[+0.0132, +0.1286]$
	32	$+13.0$ $[+9.9, +16.0]$	$+0.0068$ $[-0.1019, +0.1085]$
	64	$+8.4$ $[+3.7, +12.9]$	-0.8771 $[-1.2795, -0.5105]$
1	1	-0.7 $[-3.6, +2.1]$	-0.0090 $[-0.0163, -0.0029]$
	8	$+11.5$ $[+9.0, +13.8]$	$+0.0408$ $[+0.0014, +0.0723]$
	16	$+20.4$ $[+17.1, +23.5]$	$+0.1206$ $[+0.0508, +0.1855]$
	32	$+26.5$ $[+22.3, +30.3]$	$+0.2366$ $[+0.1154, +0.3534]$
	64	$+25.3$ $[+20.3, +30.0]$	$+0.3158$ $[+0.0886, +0.5449]$
1.5	1	$+13.1$ $[+8.2, +17.6]$	$+0.0004$ $[-0.0068, +0.0069]$
	8	$+19.0$ $[+12.5, +26.1]$	$+0.0157$ $[-0.0148, +0.0458]$
	16	$+20.2$ $[+14.1, +27.1]$	$+0.0835$ $[+0.0248, +0.1473]$
	32	$+20.7$ $[+17.5, +24.0]$	$+0.3263$ $[+0.1719, +0.4938]$
	64	$+27.4$ $[+23.5, +31.2]$	$+0.8993$ $[+0.5264, +1.2929]$
2	1	-17.4 $[-19.6, -15.3]$	-0.0303 $[-0.0382, -0.0228]$
	8	-6.2 $[-8.1, -4.3]$	-0.0551 $[-0.0786, -0.0360]$
	16	-1.7 $[-4.2, +0.9]$	$+0.0007$ $[-0.0312, +0.0316]$
	32	$+5.4$ $[+2.3, +8.8]$	$+0.2348$ $[+0.1420, +0.3356]$
	64	$+5.5$ $[-0.1, +11.4]$	$+0.7054$ $[+0.4452, +0.9728]$

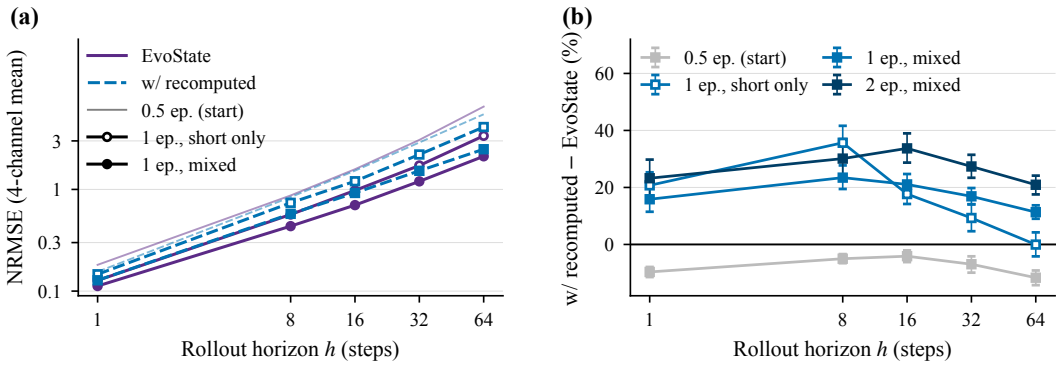

Figure 10: **Short-horizon control.** (a) Variant means at the half-epoch start (thin), after one epoch with short horizons only (open), and after one epoch with the mixed curriculum (filled). (b) Paired change of EvoState w/ recomputed coarse against EvoState with 28-file intervals. Positive values favor EvoState; negative values mean the recomputed variant has lower error, as at the half-epoch start.

Table 9: Curriculum control at one epoch, short horizons only: variant means and EvoState w/ recomputed coarse minus EvoState on 336 Valid windows with 28-file 95% intervals.

h	EvoState	w/ recomputed	Relative change (%)	Absolute change
1	0.127	0.146	$+20.7$ $[+16.2, +25.3]$	$+0.0190$ $[+0.0137, +0.0245]$
8	0.567	0.738	$+35.7$ $[+30.0, +41.6]$	$+0.1715$ $[+0.1361, +0.2111]$
16	0.977	1.203	$+17.7$ $[+14.2, +21.1]$	$+0.2260$ $[+0.1585, +0.3073]$
32	1.718	2.205	$+9.3$ $[+4.6, +14.1]$	$+0.4866$ $[+0.3197, +0.6724]$
64	3.367	4.107	-0.0 $[-4.2, +4.2]$	$+0.7402$ $[+0.4576, +1.0682]$

for EvoState w/o coarse. The u_x curves nearly coincide. At long horizons all three variants stay below Persistence for tracer but above it for pressure (3.46 against 2.08 at $h = 64$ for EvoState) and slightly for u_x. The controlled recipe trains for two epochs by pure regression; the full model's Test channel medians at $h = 64$ stay below Persistence in all four channels (Appendix D.1). In the tracer spectra (Fig. 12), the lowest band is nearly identical across rules. EvoState keeps the most middle-band coherence at $h = 16$ and 32 (0.46 and 0.44, against at most 0.43 and 0.38 for the alternatives) and the most high-band coherence at every horizon after the first.

