UNIDYNAMIC: UNIFIED SURROGATE MODELING FOR PDE AND SPDE VIA FLOW MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretraining shared models across deterministic and stochastic PDEs requires accommodating both diverse dynamics and different levels of predictive uncertainty. With sufficient observations, deterministic dynamics yield a unique continuation, whereas unobserved future stochastic forcing induces a conditional distribution of possible futures. A controlled Kolmogorov-flow experiment reveals complementary strengths of point prediction for deterministic accuracy and conditional generation for stochastic distribution matching. We therefore introduce UniDynamic, a flow-matching framework that jointly learns both modes across deterministic and stochastic PDEs. A shared backbone, conditioned on multiscale windowed log-signature history representations, supports one-pass prediction and iterative sampling. We construct a paired benchmark with repeated-future DNS ensembles for evaluating conditional moments, marginal distributions, and spatial fluctuations. UniDynamic achieves competitive accuracy across 11 deterministic PDE subsets. On the three stochastic PDEs, it attains the lowest lead-20 marginal energy distance among the compared methods. Furthermore, pretraining improves few-shot performance on unseen equations, while adding deterministic training data improves marginal-distribution accuracy on stochastic Kolmogorov flow and reaction–diffusion. These results support a unified flow-matching framework that combines accurate point prediction, conditional generation, and transfer across equations and regimes.

1 INTRODUCTION

Neural operators provide efficient approximations to PDE solution maps, evolving from task-specific architectures (Li et al., 2020; Lu et al., 2021; McCabe et al., 2023) toward models jointly pretrained across diverse PDE families (Hao et al., 2024; Chen et al., 2024; Herde et al., 2024; Zhang et al., 2025). Extending this pretraining paradigm to both deterministic and stochastic PDEs requires accounting for predictive uncertainty as well as diversity in the governing dynamics. However, deterministic point predictors do not explicitly represent the conditional variability arising from unobserved stochastic forcing.

To examine the modeling implications of this conditional variability, we compare point prediction and conditional generation in a controlled study of two-dimensional Kolmogorov flow. We keep the drift fixed across deterministic and stochastic regimes, while future stochastic forcing remains unobserved by the models. With comparable capacity and matched training updates, the point-prediction model DPOT (Hao et al., 2024) achieves lower deterministic rollout error, whereas its conditional flow-matching counterpart better matches the marginal distributions in the stochastic regime. These complementary findings motivate combining next-state supervision with conditional flow matching in a shared model. This poses two challenges: reconciling accurate, source-independent point prediction with faithful conditional generation, and learning compact history representations that retain temporal order and multiscale dynamics for both regimes.

To address these challenges, we introduce UNIDYNAMIC, a history-conditioned flow-matching framework for unified deterministic and stochastic forecasting. A common channel interface accommodates heterogeneous state variables, while a multiscale windowed log-signature encoder combines temporal increments, order-sensitive interactions, and last-frame features into shared history representations. In the gated formulation, a source gate and an explicit velocity skip yield a

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.

source-independent point prediction with one Euler step, while multistep flow integration generates conditional samples with the same weights. We jointly train both modes using velocity and next-state supervision. Both modes predict one frame at a time, and each ensemble member updates its own history during autoregressive rollout.

We evaluate the framework on 11 deterministic PDE subsets and 3 stochastic systems. Repeated-future direct numerical simulation (DNS) ensembles provide references for evaluating conditional moments, marginal distributions, and spatial fluctuations alongside pointwise accuracy. For stochastic forecasting, our baselines include deterministic predictors, probabilistic retrofits, and diffusion- and flow-matching models. UNIDYNAMIC achieves comparable deterministic forecasting accuracy and the lowest observed marginal energy distance at the 20-step forecast horizon on all three stochastic systems. Few-shot experiments support transfer to unseen equations, while a separate cross-regime study shows that joint deterministic–stochastic training improves marginal–distribution accuracy on Kolmogorov flow and reaction–diffusion. We summarize our contributions as follows.

- **Unified deterministic and stochastic forecasting.** We unify autoregressive point prediction and conditional generation in a shared backbone, supporting heterogeneous state variables, joint pretraining and source-independent one-step forecasts in the gated formulation.
- **A paired benchmark for dynamics and uncertainty.** We construct paired datasets for Kolmogorov flow, reaction–diffusion, and damped waves, combining shared drift within each family, matched histories, controlled forcing and repeated-future DNS ensembles for uncertainty evaluation.
- **Uncertainty modeling and cross-regime transfer.** Experiments show that our conditional-generation formulation better captures stochastic spread and marginal distributions than deterministic prediction, while pretraining supports few-shot transfer to unseen equations.

2 UNIFIED PDE FORECASTING

Forecasting spatiotemporal PDE dynamics. Consider a C-channel state field $u(\mathbf{x}, t) \in \mathbb{R}^C$ on $\Omega \subset \mathbb{R}^d$, governed by $\partial_t u = \mathcal{F}[u; \eta]$ with prescribed initial and boundary conditions. Here $\mathcal{F}$ denotes the governing operator and η collects the physical coefficients. Neither is supplied to the model. Let u_s denote the spatially discretized field at the s-th recorded time. We pair each valid T-frame history with the next frame: $H_s = (u_{s-T+1}, \ldots, u_s), \quad Y_s = u_{s+1}$. The model predicts Y_s from H_s and updates the history with its prediction for autoregressive rollout. Fourier neural operators and DeepONets are standard architectures for learning deterministic solution maps that return a single solution prediction per input (Li et al., 2020; Lu et al., 2021). We refer to this forecasting mode as *deterministic point prediction.*

Pretraining across PDE families. Multi-PDE pretraining shares model parameters across PDEs with different equations, coefficients, or state variables, as in DPOT, OmniArch, Poseidon, OmniFluids, and Walrus (McCabe et al., 2023; Hao et al., 2024; Chen et al., 2024; Herde et al., 2024; Zhang et al., 2025; McCabe et al., 2025). Related approaches use mixture-of-experts, heterogeneous-field transfer, or operator splitting (Wang et al., 2025; Sun et al., 2026a; Rautela et al., 2025; Sun et al., 2026b). For dataset distributions $\mathcal{D}_k$ over (H, Y), shared parameters are learned by $\min_\theta \sum_{k=1}^K \pi_k \mathbb{E}_{(H,Y)\sim\mathcal{D}_k}[\mathcal{L}_\theta^{(k)}(H, Y)]$, with $\pi_k \geq 0$ and $\sum_k \pi_k = 1$. The dataset index k is not explicitly provided to the model.

Point prediction and conditional generation. Uncertainty in PDE forecasting can arise from partial observations, unresolved scales, or stochastic forcing. We focus on a representative setting where future stochastic forcing realizations are unobserved by the model. The learning target is $\mathsf{P}_k(\cdot \mid H)$, the conditional distribution of the next state Y given the observed history H. When H uniquely determines Y, this distribution reduces to a point mass at that state. Existing approaches include diffusion-based ACDM and GenCFD (Kohl et al., 2023; Molinaro et al., 2024), latent flow-matching methods (Li et al., 2025; Chen & Deng, 2026), and probabilistic retrofitting or generative fine-tuning of point predictors (Diaconu et al., 2026; Couairon et al., 2026). Latent methods may incur autoencoder reconstruction error, while marginal CRPS does not determine joint spatial dependence. We study a shared model supporting both *deterministic point prediction* and *conditional distribution generation* across deterministic and stochastic PDEs.

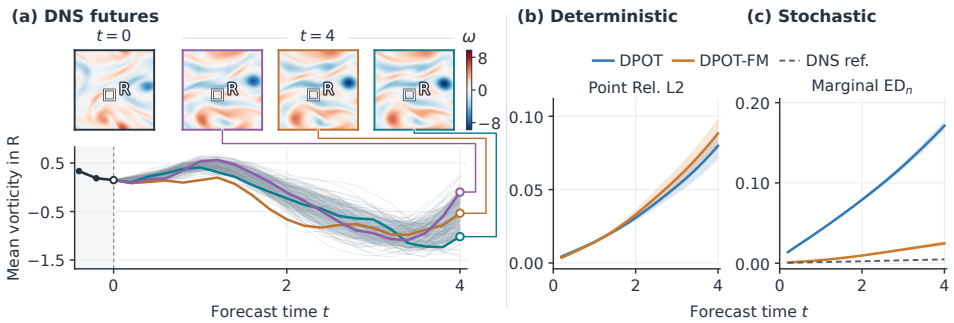

Figure 1: **Point accuracy and conditional variability.** (a) Same-history DNS futures (10–90% band). (b) Deterministic point error. DPOT-FM uses its ensemble mean. (c) Stochastic marginal ED_n. Bands in (b)–(c): standard deviation over 3 seeds. Dashed: independent DNS–DNS reference.

3 Motivating Example

We use two-dimensional Kolmogorov flow with shared drift across deterministic and stochastic regimes. The models take ten 128×128 vorticity frames as input and forecast the next 20 time steps. Each stochastic history is paired with 128 DNS continuations from the same state at its final observed time, under independent realizations of the unobserved future forcing. We use one continuation per deterministic history. We compare the deterministic field predictor DPOT with its conditional flow-matching counterpart, DPOT-FM, matching model capacity and training-update counts. Both models are trained separately in each regime using three random seeds. DPOT-FM generates 128 trajectories using 32 Euler integration steps per forecast frame.

Figure 1(a) shows distinct DNS trajectories and future fields conditioned on the same history, illustrating conditional variability that a single field forecast cannot represent. In the stochastic regime, DPOT-FM's lower *marginal energy distance* (ED) relative to DPOT indicates a closer match to pointwise DNS distributions, not merely more diverse samples. In the deterministic regime, DPOT achieves a lower rollout-averaged point-prediction error even against DPOT-FM's 128-member ensemble mean (Figure 1(b,c)). Detailed protocols and additional results are provided in Appendix B. These results motivate a unified model that combines direct next-state supervision with conditional flow matching. The shared drift also provides a controlled setting for evaluating whether joint training improves cross-regime transfer (Section 6.4).

4 UniDynamic

In this section, we introduce UniDynamic, which combines field preprocessing, windowed log-signature history encoding, and a shared flow-time-conditioned diffusion Transformer (DiT) for one-step deterministic prediction and multi-step conditional generation (Figure 2).

4.1 Field Preprocessing

Given a T-frame history $H \in \mathbb{R}^{T \times N_x \times N_y \times C}$ on an $N_x \times N_y$ grid with C physical channels, we pad the channel dimension to $C_{\max}$. History-only reversible instance normalization (Kim et al., 2022) uses channelwise statistics over observed frames and spatial locations to produce the normalized history $\widetilde{H}$ and next-state target y. $\mathcal{D}_H$ denotes the inverse transformation. We perform flow matching directly in normalized field space, without requiring a separately trained variational autoencoder.

4.2 Encoding Dynamics with Windowed Log Signatures

After patch embedding, prior work combines frame features across time using learned temporal aggregation or attention (Hao et al., 2024; McCabe et al., 2025). We instead use multiscale windowed log signatures to summarize changes in each patch's features and capture their temporal order.

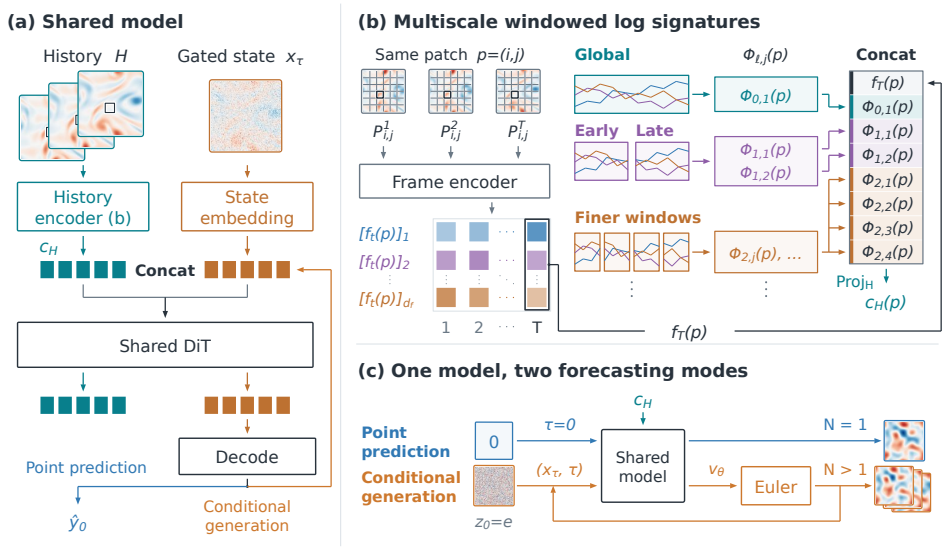

Figure 2: **UniDynamic architecture.** (a) Shared DiT with state-only decoding. (b) Two-scale log-signature history encoding. Finer windows are optional. (c) One-step point prediction and multi-step generation share weights. Outputs are illustrative, not predictions.

Encoding ordered changes. Let $f_1(p), \ldots, f_T(p)$ be the frame features at patch p. We group feature channels to limit the number of within-group interactions. We append normalized observation time $s_i = \frac{i-1}{T-1}$ to each group, forming a piecewise-linear path. Within a window $w = \{a, \ldots, b\}$, define the feature difference between consecutive frames as $d_i = x_{i+1} - x_i$, for $i = a, \ldots, b-1$. The log signature truncated at level two is given by (Chevyrev & Kormilitzin, 2016):

$$S_w = \sum_{i=a}^{b-1} d_i = x_b - x_a, \qquad A_w = \tfrac{1}{2} \sum_{a \le i < j < b} \left(d_i d_j^\top - d_j d_i^\top \right). \tag{1}$$

Here, S_w records the overall change between the window endpoints, whereas A_w contains signed-area terms that capture order-sensitive interactions between increments. These terms can distinguish paths with the same overall change but different orders of changes.

Combining the latest state with multiscale evolution. Since increments omit absolute feature values, we retain the last-frame feature $f_T(p)$ and summarize evolution over windows $w_{\ell,j}$, where ℓ indexes window-length scales and j indexes temporal positions within each scale. At $\ell = 0$, one window spans the full history. Larger ℓ correspond to shorter, overlapping windows (Morrill et al., 2020). All windows share observation-time coordinates normalized over the full history. For $w = w_{\ell,j}$, $\Phi_{\ell,j}(p)$ concatenates group-wise endpoint differences S_w and the strictly upper-triangular entries of A_w. The history-token content is $c_H(p) = \text{Proj}_H\left[f_T(p); \text{Concat}_{\ell,j}\, \Phi_{\ell,j}(p)\right]$. This yields one history token per spatial patch for point prediction and conditional generation.

4.3 Unified point prediction and conditional generation

Shared field network. The shared field network r_θ supports both forecasting regimes by jointly processing history tokens c_H and patch tokens of the current flow state. Since history receives no flow-matching noise, adaptive normalization uses a fixed time embedding for history tokens and an embedding of the current flow time for state tokens. The two token sequences are concatenated and processed by a shared DiT, where every token can attend to all history and state tokens. History and state tokens at matching spatial patches share the same rotary positional embedding (RoPE) coordinates. Only state tokens are decoded to the field grid.

From conditional generation to deterministic field prediction. Conditioned on c_H, we construct a velocity field v_θ from the shared network r_θ and train it by flow matching along a linear path from

Gaussian noise e to the normalized next-state target y (Liu et al., 2022):

$$z_\tau = (1-\tau)e + \tau y, \qquad e \sim \mathcal{N}(0, I), \quad \tau \sim \mathcal{U}[0,1],$$
$$\mathcal{L}_{\mathrm{FM}}(\tau) = \mathrm{MSE}(v_\theta(c_H, z_\tau, \tau), y - e)\,. \tag{2}$$

Here τ denotes flow time for generating a single next state, rather than physical rollout time. To connect velocity fitting with field accuracy, define the implied next-state estimate $\hat{y}_\tau = z_\tau + (1-\tau)v_\theta(c_H, z_\tau, \tau)$. Its reconstruction error satisfies the exact identity (Gagneux et al., 2026):

$$\mathrm{MSE}(\hat{y}_\tau, y) = (1-\tau)^2 \mathcal{L}_{\mathrm{FM}}(\tau), \qquad 0 \le \tau < 1. \tag{3}$$

Thus, the flow-matching objective is equivalently a reconstruction loss weighted by $(1-\tau)^{-2}$.

When the history uniquely determines the next state, $y = f(H)$, the final prediction should be independent of the Gaussian initialization. The ideal conditional velocity $v^\star(H, z, \tau) = (f(H) - z)/(1-\tau)$ is increasingly sensitive to perturbations in z as $\tau \to 1$. In practice, model approximation and numerical integration errors can leave spurious dependence on the initial noise and reduce deterministic field accuracy. We therefore construct a source-independent one-step prediction using the same network and supervise it directly.

Source cancellation and zero-time supervision. At $\tau = 0$, we set the network's state input to zero and introduce an explicit cancellation term:

$$v_\theta(c_H, z, \tau) = \begin{cases} r_\theta(c_H, 0, 0) - z, & \tau = 0, \\ r_\theta(c_H, z, \tau), & 0 < \tau < 1. \end{cases} \tag{4}$$

Since $z_0 = e$, a single Euler step gives $\hat{y}_0 = e + v_\theta(c_H, e, 0) = r_\theta(c_H, 0, 0)$. The prediction is therefore independent of the random source by construction.