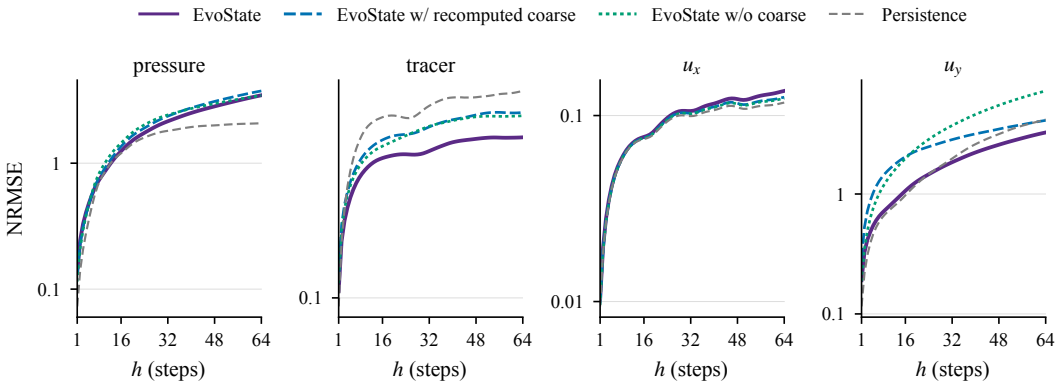

Figure 11: **Per-channel NRMSE of the three variants at two epochs** on the 336-window Valid evaluation set, with Persistence (gray dashed) for reference.

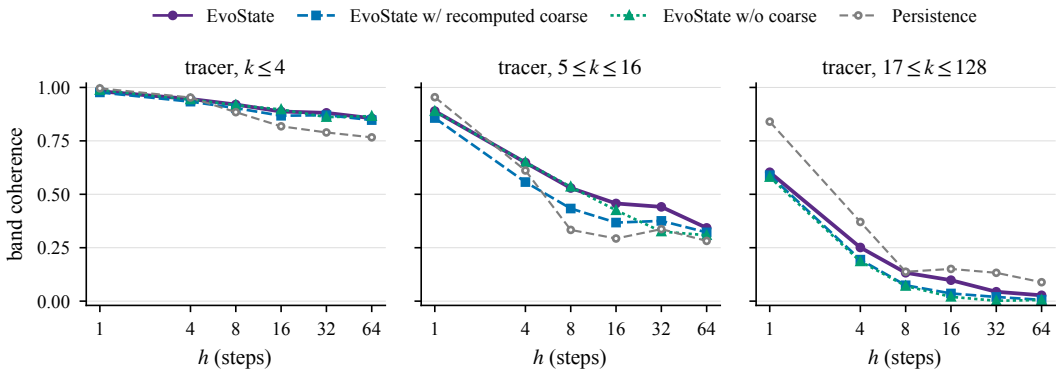

Figure 12: **Tracer band coherence of the three variants at two epochs**, mean over the 336 Valid windows, with Persistence (open circles).

D Additional Results

D.1 Test Rollout Profile

Fig. 13 extends the official Test rollout to 64 steps. The median channel-mean VRMSE of EvoState stays near 0.4 after $h = 16$, while Persistence approaches 1, the error of a spatial-mean prediction. At $h = 64$ the medians are 0.408 (IQR 0.293–1.319) and 0.884 (0.660–2.728). Means sit above medians because nearly uniform pressure and u_y frames inflate VRMSE. Per channel, EvoState's $h = 64$ medians are 0.465, 0.315, 0.059, and 0.508 for pressure, tracer, u_x, and u_y, against 0.997, 0.464, 0.085, and 1.001 for Persistence.

Table 10 breaks the official Test scores down by channel. EvoState is below Persistence in every channel over both rollout intervals, with the largest reductions for pressure and u_y (61–66% and 64–76%). At one step, pressure is the exception: 0.0530 against 0.0340.

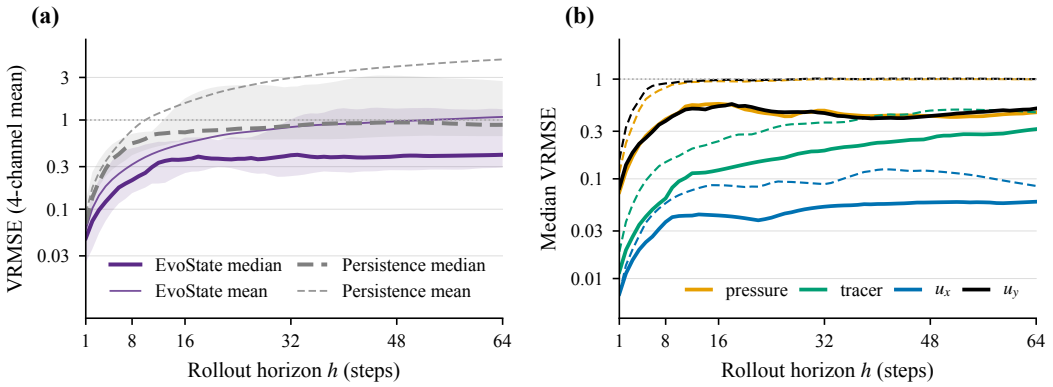

Figure 13: **Official Test rollout over 64 steps.** (a) Channel-mean VRMSE across 112 trajectories: median (thick), interquartile range (band), and mean (thin); the dotted line is VRMSE $= 1$. (b) Channel medians for EvoState (solid) and Persistence (dashed).

Table 10: Official Test VRMSE by channel (mean over 21,952 one-step windows or 112 rollouts, averaged over the stated horizons).

Interval	Model	Pressure	Tracer	u_x	u_y
One step	EvoState	0.0530	0.0458	0.0054	0.0620
	Persistence	0.0340	0.0826	0.0079	0.0828
6–12	EvoState	0.711	0.144	0.0616	0.479
	Persistence	1.841	0.229	0.0931	1.342
13–30	EvoState	1.522	0.274	0.0812	0.752
	Persistence	4.517	0.393	0.151	3.179

D.2 Per-condition results

Fig. 5 divides EvoState's Test VRMSE by that of Persistence for each Reynolds–Schmidt condition. The ratio is below one in all 28 conditions at $h = 8$, 32, and 64, with medians 0.50, 0.34, and 0.41. The smallest gain occurs at $\mathrm{Sc} = 0.1$ with $\mathrm{Re} = 5 \times 10^5$, where the ratio is 0.98 at $h = 32$.

Tables 11 and 12 report four-channel VRMSE for the twelve pre-specified Test conditions. Gray values in these tables mark truth spatial standard deviation below 0.01, the VRMSE denominator. The corresponding Valid rollouts are reported separately as NRMSE in Table 13; its gray values instead mark truth RMS below 0.01. All thresholds use the matching split, condition, and horizon. Test scales were calculated from the retained full-precision field export.

D.3 Token-grid sweep at other training sizes

Fig. 14 repeats the Valid evaluation set comparison with the Walrus-SF checkpoints at all three training budgets. Additional training does not close the long-horizon gap: from 8,192 to 24,576 windows, the $h = 64$ NRMSE of SF16 rises from 0.926 to 1.161, of SF32 from 0.839 to 1.408, and of SF64 from 1.242 to 2.412. EvoState's 0.608 is below all nine checkpoints at $h = 16$, 32, and 64. The foundation-model scores in Table 1 are the shear-flow entries of Walrus Table 1 (McCabe et al., 2025), which reports median VRMSE after fine-tuning. Parameter counts follow the original releases of MPP-AViT-L (McCabe et al., 2024), Poseidon-L (Herde et al., 2024), and DPOT-H (Hao et al., 2024); Walrus Table 4 lists the evaluated variants as 407M, 628M, and 1.15B. EvoState's one-step median comes from all 21,952 Test windows, and its interval medians from the 112 Test rollouts.

Table 14 lists the values behind Figs. 4 and 14.

Table 11: Four-channel VRMSE on the twelve pre-specified official Test rollouts at $h = 1$ and $h = 4$. Gray entries have truth spatial standard deviation below 0.01 at the indicated horizon.