We directly supervise this prediction with the zero-time flow-matching loss:

$$\mathcal{L}_{\mathrm{FM}}(0) = \mathrm{MSE}(v_\theta(c_H, e, 0), y - e) = \mathrm{MSE}(\hat{y}_0, y) \equiv \mathcal{L}_{\mathrm{pred}}. \tag{5}$$

At positive flow times, the same network retains its state input and learns the velocity field for conditional generation.

4.4 Training and inference

We train the shared model on history–target pairs (H, Y) from both deterministic and stochastic dynamical systems by minimizing:

$$\mathcal{L}_\theta = \lambda_{\mathrm{pred}}\, \mathbb{E}_{k,H,Y}[\mathcal{L}_{\mathrm{pred}} + \mathcal{L}_{\mathrm{rel}}(0)] + \lambda_{\mathrm{FM}}\, \mathbb{E}_{k,H,Y,e,\tau}[\mathcal{L}_{\mathrm{FM}}(\tau) + \mathcal{L}_{\mathrm{rel}}(\tau)]\,. \tag{6}$$

Here $\mathcal{L}_{\mathrm{rel}}(\tau)$ is the unsquared relative-L_2 error between the inverse-normalized prediction $\mathcal{D}_H(\hat{y}_\tau)$ and Y, averaged over valid physical channels. Both branches are evaluated on every batch, with nonnegative weights λ_{pred} and λ_{FM}. During autoregressive forecasting, each ensemble member feeds back its own predicted state rather than the ensemble mean.

5 A Multi-System Benchmark with Conditional DNS Ensembles

Scores such as CRPS can evaluate probabilistic forecasts against a single realized future, but this alone does not directly characterize the conditional variability. We construct a three-system benchmark with repeated DNS continuations from the same saved state and history under independent future stochastic forcing, enabling direct evaluation of conditional means, spread, marginal distributions, and spatial fluctuations throughout autoregressive rollout.

5.1 Repeated-future DNS reference ensembles

For each selected ten-frame reference history H, we generate M DNS continuations from the same saved high-resolution state at the final observed frame. Across continuations, we independently resample only the future stochastic forcing, whose realizations remain unobserved by the model. This procedure yields the reference ensemble $\mathcal{R}(H) = \{Y_{1:L}^{(m)}\}_{m=1}^{M}$, where each member contains L future frames. We construct 20- and 50-step reference ensembles, with larger ensembles for ensemble-size convergence checks.

5.2 PHYSICAL SYSTEMS AND FORCING REGIMES

The benchmark covers two-dimensional Kolmogorov flow, reaction–diffusion, and damped nonlinear waves (Figure 3). Within each system, we keep the drift fixed and vary the stochastic forcing strength across zero, low, mid, and high levels. For diagnostic comparisons, paired deterministic continuations reuse stochastic histories and saved states but disable only future stochastic forcing.

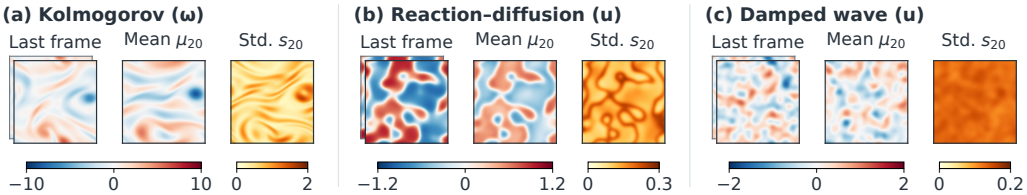

Figure 3: **Dataset preview.** For one reference history per system, we show the last observed frame and the DNS ensemble mean and standard deviation at lead 20 under mid-noise forcing. The displayed fields are Kolmogorov vorticity ω, reaction–diffusion activator u, and wave displacement u.

All physical state channels are observed. Zero- and mid-noise training pools are size-matched, while additional reference branches are excluded from training. Splits separate chains or initial conditions, keeping each trajectory and all its windows within the same split. Detailed dataset sizes, resolutions, and reference allocations are provided in Appendix D.

5.3 EVALUATING CONDITIONAL GENERATION OVER ROLLOUT

At each forecast step $\ell = 1, \ldots, L$, we compare model and DNS ensembles conditioned on the same initial history H. Each metric is computed per history and channel before averaging. Lower values indicate closer agreement with the DNS reference.

We measure errors in the ensemble mean and standard deviation relative to DNS as $E_\mu(\ell) = \|\hat{\mu}_\ell - \mu_\ell\|_2 / \|\mu_\ell\|_2$ and $E_s(\ell) = \|\hat{s}_\ell - s_\ell\|_2 / \|s_\ell\|_2$. For $s_\ell = 0$, we replace E_s with the absolute root-mean-square (RMS) predictive spread. To compare marginal distributions beyond their first two moments, we use pointwise Energy Distance (ED) following Pfister et al. (2026). For independent model draws X, X' and DNS draws Y, Y' at a fixed history, lead, channel, and grid point, its population form is

$$D_{\mathrm{ED}} = 2\,\mathbb{E}|X - Y| - \mathbb{E}|X - X'| - \mathbb{E}|Y - Y'|. \tag{7}$$

We spatially average this distance and normalize by the RMS of the DNS ensemble fields to obtain ED_n, which we then average over channels and histories.

To complement these marginal diagnostics, the relative fluctuation-spectrum error compares fluctuation power across spatial scales after removing each ensemble's mean field. For K model fields X_i and M DNS fields Y_j at a fixed history, lead, and channel, let $\mathcal{S}(f; k)$ denote shell-summed Fourier power, and define $S'_X(k) = K^{-1} \sum_i \mathcal{S}(X_i; k) - \mathcal{S}(\hat{\mu}_\ell; k)$ and $S'_Y(k) = M^{-1} \sum_j \mathcal{S}(Y_j; k) - \mathcal{S}(\mu_\ell; k)$. For stochastic references, we report

$$E_{\mathrm{fluc}}(\ell) = \frac{\sum_k |S'_X(k) - S'_Y(k)|}{\max\{\sum_k |S'_Y(k)|, 10^{-12}\}}. \tag{8}$$

Reference validation. We check DNS numerical accuracy using coupled-noise time-step refinement, spatial resolution, numerical-precision checks, and equation-specific diagnostics. Convergence checks and DNS–DNS comparisons assess finite-ensemble effects. Detailed protocols, metric estimators, and validation results are provided in Appendix D.

6 EXPERIMENTS

Datasets and evaluation. Deterministic evaluation covers 11 subsets from the FNO Navier–Stokes data, PDEBench, and PDEArena, totaling 5,654 held-out trajectories with dataset-specific rollout horizons of 4–91 steps. Our benchmark in Section 5 provides deterministic and stochastic versions of the three selected equation systems. The main stochastic comparison uses 20-step rollouts

Table 1: **Deterministic forecasting on 11 PDE datasets.** Rollout L2RE (↓). Boldface marks the lowest reported error in each column. DR: reaction–diffusion. SWE: shallow water. "–": unreported.

Method	Params (M)	FNO			PDEBench						PDEArena		Mean↓
		NS (ν)			CNS (Ma, $\eta = \zeta$)				DR	SWE	NS	NS-cond	
		10^{-5}	10^{-4}	10^{-3}	$(1, .1)$	$(1, .01)$	$(.1, .1)$	$(.1, .01)$					
Single-dataset models													
FNO	0.5	0.1560	0.0834	0.0128	0.0980	0.0960	0.3600	0.1700	0.1200	0.0044	0.0912	0.3190	0.13735
UNet	25	0.1980	0.1190	0.0245	0.3340	0.2910	0.5690	0.3570	0.0971	0.0521	0.1020	0.3370	0.22552
F-FNO	1.3	0.1210	0.0503	0.0099	0.0212	0.0520	0.1620	0.0452	0.0571	0.0116	0.0839	0.6020	0.11056
GK-T	1.6	0.1340	0.0792	0.0098	0.0341	0.0377	0.0274	0.0366	0.0359	0.0069	0.0952	0.4230	0.08362
GNOT	1.8	0.1570	0.0443	0.0125	0.0325	0.0420	0.0228	0.0341	0.0311	0.0068	0.1720	0.3250	0.08001
OFormer	1.9	0.1710	0.0645	0.0104	0.0417	0.0625	0.0254	0.0205	0.0192	0.0072	0.1350	0.3320	0.08085
Pretrained models													
MPP-S	30	–	–	–	–	–	–	–	**0.0112**	**0.0024**	–	–	–
DPOT-S	30	0.0553	0.0442	0.0131	0.0153	0.0337	0.0119	0.0188	0.0379	0.0066	0.0991	0.3160	0.05926
MoE-POT-S	90	0.0527	0.0417	0.0126	0.0143	**0.0323**	0.0113	0.0181	0.0354	0.0063	0.1060	0.2980	0.05715
UNIDYNAMIC	30.7	**0.0212**	**0.0162**	**0.0024**	**0.0090**	0.0348	**0.0090**	**0.0161**	0.0430	0.0053	**0.0597**	**0.2779**	**0.04496**

against 128 same-history DNS futures. Models in our evaluations forecast autoregressively from ten observed frames. We report full-horizon relative L_2 error for deterministic forecasting, averaged equally across subsets, and the distributional metrics in Section 5.3 for stochastic forecasting.