		$h = 1$				$h = 4$			
Re	Sc	p	tracer	u_x	u_y	p	tracer	u_x	u_y
10^4	0.1	0.03	0.02	0.01	0.04	0.11	0.05	0.02	0.12
10^4	1	0.18	0.01	0.00	0.10	1.21	0.03	0.02	0.45
10^4	10	0.30	0.03	0.02	0.21	0.92	0.08	0.05	0.56
5×10^4	0.1	0.14	0.03	0.03	0.13	0.45	0.12	0.11	0.41
5×10^4	1	0.01	0.01	0.00	0.03	0.04	0.02	0.01	0.08
5×10^4	10	0.04	0.01	0.01	0.04	0.15	0.03	0.02	0.13
10^5	0.1	0.04	0.01	0.01	0.04	0.17	0.03	0.02	0.12
10^5	1	0.02	0.00	0.00	0.03	0.06	0.01	0.01	0.07
10^5	10	0.21	0.03	0.02	0.18	0.71	0.07	0.05	0.56
5×10^5	0.1	0.19	0.01	0.00	0.10	1.21	0.03	0.02	0.46
5×10^5	1	0.40	0.06	0.05	0.42	0.69	0.26	0.19	0.65
5×10^5	10	0.18	0.02	0.01	0.10	0.98	0.07	0.03	0.43

Table 12: Four-channel VRMSE on the twelve pre-specified official Test rollouts at $h = 16$ and $h = 64$. Gray entries have truth spatial standard deviation below 0.01 at the indicated horizon.

		$h = 16$				$h = 64$			
Re	Sc	p	tracer	u_x	u_y	p	tracer	u_x	u_y
10^4	0.1	0.16	0.21	0.10	0.21	0.28	1.64	0.11	0.28
10^4	1	5.81	0.05	0.02	1.95	6.00	0.06	0.02	8.58
10^4	10	4.34	0.13	0.07	1.17	8.12	0.21	0.07	3.76
5×10^4	0.1	0.27	0.39	0.08	0.27	0.33	0.34	0.08	0.32
5×10^4	1	0.09	0.05	0.02	0.10	0.25	0.37	0.05	0.31
5×10^4	10	0.25	0.23	0.10	0.24	0.37	0.73	0.09	0.39
10^5	0.1	0.35	0.29	0.06	0.31	0.27	0.23	0.04	0.29
10^5	1	0.12	0.05	0.02	0.11	0.13	0.40	0.03	0.12
10^5	10	1.31	0.12	0.04	1.31	3.49	0.25	0.03	3.14
5×10^5	0.1	5.86	0.06	0.02	1.81	6.08	0.07	0.02	7.07
5×10^5	1	0.82	0.78	0.32	0.80	1.71	0.99	0.80	1.71
5×10^5	10	4.80	0.11	0.04	1.28	5.88	0.27	0.03	5.57

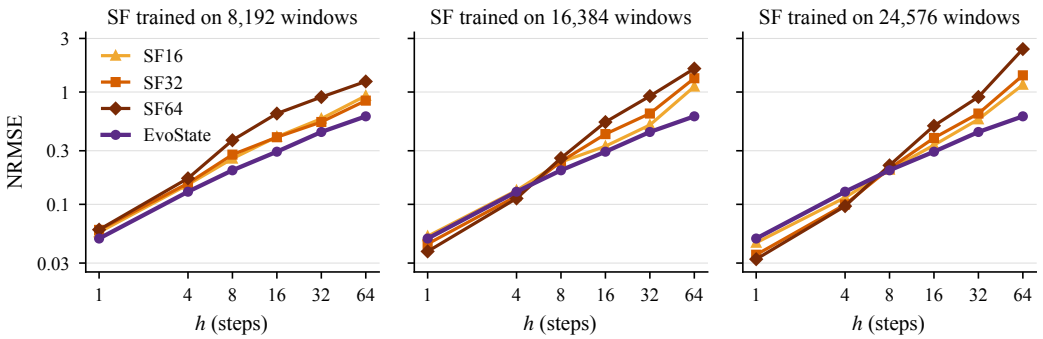

Figure 14: **Walrus-SF at three training budgets on the common Valid panel.** EvoState (purple) is the same curve in every panel.

Table 13: Per-channel NRMSE of the twelve pre-specified Valid rollouts. Gray entries have a truth RMS below 0.01 at the indicated horizon (nearly uniform field), where NRMSE is dominated by its denominator.

		$h = 16$				$h = 64$			
Re	Sc	p	tracer	u_x	u_y	p	tracer	u_x	u_y
10^4	0.1	4.33	0.11	0.07	1.17	8.12	0.13	0.07	3.76
10^4	1	4.73	0.09	0.03	1.35	5.81	0.11	0.04	7.14
10^4	10	0.40	0.51	0.21	0.47	0.16	0.73	0.09	0.20
5×10^4	0.1	0.53	0.04	0.01	0.55	0.84	0.08	0.02	0.79
5×10^4	1	0.32	0.66	0.15	0.33	0.30	1.13	0.15	0.45
5×10^4	10	1.19	0.06	0.02	1.20	3.72	0.12	0.02	3.73
10^5	0.1	0.32	0.61	0.15	0.33	0.33	1.14	0.15	0.46
10^5	1	0.32	0.67	0.15	0.33	0.32	1.04	0.14	0.43
10^5	10	0.32	0.69	0.15	0.33	0.32	1.02	0.14	0.43
5×10^5	0.1	0.22	0.26	0.12	0.26	0.13	0.73	0.09	0.16
5×10^5	1	4.80	0.11	0.04	1.28	5.88	0.24	0.03	5.57
5×10^5	10	0.22	0.21	0.08	0.22	0.22	0.69	0.07	0.25

Table 14: Four-channel mean NRMSE on the common 336-window Valid evaluation set for the full EvoState, the nine Walrus-SF checkpoints (grid, training windows), and Persistence.

Model	Params	$h = 1$	4	8	16	32	64
EvoState	11.2M	0.050	0.130	0.201	0.294	0.441	0.608
SF16, 8,192	140M	0.056	0.150	0.255	0.399	0.579	0.926
SF32, 8,192	140M	0.059	0.154	0.277	0.396	0.541	0.839
SF64, 8,192	140M	0.060	0.170	0.372	0.643	0.902	1.242
SF16, 16,384	140M	0.052	0.133	0.238	0.329	0.509	1.118
SF32, 16,384	140M	0.044	0.118	0.241	0.422	0.643	1.326
SF64, 16,384	140M	0.038	0.113	0.258	0.540	0.919	1.618
SF16, 24,576	140M	0.046	0.115	0.203	0.337	0.572	1.161
SF32, 24,576	140M	0.036	0.100	0.209	0.390	0.643	1.408
SF64, 24,576	140M	0.032	0.096	0.222	0.500	0.904	2.412
Persistence	–	0.073	0.233	0.419	0.660	1.077	1.712

D.4 Coarse-state replacement

The probe runs the trained EvoState transition on the Valid evaluation set twice from the same history. The first run carries Middle and Coarse as trained; the second overwrites them after every step with $f_{\mathrm{m}}(\hat{z}^{\mathrm{f}})$ and $f_{\mathrm{c}}(f_{\mathrm{m}}(\hat{z}^{\mathrm{f}}))$, computed by the frozen Stage A encoders. Fig. 7a compares the carried states with the recomputed ones against the encodings of the true frames, and Fig. 7b compares the two rollouts. Table 15 gives the paired changes. The recomputed Coarse state is 60–75% closer to the truth at every horizon, and the rollout that uses it is worse from $h = 4$ onward, with the gap widening with horizon.

Table 15: Paired relative change (%) from replacing carried Middle/Coarse with the quantity derived from Fine, mean over 336 windows with 28-file bootstrap 95% intervals. Negative latent change means the recomputed state is closer to the true-frame encoding; positive NRMSE change means a worse rollout.

h	Middle latent MSE	Coarse latent MSE	Rollout NRMSE
1	-44.9 $[-49.8, -39.4]$	-60.5 $[-66.3, -54.4]$	0 (identical)
4	-50.1 $[-53.4, -46.4]$	-72.2 $[-75.0, -69.3]$	$+4.6$ $[3.7, 5.6]$
8	-47.3 $[-50.9, -43.3]$	-74.4 $[-77.0, -71.6]$	$+9.1$ $[7.4, 10.8]$
16	-43.4 $[-47.5, -39.1]$	-75.3 $[-78.2, -72.2]$	$+15.9$ $[13.3, 18.9]$
32	-38.9 $[-43.4, -34.5]$	-74.8 $[-78.3, -71.0]$	$+33.6$ $[26.5, 42.2]$
64	-31.3 $[-35.9, -26.7]$	-70.0 $[-74.3, -65.7]$	$+112.6$ $[76.6, 154.2]$

D.5 SPECTRAL DIAGNOSTICS

Fig. 15 resolves the Valid evaluation set forecasts by spatial scale for tracer and both velocity components. Coherence measures phase agreement within a band, and the energy ratio measures how much of the true band energy the forecast carries. In every channel the lowest band keeps high coherence through $h = 64$ (0.90 for tracer, above 0.97 for u_x), while the highest band loses coherence first. EvoState's high-band energy falls below the truth for tracer (0.29 at $h = 64$); Persistence keeps the initial frame's fine-scale energy while the true field's decays, so its high-band ratio rises toward 2 despite low coherence. The radial spectra show the same pattern: the forecast follows the true spectrum at large scales and drops below it at the smallest ones.