Baselines. For deterministic forecasting, we compare with single-dataset models FNO, UNet, F-FNO, GK-T, GNOT, and OFormer (Li et al., 2020; Ronneberger et al., 2015; Tran et al., 2023; Cao, 2021; Hao et al., 2023; Li et al., 2023), and pretrained MPP-S, DPOT-S, and MoE-POT-S (McCabe et al., 2023; Hao et al., 2024; Wang et al., 2025). For stochastic forecasting, we compare FNO/DPOT point predictors, their CRPS-retrofitted variants, and DPOT-FM, ACDM, and GenCFD (Hao et al., 2024; Lipman et al., 2022; Kohl et al., 2023; Molinaro et al., 2024) on the same DNS references. FNO-CRPS and DPOT-CRPS follow Diaconu et al. (2026), adding noise conditioning and CRPS fine-tuning.

6.1 Deterministic PDE Forecasting

We first evaluate deterministic point accuracy within our unified forecasting framework. Using its one-step point prediction mode, UNIDYNAMIC achieves a mean rollout L2RE of 0.04496 across eleven subsets (Table 1), lower than the reported results of DPOT-S and MoE-POT-S. It attains the lowest reported error among the compared models on $8/11$ subsets, including all three FNO Navier–Stokes cases and both PDEArena subsets. These results show that the unified architecture supports accurate deterministic forecasting through a one-step point readout.

6.2 Conditional Generation across Stochastic PDEs

We evaluate conditional generation on the three stochastic systems using 20-frame autoregressive rollouts against same-history DNS ensembles. UNIDYNAMIC achieves the lowest ED_n on all three systems (Table 2). Its lead-20 scores are approximately 62%, 40%, and 44% lower than the best baseline scores on Kolmogorov, reaction–diffusion, and damped waves, respectively. These gains are accompanied by the lowest E_s and E_{fluc} on Kolmogorov and reaction–diffusion. On damped waves, DPOT-FM has lower spread and fluctuation-spectrum errors, while UNIDYNAMIC achieves lower marginal-distribution error.

The differences cannot be explained by conditional-mean accuracy alone. On reaction–diffusion, DPOT-CRPS achieves the lowest E_μ, but its E_s is 0.8099, compared with 0.1066 for UNIDYNAMIC. The evaluated CRPS retrofits produce underdispersed ensembles over rollout despite competitive mean predictions. Figure 4 further shows that UNIDYNAMIC's lower distributional errors persist over the forecast horizon, rather than being confined to the final lead.

Table 2: **Stochastic forecasting at lead time 20.** All metrics ↓. Best values in bold. "–": omitted $E_s = 1$ for point forecasts. †: point-prediction pretraining before CRPS fine-tuning.

Method	Params (M)	Kolmogorov				Reaction–diffusion				Damped wave			
		E_μ	E_s	ED_n	E_{fluc}	E_μ	E_s	ED_n	E_{fluc}	E_μ	E_s	ED_n	E_{fluc}
Deterministic point predictors													
FNO	30.1	0.1521	–	0.1693	1.0000	0.0638	–	0.1452	1.0000	**0.0317**	–	0.1250	1.0000
DPOT	30.2	0.1477	–	0.1658	1.0000	0.0560	–	0.1424	1.0000	0.0606	–	0.1332	1.0000
CRPS-retrofitted probabilistic surrogates													
FNO-CRPS†	30.2	0.1381	0.5908	0.0685	0.8121	0.0656	0.4615	0.0311	1.2247	0.0500	0.6787	0.0605	0.9885
DPOT-CRPS†	30.9	0.1137	0.4819	0.0462	0.6883	**0.0508**	0.8099	0.0847	1.0304	0.0561	0.6172	0.0543	1.1039
Explicit flow/diffusion generators													
ACDM	30.3	0.7225	0.5590	0.5713	0.6814	0.1871	0.4235	0.0916	0.7244	0.5486	0.5333	0.3766	1.4063
GenCFD	29.8	0.1945	0.2080	0.0697	0.1683	0.1209	0.1771	0.0303	0.2623	0.1634	0.1796	0.0624	0.4731
DPOT-FM	30.4	0.1797	0.2073	0.0575	0.1882	0.0873	0.1243	0.0171	0.1508	0.0606	**0.0914**	0.0101	**0.0645**
UniDynamic	30.7	**0.0950**	**0.1290**	**0.0175**	**0.0844**	0.0639	**0.1066**	**0.0102**	**0.0987**	0.0402	0.0934	**0.0057**	0.0977

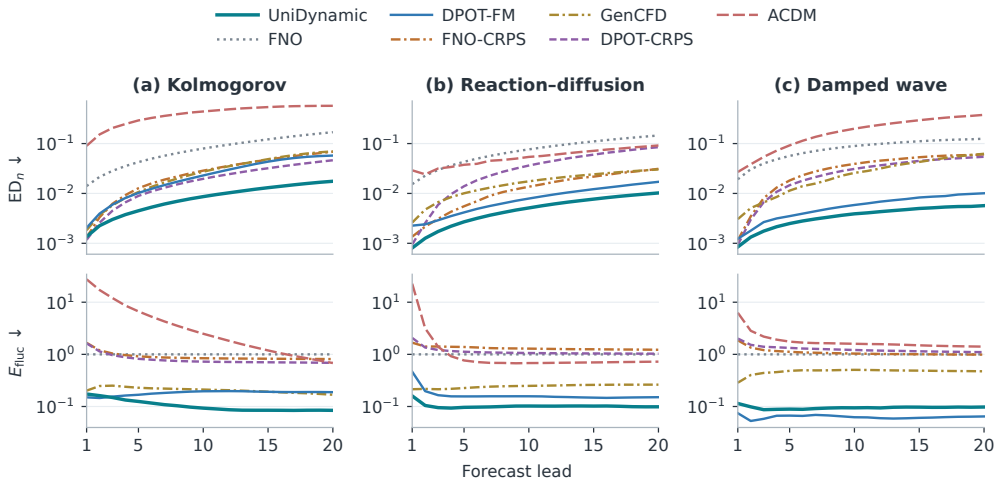

Figure 4: **Distributional errors over autoregressive rollout.** ED_n (top) and E_{fluc} (bottom) for selected methods in Table 2.

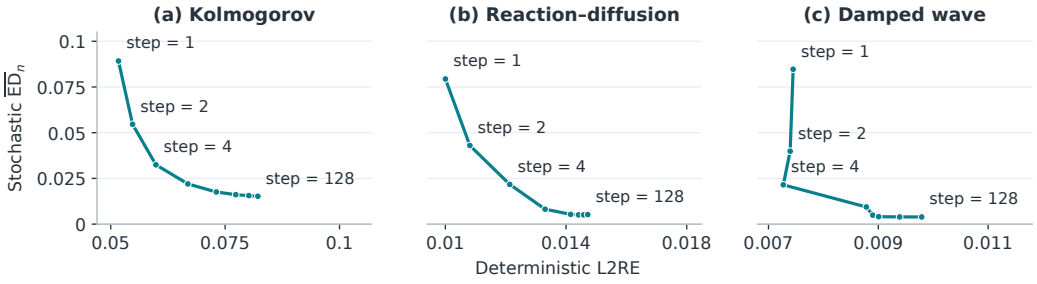

Figure 5: **One checkpoint, two inference modes.** Labels indicate Euler steps. Axes show per-trajectory deterministic L2RE and horizon-averaged stochastic $\overline{\mathrm{ED}}_n$. Both ↓.

6.3 One Checkpoint, Two Inference Modes

We train one UniDynamic jointly on deterministic and stochastic data from all three systems, then vary only the number of Euler steps per predicted frame. Figure 5 shows that the step count controls the trade-off between point accuracy and distributional fidelity. One-step point prediction favors low-cost, accurate deterministic forecasts, while multi-step conditional generation improves stochastic distribution matching. Both modes use the same weights, without retraining or a separate prediction head. Additional history-encoder ablations are reported in Appendix A.