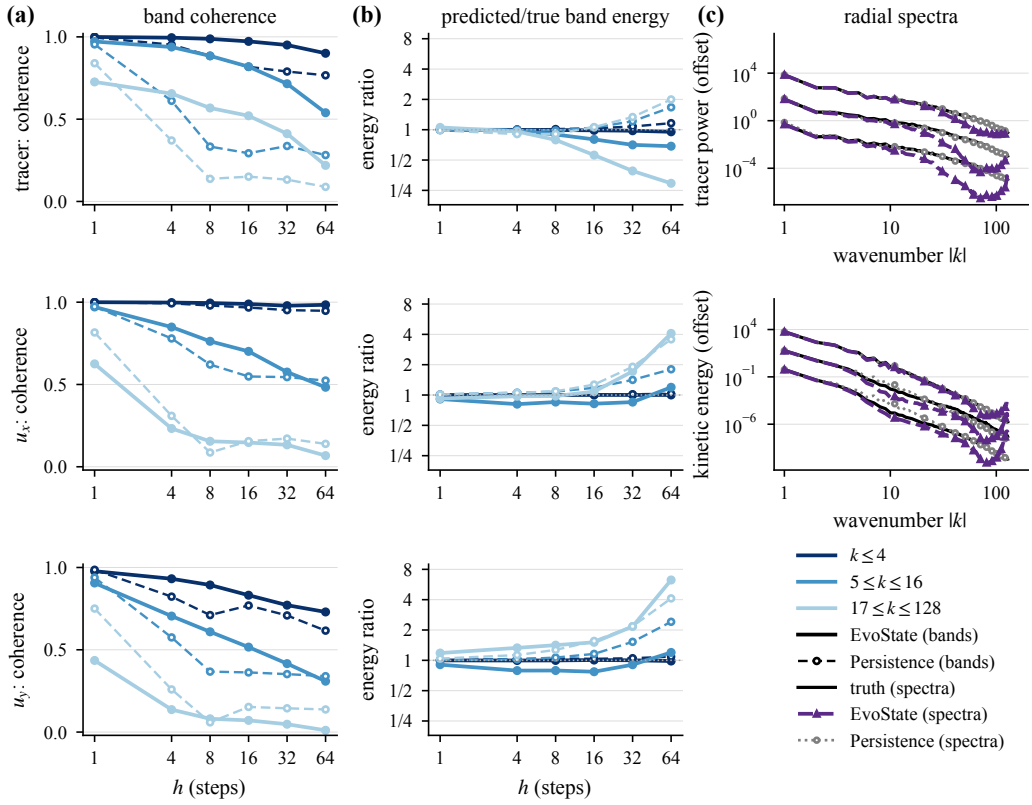

Figure 15: **Scale-resolved diagnostics on the 336-window Valid evaluation set.** Rows show tracer, u_x, and u_y; columns show band coherence, predicted-to-true energy ratio, and radial spectra.

E QUALITATIVE RESULTS

The displayed fields use the official Test trajectory 0 and origin 0 of each selected condition, fixed before predictions were viewed. Four observed frames initialize each autonomous 64-step rollout. The plots embed native 256×512 pixels. Ground Truth and EvoState share a symmetric color scale within each field pair; $|error|$ uses black at zero.

E.1 FOUR-CHANNEL FIELDS

Fig. 16 pairs truth and EvoState for all four channels at $\mathrm{Re} = 10^5$, $\mathrm{Sc} = 1$. The broad velocity bands and tracer interfaces remain recognizable through $h = 64$. The fine vortex contours soften as the rollout proceeds; the tracer VRMSE reaches 0.40 at $h = 64$ (Table 12), while the u_x VRMSE remains 0.03.