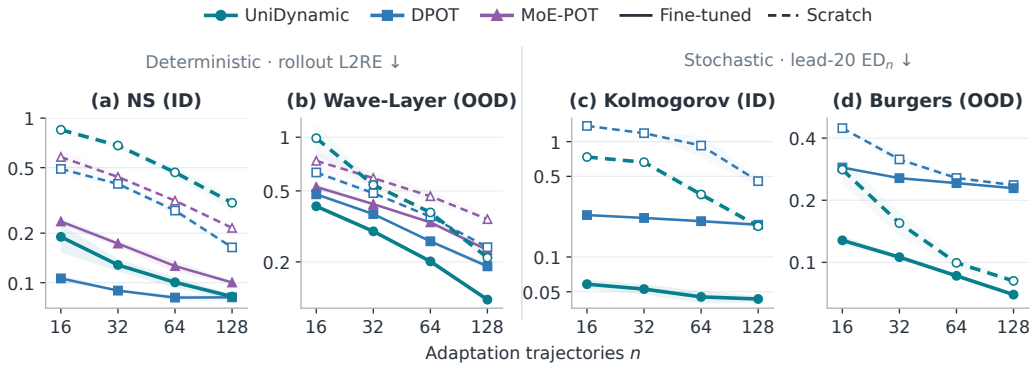

Figure 6: **Few-shot transfer.** Solid/dashed: fine-tuned/from scratch. Shading: three-seed SD. ID/OOD denote pretraining status. NS uses $\nu = 10^{-4}$.

6.4 Joint Training and Few-Shot Transfer

We examine whether deterministic data can improve stochastic forecasting, and whether pretraining supports adaptation to target PDEs with limited data.

Deterministic data for stochastic forecasting. We compare stochastic-only training with an equal deterministic–stochastic mixture while fixing the stochastic dataset, model, loss, and update budget. Joint training therefore receives roughly half the stochastic exposure. Table 3 reports horizon-averaged $\overline{\mathrm{ED}}_n$.

Joint training reduces $\overline{\mathrm{ED}}_n$ by 13.99% on Kolmogorov flow and 11.78% on reaction–diffusion, while damped-wave performance is nearly unchanged. These improvements are obtained while keeping the stochastic dataset and total update budget fixed, so the joint model receives less stochastic-only exposure than its S-only counterpart. The result suggests that deterministic trajectories can provide useful complementary training signals for stochastic forecasting.

Table 3: **Cross-regime joint training.** Mean $\overline{\mathrm{ED}}_n \downarrow$. Δ: Joint vs. S-only (%).

System	S-only	Joint	Δ (%)
Kolmogorov	0.011652	**0.010022**	−13.99
Reaction–diffusion	0.004879	**0.004304**	−11.78
Damped wave	**0.003513**	0.003579	+1.90

Few-shot adaptation across PDEs. Figure 6 compares fine-tuning with scratch training using 16–128 target trajectories. Pretraining improves performance at every displayed budget. On NS, UniDynamic closes the small-budget gap to DPOT and reaches comparable accuracy with 128 trajectories. On the unseen Wave-Layer and Burgers tasks, fine-tuning with 16 trajectories outperforms scratch training with 32. On Wave-Layer, it also outperforms the compared fine-tuned baselines throughout the displayed range. Overall, pretraining enables data-efficient adaptation for both point prediction and conditional generation, including unseen PDEs.

7 Conclusion

We introduced UniDynamic, a unified framework for deterministic point prediction and stochastic conditional generation. Zero-time supervision and source cancellation provide a source-independent one-step readout, while conditional flow matching enables multi-step sampling with the same network weights. Windowed log-signature history encoding summarizes ordered changes across temporal scales to condition both modes. Our same-history DNS benchmark supports direct evaluation of conditional means, spread, marginal distributions, and fluctuation spectra. Experiments on two-dimensional regular grids show competitive accuracy across eleven deterministic PDE subsets and the lowest lead-20 marginal energy distance among the compared methods on all three stochastic systems. Pretraining improves few-shot adaptation to unseen equations, while joint deterministic–stochastic training improves stochastic marginal-distribution accuracy on Kolmogorov flow and reaction–diffusion.

AI USE STATEMENT

The authors used general-purpose large language model (LLM) assistants for language editing of author-written text, literature retrieval and discovery, research brainstorming, and routine implementation support. In particular, LLMs were used to improve grammar, wording, clarity, and conciseness; to suggest search directions and potentially relevant prior work; and to assist with tasks such as plotting, debugging, and refactoring experiment code. All references suggested through AI-assisted retrieval were independently located and verified by the authors. LLM outputs were treated as suggestions rather than authoritative results. The authors made the final decisions regarding the research questions, methodology, mathematical claims, experimental design, and interpretation of results, and reviewed and verified all AI-assisted text, code, derivations, and citations. The authors take full responsibility for the contents of this paper, including any remaining errors.

REPRODUCIBILITY STATEMENT

We provide detailed information required to reproduce the experiments and validate the methodological claims throughout the paper and appendices. Appendix C specifies the field normalization, token construction, history representation, model architecture, and field readout used in the principal experiments. Appendix C.2 details the training losses, source-independent point-readout identity, Euler sampling procedure, and autoregressive feedback mechanism. Appendix D defines the stochastic benchmark, including the governing equations and numerical discretization, DNS data-generation pipeline, training/validation/test partitioning, repeated-future reference construction, and numerical and statistical validation checks. Experiment-specific training budgets, checkpoint-selection rules, evaluation settings, and random-seed aggregation are reported in the corresponding experimental sections and appendices. The public benchmark datasets used for deterministic evaluation are cited in the paper. The implementation code and scripts for data preparation, training, and evaluation will be released publicly upon acceptance.

REFERENCES

Shuhao Cao. Choose a transformer: Fourier or galerkin. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Tianyu Chen, Haoyi Zhou, Ying Li, Hao Wang, Chonghan Gao, Rongye Shi, Shanghang Zhang, and Jianxin Li. OmniArch: Building foundation model for scientific computing. *arXiv preprint arXiv:2402.16014*, 2024.

Zituo Chen and Sili Deng. Latent generative solvers for generalizable long-term physics simulation. *arXiv preprint arXiv:2602.11229*, 2026.

Ilya Chevyrev and Andrey Kormilitzin. A primer on the signature method in machine learning. *arXiv preprint arXiv:1603.03788*, 2016.

Guillaume Couairon, Alexis Jacq, Yu-Han Wu, Renu Singh, Yana Hasson, Quentin Berthet, and Romuald Elie. Kastor: An efficient fine-tuning strategy for generative emulation of PDE simulations. *arXiv preprint arXiv:2608.06107*, 2026.

Cristiana Diaconu, Miles Cranmer, Richard E. Turner, Tanya Marwah, and Payel Mukhopadhyay. Probabilistic retrofitting of learned simulators. *arXiv preprint arXiv:2603.01949*, 2026.

Anne Gagneux, Ségolène Martin, Rémi Gribonval, and Mathurin Massias. Training flow matching: The role of weighting and parameterization. *arXiv preprint arXiv:2603.06454*, 2026.

Zhongkai Hao, Zhengyi Wang, Hang Su, Chengyang Ying, Yinpeng Dong, Songming Liu, Ze Cheng, Jian Song, and Jun Zhu. GNOT: A general neural operator transformer for operator learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 12556–12569. PMLR, 2023.

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. *arXiv preprint arXiv:2403.03542*, 2024.

Maximilian Herde, Bogdan Raonic, Tobias Rohner, Roberto Käppeli, Roberto Molinaro, Emmanuel de Bézenac, and Siddhartha Mishra. Poseidon: Efficient foundation models for PDEs. *Advances in Neural Information Processing Systems*, 37, 2024.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*, 2022.

Georg Kohl, Li-Wei Chen, and Nils Thuerey. Benchmarking autoregressive conditional diffusion models for turbulent flow simulation. *arXiv preprint arXiv:2309.01745*, 2023.

Zijie Li, Kazem Meidani, and Amir Barati Farimani. Transformer for partial differential equations' operator learning. *Transactions on Machine Learning Research*, 2023.

Zijie Li, Anthony Zhou, and Amir Barati Farimani. Generative latent neural PDE solver using flow matching. *arXiv preprint arXiv:2503.22600*, 2025.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via DeepONet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 2021.

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.

Michael McCabe et al. Multiple physics pretraining for physical surrogate models. *arXiv preprint arXiv:2310.02994*, 2023.

Roberto Molinaro, Samuel Lanthaler, Bogdan Raonić, Tobias Rohner, Victor Armegioiu, Stephan Simonis, Dana Grund, Yannick Ramic, Zhong Yi Wan, Fei Sha, Siddhartha Mishra, and Leonardo Zepeda-Núñez. Generative ai for fast and accurate statistical computation of fluids. *arXiv preprint arXiv:2409.18359*, 2024.

James Morrill, Adeline Fermanian, Patrick Kidger, and Terry Lyons. A generalised signature method for multivariate time series feature extraction. *arXiv preprint arXiv:2006.00873*, 2020.

Sebastian Pfister, Benjamin Holzschuh, and Nils Thuerey. StocBench: A benchmark for generative modeling of stochastic dynamics. *arXiv preprint arXiv:2608.22309*, 2026.

Mahindra Singh Rautela, Alexander Most, Siddharth Mansingh, Bradley C. Love, Alexander Scheinker, Diane Oyen, Nathan Debardeleben, Earl Lawrence, and Ayan Biswas. MORPH: PDE foundation models with arbitrary data modality. *arXiv preprint arXiv:2509.21670*, 2025.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015.

Dengdi Sun, Xiaoya Zhou, Xiao Wang, Hao Si, Wanli Lyu, Jin Tang, and Bin Luo. NESTOR: A nested MOE-based neural operator for large-scale PDE pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6147–6156, 2026a.

Li Sun, Hongbo Lv, Zhikai Jiang, Zhongtian Sun, Lanxu Yang, and Philip Yu. Origo: Interpretable multi-physics PDE foundation model through neural operator splitting. In *International Conference on Machine Learning*, 2026b.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Dan MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. PDEBench: An extensive benchmark for scientific machine learning. *Advances in Neural Information Processing Systems*, 35:1596–1611, 2022.

Alasdair Tran, Alexander Mathews, Lexing Xie, and Cheng Soon Ong. Factorized fourier neural operators. In *International Conference on Learning Representations*, 2023.

Hong Wang, Haiyang Xin, Jie Wang, Xuanze Yang, Fei Zha, Huanshuo Dong, and Yan Jiang. Mixture-of-experts operator transformer for large-scale PDE pre-training. In *Advances in Neural Information Processing Systems*, volume 38, pp. 31498–31527, 2025. doi: 10.52202/085713-1057.

Rui Zhang, Qi Meng, Han Wan, Yang Liu, Zhi-Ming Ma, and Hao Sun. OmniFluids: Unified physics pre-trained modeling of fluid dynamics. *arXiv preprint arXiv:2506.10862*, 2025.

APPENDIX CONTENTS

A **History Encoder Ablations** . . . 14

B **Motivating Experiment: Stochastic Results** . . . 14

B.1 Data generation and evaluation . . . 14

B.2 DPOT and its flow-matching adaptation . . . 15

B.3 Full quantitative results . . . 16

B.4 Summary . . . 17

C **Model Details and Training Identities** . . . 17

C.1 Field coordinates and token construction . . . 17

C.2 Losses, source cancellation, and the sampler . . . 19

D **Benchmark Definition and Numerical Checks** . . . 19

D.1 Equations, forcing, and observations . . . 19

D.2 DNS data-generation and validation pipeline . . . 20

D.3 Dataset splits and reference allocations . . . 21

D.4 Numerical and physical checks . . . 22

E **Qualitative Visualizations** . . . 23

E.1 Deterministic autoregressive rollouts . . . 23

E.2 Stochastic conditional samples . . . 31

E.3 Out-of-distribution fine-tuning . . . 34

F **Additional Long-Horizon Experiments** . . . 36

F.1 Evaluation and quantitative results . . . 36

F.2 Error trajectories over 50 forecast frames . . . 37

F.3 Autoregressive field evolution . . . 38

G **Additional Deep-Ensemble Experiments** . . . 41

G.1 Evaluation and quantitative results . . . 41

G.2 Error trajectories over 20 forecast frames . . . 42

G.3 Conditional-ensemble mean and spread fields . . . 43

G.4 Sensitivity to DNS reference-ensemble size . . . 46

A History Encoder Ablations

Matched experimental setup. We compare history encoders under a fixed generation-only control. Every variant uses the same DiT and detail readout, 256 history tokens, history/data interface, 96k-update budget, global batch 32, optimizer schedule, and one training seed. Training pools all three systems under mid-strength stochastic forcing. Evaluation uses the same development references, 128 model/DNS members, 32 Euler steps, and lead-20 scores for every variant. The macro gives equal weight to the three equations.

Table 4: **History encoder ablations.** $\mathrm{ED}_n \downarrow$ averages the three PDEs at lead 20 (NFE32). Increases are relative to the full encoder.

History encoder	Params (M)	$\mathrm{ED}_n \downarrow$	Increase (%)
Full history encoder	30.7	**0.010595**	0.0
Component removals and replacements			
w/o local windows	28.8	0.013153	+24.1
w/o second-order path features	28.2	0.012816	+21.0
w/o both	28.0	0.018943	+78.8
Signed areas → feature covariance	30.7	0.012041	+13.6
Alternative history encoders			
DPOT pooling, aligned frame encoder	30.9	0.011960	+12.9
DPOT pooling, original history encoder	30.4	0.012131	+14.5

Removing local windows or second-order path features increases macro ED_n by 24.1% and 21.0%, respectively (Table 4). To test whether generic second-order statistics offer comparable benefits, we replace signed areas with within-window feature covariance, which is insensitive to permutation of the feature values within a fixed window. This increases macro ED_n by 13.6% with the same parameter count and initialization. Both temporal-pooling alternatives also yield higher macro ED_n than the full encoder. These results favor the full encoder in macro marginal-distribution accuracy under the common generation-only control.

Both temporal-pooling alternatives use the same DPOT `exp_mlp` temporal aggregator. The original control retains its coordinate-aware history stem and pre-aggregation position handling. The aligned control uses our 192-dimensional frame features, lifts them to width 512, and fuses the temporal aggregate with the last-frame features. Together, the two controls compare complete history-encoding interfaces, including frame encoding and temporal aggregation.

B Motivating Experiment: Stochastic Results

This section documents the stochastic Kolmogorov experiment in Section 3, comparing the original DPOT point-prediction architecture with its conditional flow-matching adaptation, DPOT-FM. The two models share the DPOT backbone family and matched parameter and training-update budgets.

B.1 Data generation and evaluation

Physical system and generated data. We generate data from the stochastically forced two-dimensional Kolmogorov vorticity equation on the periodic domain $[0, 2\pi)^2$, with viscosity $\nu = 0.01$ and vorticity forcing $-2\sin(4y)$. The physical stochastic-forcing amplitude is 0.10874644. The stochastic forcing is additive, white in time, and spatially colored, as specified in Appendix D.1. DNS uses a 256×256 grid, double precision, and an IMEX Runge–Kutta solver with time step 0.001. Amplitude-preserving Fourier restriction produces the observed 128×128 vorticity fields, saved every $\Delta t = 0.2$. The dataset contains 1,024 training and 128 validation trajectories, each containing 64 frames. Training uses ten consecutive observed frames to predict the next frame. Evaluation rolls out 20 frames, reaching physical time 4.0 after the last observation. Neither model receives the high-resolution state, future physical forcing, or an explicit forcing-amplitude input.

Held-out histories and DNS references. The independent test contains 48 held-out histories, separate from the training and ordinary validation cohorts. For each history, we branch from its final

high-resolution DNS state and independently resample the unobserved future forcing to obtain 128 continuations (ensemble A). All models are evaluated against this same fixed ensemble. A second, independently generated 128-member ensemble B is evaluated against A to provide a DNS–DNS sampling reference on the same test histories. DPOT produces one autoregressive trajectory, evaluated as a point distribution. DPOT-FM produces 128 trajectories: each member maintains its own rolling history and draws a fresh generative source at every forecast step. There is no teacher forcing during the test rollout.

Checkpoint selection and aggregation. For each training run, we select the checkpoint with the lowest validation fair CRPS under a predeclared rule, using eight DPOT-FM members and 32 Euler steps per frame. Checkpoint files are frozen before the independent DNS test. All test metrics use the same selected checkpoint. For each seed, scores are computed per history and lead, averaged over the 48 histories, and then, for rollout averages, over leads 1–20. Model scores are reported as the mean and sample standard deviation across three training seeds.

Metrics. We use E_μ, E_s, ED_n, and E_{fluc} as defined in Section 5.3.

B.2 DPOT AND ITS FLOW-MATCHING ADAPTATION

DPOT: the original point-prediction path. DPOT retains its coordinate-aware history patch embedding, learned spatial position embedding, `exp_mlp` temporal aggregator, AFNO blocks, and convolutional field-reconstruction head. The history encoder appends spatial coordinates and the normalized history-frame position to each vorticity frame. An 8×8 strided patch convolution and a channel projection produce width-448 features on a 16×16 latent grid. The original `exp_mlp` combines learned cosine temporal modulation with position-specific linear aggregation. It is not plain temporal averaging. Six AFNO blocks process the aggregate, and the reconstruction head directly predicts the full next frame y, not a last-frame residual. Only the prediction path is retained. There is no auxiliary classification head or classification loss. For history H, DPOT is trained with the pixel-averaged squared loss on $f_\theta(H) - y$ in physical field coordinates.