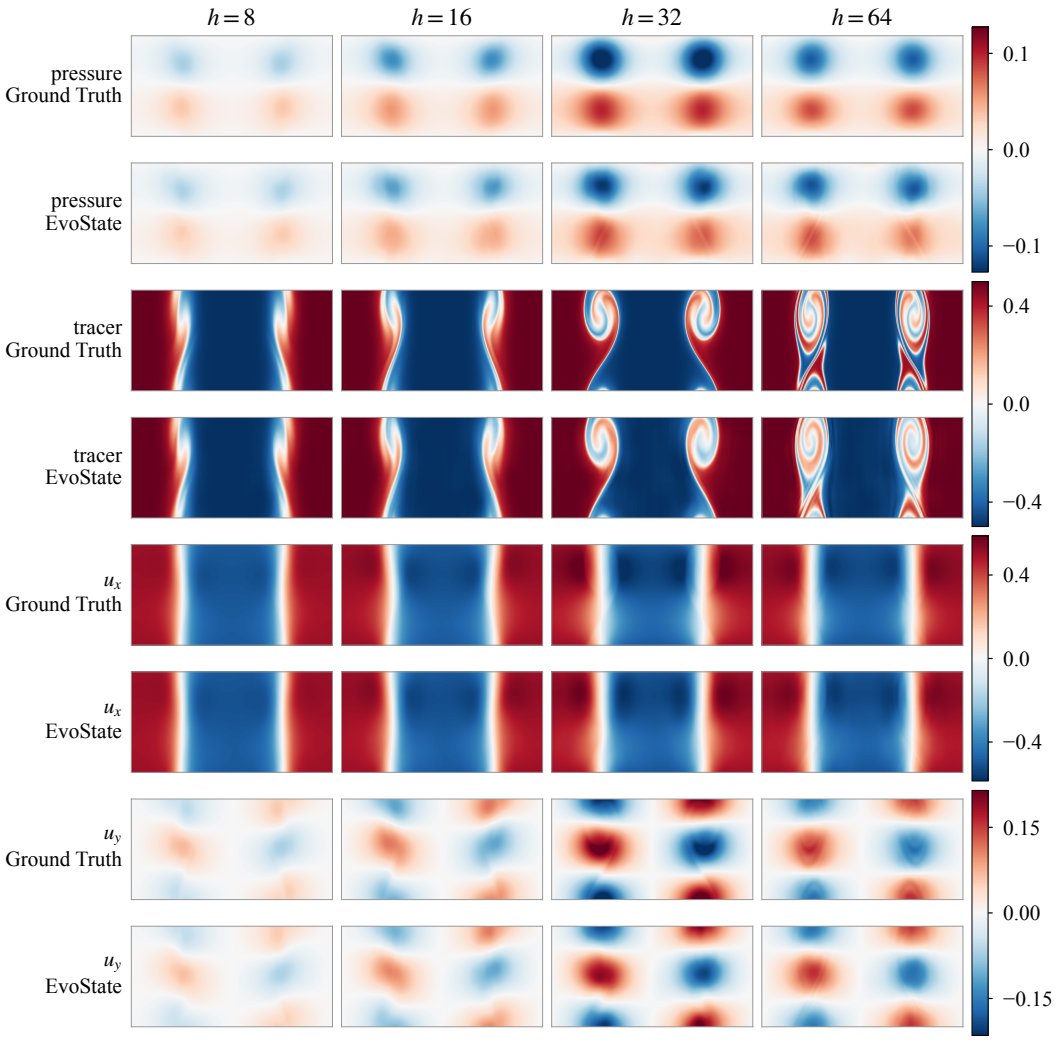

Figure 16: **Four-channel Test fields at** $\mathrm{Re} = 10^5$**,** $\mathrm{Sc} = 1$ **and** $h = 8, 16, 32, 64$**.**

E.2 Test trajectories by Reynolds number

At $\mathrm{Sc} = 10$, Figs. 8 and 17 show the three selected Reynolds regimes. The medium-Reynolds forecast follows the main rolled-up tracer regions through $h = 32$; at $h = 64$ the interfaces shift and fine filaments are less distinct. The low- and high-Reynolds examples keep their broad band layout, with absolute error concentrated around displaced interfaces. These examples illustrate the phase and detail errors behind the pointwise VRMSE in Table 12.

At $h = 32$, tracer VRMSE is 0.18, 0.52, and 0.18 for the low-, medium-, and high-Reynolds examples, respectively. By $h = 64$ it reaches 0.21, 0.73, and 0.27 (Table 12). The medium-Reynolds example has the largest increase because its predicted interface positions diverge as the vortices develop. In the other two examples, broad bands still align, while the $|error|$ rows show localized phase errors along thinner folds. At high Reynolds number the forecast also smooths fine tracer filaments that remain visible in Ground Truth. Thus the smaller aggregate VRMSE there coexists with a loss of local detail, which the paired field panels expose.