DPOT-FM: noisy-state input and flow-time conditioning. DPOT-FM keeps the same history-conditioning interface and AFNO backbone family, but additionally receives a noisy next-state field z_τ and a flow time τ. To avoid discarding source information at the patch encoder, an 8×8 pixel-unshuffle operation directly retains all 64 scalar values of each noisy-state patch. These are concatenated with 384 learned features projected from the full patch and its coordinates, yielding 448 channels. The noisy-state and history features are then concatenated and fused by a 1×1 convolution from 896 to 448 channels. A sinusoidal flow-time embedding followed by a two-layer MLP conditions each AFNO block through learned shifts, scales, and residual gates. The underlying block normalization remains GroupNorm. The final time-modulation projections are zero-initialized. The field-reconstruction head now predicts a velocity $g_\theta(H, z_\tau, \tau)$ rather than the next state itself.

Flow-matching objective and sampling. Let H_{last} be the most recent observed frame. The generative source scale depends only on this conditioning frame:

$$s_H = \mathrm{std}_{\mathrm{space}}(H_{\mathrm{last}}) + 10^{-6}, \qquad e = s_H \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I),$$
$$\tau \sim \mathcal{U}[0, 1), \qquad z_\tau = (1 - \tau)e + \tau y,$$
$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\Big[\mathrm{mean}_{\mathrm{pixels}} \, |g_\theta(H, z_\tau, \tau) - (y - e)|^2\Big].$$

The spatial standard deviation uses the sample estimator. The source scale s_H is neither the physical forcing amplitude nor a statistic of the target. Training samples one source and one flow time per example. It does not backpropagate through an ODE rollout. There is no auxiliary point-prediction loss, endpoint parameterization, or source-cancellation module from UNIDYNAMIC. At inference, each next frame is obtained with 32 explicit Euler steps:

$$z^{(0)} = e, \qquad z^{(j+1)} = z^{(j)} + \frac{1}{32} g_\theta\left(H, z^{(j)}, \frac{j}{32}\right), \quad j = 0, \ldots, 31.$$

The output $z^{(32)}$ is appended to that member's history. The history encoding is cached within these 32 steps, but recomputed after the autoregressive history changes.

Capacity and training controls. Both models use width 448, six AFNO blocks, and 8×8 patches. To account for the added noisy-state and time-conditioning modules, the block FFN hidden widths are 1,737 for DPOT and 448 for DPOT-FM, giving 13,625,954 and 13,625,964 prediction parameters, respectively. The parameter budgets are therefore closely matched. Each model is trained from scratch on this stochastic dataset for 128,000 updates, with global batch 32 and the same three training seeds, data-window streams, and optimizer schedule. Both use Adam with $\beta = (0.9, 0.9)$, weight decay 10^{-6}, and gradient-norm clipping at 1. The learning rate warms up from 10^{-5} to 10^{-3} over 6,400 updates, then follows cosine decay back to 10^{-5}. Both objectives operate on unnormalized physical fields. At inference, DPOT uses one network call per frame. DPOT-FM generates 128 members, each with 32 velocity evaluations per frame.

B.3 Full quantitative results

Table 5 reports the rollout-average and terminal-lead results from the validation-selected checkpoints, including the independent DNS B128-versus-A128 sampling reference for every diagnostic.

Table 5: **Stochastic motivating experiment: full numerical diagnostics.** Mean $\pm$ sample SD over three training seeds, each evaluated on 48 histories. Panels report the 20-step average and terminal lead (physical time 4.0). The DNS column provides an independent B128-versus-A128 sampling reference using the same target ensemble as the model comparisons.

Metric	DPOT	DPOT-FM	DNS B128 vs A128
Mean over leads 1–20			
$E_\mu \downarrow$	0.06285 ± 0.00587	0.05813 ± 0.00353	0.02098
$E_s \downarrow$	1.00000 ± 0.00000	0.11282 ± 0.00098	0.08862
Raw ED $\downarrow$	0.16662 ± 0.00480	0.02170 ± 0.00235	0.00492
$\mathrm{ED}_n \downarrow$	0.08647 ± 0.00263	0.01127 ± 0.00127	0.00258
Std. RMSE $\downarrow$	0.33449 ± 0.00000	0.04075 ± 0.00060	0.02968
Std. RMSE / DNS RMS $\downarrow$	0.17470 ± 0.00000	0.02123 ± 0.00032	0.01558
Mean-spectrum error $\downarrow$	0.04169 ± 0.00239	0.01864 ± 0.00418	0.00660
Member-spectrum error $\downarrow$	0.01983 ± 0.00233	0.01763 ± 0.00383	0.00601
$E_{\mathrm{fluc}} \downarrow$	1.00000 ± 0.00000	0.06799 ± 0.00373	0.04820
Predicted spread RMS	0.00000 ± 0.00000	0.33028 ± 0.00160	0.33505
DNS spread RMS	0.33449 ± 0.00000	0.33449 ± 0.00000	0.33449
Spread / DNS spread (1 is ideal)	0.00000 ± 0.00000	0.98116 ± 0.00378	1.00099
Lead 20			
$E_\mu \downarrow$	0.15654 ± 0.01263	0.12184 ± 0.00893	0.04050
$E_s \downarrow$	1.00000 ± 0.00000	0.14141 ± 0.00470	0.09286
Raw ED $\downarrow$	0.33406 ± 0.01126	0.04733 ± 0.00618	0.00921
$\mathrm{ED}_n \downarrow$	0.17113 ± 0.00619	0.02468 ± 0.00316	0.00477
Std. RMSE $\downarrow$	0.63651 ± 0.00000	0.08998 ± 0.00339	0.05824
Std. RMSE / DNS RMS $\downarrow$	0.32882 ± 0.00000	0.04665 ± 0.00174	0.03033
Mean-spectrum error $\downarrow$	0.12077 ± 0.00668	0.03911 ± 0.00895	0.01307
Member-spectrum error $\downarrow$	0.04246 ± 0.00424	0.03549 ± 0.00817	0.01071
$E_{\mathrm{fluc}} \downarrow$	1.00000 ± 0.00000	0.06482 ± 0.00290	0.05669
Predicted spread RMS	0.00000 ± 0.00000	0.63101 ± 0.00376	0.63723
DNS spread RMS	0.63651 ± 0.00000	0.63651 ± 0.00000	0.63651
Spread / DNS spread (1 is ideal)	0.00000 ± 0.00000	0.99045 ± 0.00591	1.00187

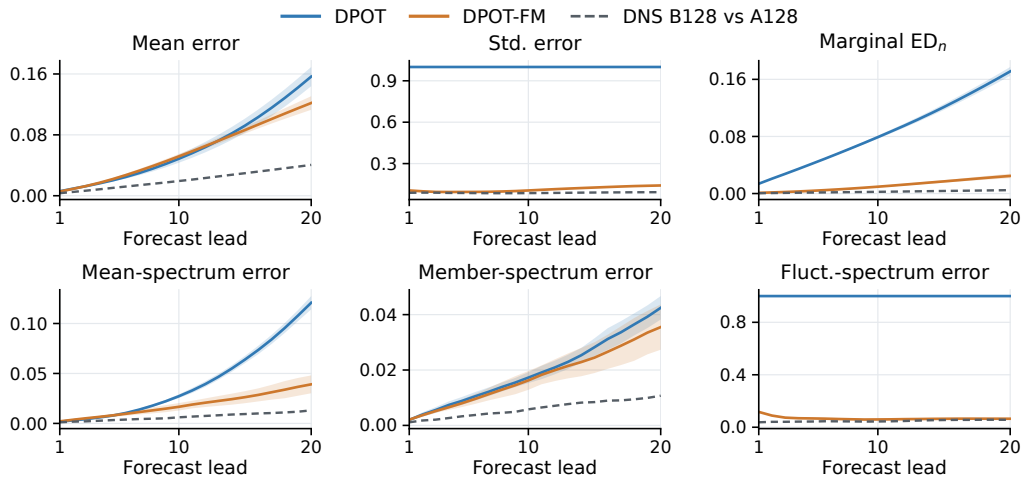

Figure 7: **Stochastic motivating experiment: lead-wise results.** Means over 48 histories, with SD shading over three training seeds. DPOT-FM uses 128 members and 32 Euler steps. Dashed curves compare independent DNS B128 with the common A128 target.

Conditional-distribution accuracy. Under matched parameter and training-update budgets, DPOT-FM reduces rollout-averaged ED_n from 0.08647 to 0.01127, an 87.0% reduction relative to DPOT (Table 5). At lead 20, ED_n decreases from 0.17113 to 0.02468, a reduction of 85.6%. Figure 7 shows lower seed-averaged ED_n at every forecast lead. The rollout-averaged reduction holds for each of the three training seeds and is retained when both methods use their final 128k checkpoints.