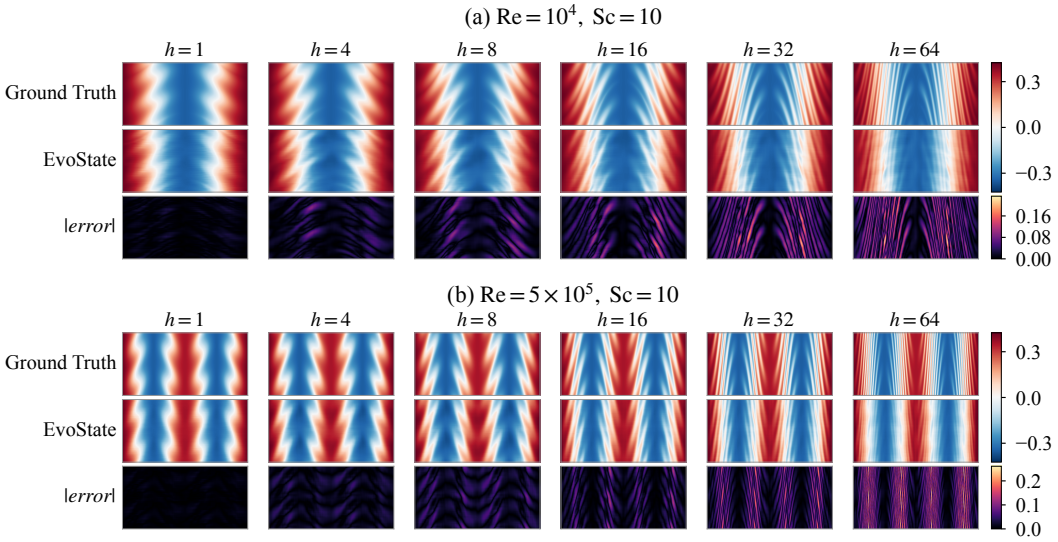

Figure 17: **Test tracer rollouts at two Reynolds numbers and** $\mathrm{Sc} = 10$**.**

E.3 TRACER ACROSS CONDITIONS

Figs. 18 and 19 cover the twelve pre-specified Test conditions. At $h = 32$, large tracer structures remain in most predictions, while thin interfaces already differ where the truth rolls into compact vortices.

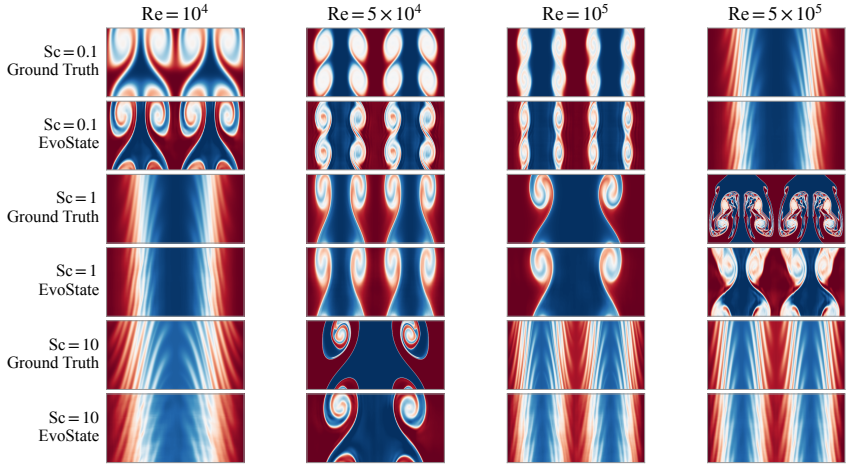

Figure 18: **Test tracer fields in twelve conditions at** $h = 32$**.**

By $h = 64$, the separation is clearest at $\mathrm{Re} = 5 \times 10^5$, $\mathrm{Sc} = 1$: the forecast retains the broad vortex layout but misses part of the internal fine structure. Its tracer VRMSE is 0.99, compared with 0.07 and 0.27 at the same Reynolds number for $\mathrm{Sc} = 0.1$ and 10 (Table 12).

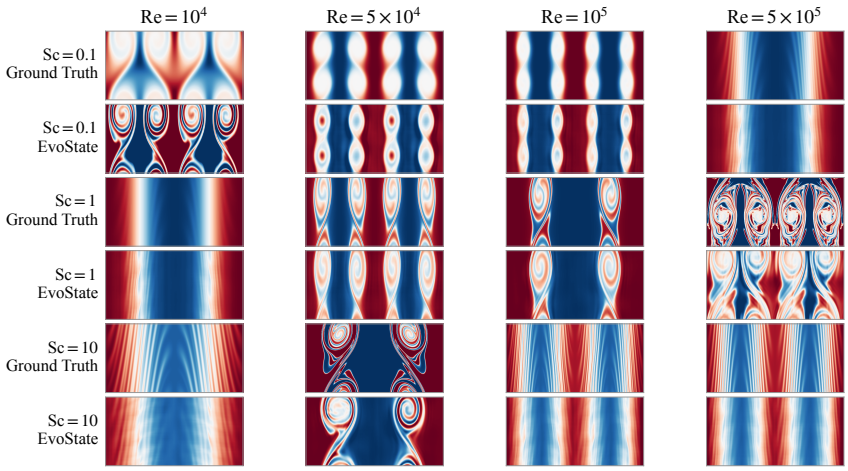

Figure 19: **Test tracer fields in the same twelve conditions at $h = 64$.**