Recovery of stochastic variability. DPOT-FM achieves a predicted-to-DNS RMS spread ratio of 0.98116 averaged over the rollout and 0.99045 at lead 20, recovering the conditional variability absent from DPOT's point forecasts. Its rollout-averaged fluctuation-spectrum error falls to 0.06799, compared with 1.00000 for DPOT. At lead 20, this error is 0.06482, with an independent DNS–DNS reference of 0.05669. The distributional gains are accompanied by a reduction in rollout-averaged mean-field error E_μ, from 0.06285 to 0.05813.

B.4 Summary

On stochastic Kolmogorov flow, DPOT-FM more closely matches DNS marginal distributions, ensemble spread, and fluctuation spectra than DPOT, while also reducing rollout-averaged conditional-mean error. These results motivate conditional generation alongside point prediction for forecasting under unobserved future forcing.

C Model Details and Training Identities

C.1 Field coordinates and token construction

For each history and physical channel, let $\mu_{H,c}$ and $s_{H,c}$ be the mean and population standard deviation over the observed frames and spatial locations. The implementation uses

$$a_{H,c} = \max\left\{ s_{H,c}, \epsilon_{\mathrm{norm}} \left(1 + \sqrt{\mathrm{mean}(H_c^2)}\right) \right\},$$
$$\widetilde{H}_c = \frac{H_c - \mu_{H,c}}{a_{H,c}}, \qquad y_c = \frac{Y_c - \mu_{H,c}}{a_{H,c}}.$$

Here $\epsilon_{\mathrm{norm}} = 10^{-6}$ sets a lower bound on the normalization scale for nearly constant histories. These statistics are computed exclusively from the observed history. The inverse map is $\mathcal{D}_H(x)_c =$

$a_{H,c}x_c + \mu_{H,c}$. Unused channels are padded with ones in physical coordinates and hence zeros in normalized coordinates. Gaussian sources and velocity MSE cover all four model channels. The physical relative error and reported metrics cover only the observed physical channels.

Table 6: **Architecture used by the principal UNIDYNAMIC results.** Tables 1 and 2, and Figure 5, use this architecture.

Component	Setting	Value
Input and tokenization	History length	10 frames
	Observed grid	128×128
	Padded channels	4
	Patch size	8×8
	History tokens	256
	State tokens	256
History encoder	Frame-feature width	192
	Number of feature groups	12
	Features per group	16
	Windows (inclusive frame indices)	$[1, 10]$, $[1, 6]$, $[5, 10]$
	Projection input dimension	5,700 per patch
Shared Transformer	Depth	8 blocks
	Hidden width	512
	Attention heads	4
	Positional encoding	2D RoPE
	Feed-forward nonlinearity	SwiGLU
	Feed-forward hidden width	344
	Attention dropout	0
	Projection dropout	0
Output head	Decoder intermediate channels	32
Model size	Trainable parameters	30,702,088

History path features. Pixel unshuffle packs each frame into 256-dimensional patch vectors. A shared linear map followed by GELU produces 192 features per patch. After grouping into twelve 16-dimensional paths, a normalized observation time coordinate makes each path 17-dimensional. Each window contributes $17 + \binom{17}{2} = 153$ statistics per group: 17 first-level increments and 136 signed-area terms. The concatenation in Figure 2 therefore has dimension $192 + 3 \times 12 \times 153 = 5{,}700$. The two shorter windows overlap and reuse the full-history normalized time coordinate. The reported principal models use these three windows.

The signed areas can be accumulated without constructing a full second-order signature tensor. With $P_i = \sum_{j=a}^{i-1} d_j$, Eq. (1) is equivalent to

$$A_w = \frac{1}{2} \sum_{i=a}^{b-1} \left(P_i d_i^\top - d_i P_i^\top \right).$$

Reversing two non-collinear increments reverses their contribution to A_w, while leaving their sum unchanged. The signed areas therefore encode chronological structure beyond the endpoint increments.

State tokens and field readout. The state embedding retains the 256 raw values of each normalized patch as a prefix and concatenates a learned suffix to reach width 512. The raw-value prefix preserves every patch coordinate at the backbone input. History tokens use the fixed clean-history flow-time embedding at one. State tokens use the current τ. The two streams share spatial RoPE coordinates but have distinct role embeddings and joint routing positions. The shared Transformer decodes only state tokens. If r_{base} is the decoded field and $g(h)$ is a learned patchwise, channelwise gate, the detail-adaptive output is

$$r_\theta = r_{\text{base}} + \tanh(g(h)) \odot \big(r_{\text{base}} - \text{PatchMean}(r_{\text{base}}) \big).$$

Here h denotes decoded state-token features. The patch mean is broadcast back to the field grid.

C.2 Losses, source cancellation, and the sampler

For a valid physical-channel set $\mathcal{C}$, the relative loss is

$$\mathcal{L}_{\mathrm{rel}}(\tau) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{\|\mathcal{D}_H(\hat{y}_\tau)_c - Y_c\|_{2,\mathrm{space}}}{\max\{\|Y_c\|_{2,\mathrm{space}}, \epsilon_{\mathrm{rel}}\}}.$$

Here $\epsilon_{\mathrm{rel}} = 10^{-8}$ is a denominator floor that prevents division by zero. The loss is unsquared and is computed per example and channel before batch averaging. Velocity MSE is evaluated in normalized field coordinates. The reconstruction and velocity losses use the same positive-time forward evaluation.

Velocity and point-readout identities. From $z_\tau = (1-\tau)e + \tau y$,

$$\hat{y}_\tau - y = (1-\tau)\big[v_\theta(c_H, z_\tau, \tau) - (y - e)\big].$$

Squaring and averaging proves Eq. (3). At zero time, Eq. (4) gives $e + v_\theta(c_H, e, 0) = r_\theta(c_H, 0, 0)$ for every e. The point readout is therefore exactly source-independent by construction.

Euler sampling and autoregressive feedback. For N steps per forecast frame, we draw $z_0 = e$ and use

$$z_{j+1} = z_j + \frac{1}{N} v_\theta(c_H, z_j, j/N), \qquad j = 0, \ldots, N-1.$$

The first evaluation uses the exact-zero branch, so $z_1 = (1 - N^{-1})e + N^{-1} r_\theta(c_H, 0, 0)$. For $N = 1$ this is the source-independent point prediction. After the first Euler step, the same network takes the current flow state as input at positive flow times. After inverse normalization, each member appends its own predicted frame, discards the oldest frame, recomputes history statistics, and draws fresh Gaussian noise for the next physical step.

D Benchmark Definition and Numerical Checks

D.1 Equations, forcing, and observations

All three systems use periodic boundaries and additive stochastic forcing interpreted in the Itô sense. All physical state channels are observed on the coarse grid. Future stochastic-forcing realizations are not inputs. Forcing levels denote the stochastic amplitude σ, with σ_{mid} denoting each system's mid-strength amplitude. Setting $\sigma = 0$ removes only stochastic forcing and leaves the deterministic drift unchanged.

Kolmogorov flow. On $[0, 2\pi)^2$, vorticity obeys

$$d\omega = [-\mathbf{u} \cdot \nabla\omega + 0.01\Delta\omega - 2\sin(4y)]\, dt + \sigma Q^{1/2} dW_t.$$

The incompressible velocity is recovered spectrally from vorticity. We use four forcing strengths $\sigma \in \{0, \sigma_{\mathrm{mid}}/2, \sigma_{\mathrm{mid}}, 2\sigma_{\mathrm{mid}}\}$, where $\sigma_{\mathrm{mid}} \approx 0.109$ is selected through DNS-based calibration of conditional spread. State preparation and burn-in diagnostics are described in Appendix D.2.

Reaction–diffusion. On $[0, 2)^2$, the activator and inhibitor satisfy

$$\begin{aligned} du &= [0.001\Delta u + u - u^3 - 0.005 - v]dt + \sigma Q^{1/2} dW_t, \\ dv &= [0.005\Delta v + u - v]dt, \end{aligned} \qquad \sigma \in \{0, 0.05, 0.1, 0.2\}.$$

Stochastic forcing acts only on u. Initial perturbations about $u = v = -\sqrt[3]{0.005}$ use two independent, spectrally filtered Gaussian draws, each rescaled to unit spatial RMS. The activator uses the first draw. The inhibitor mixes both with weights ρ and $\sqrt{1-\rho^2}$, where $\rho = 0.5$. The channel amplitude scales are $0.8/0.4$, with filter parameters $(\alpha_{\mathrm{init}}, k_{\mathrm{init}}) = (0.01, 20)$. The reaction terms follow the PDEBench setting (Takamoto et al., 2022). Our benchmark combines them with periodic boundaries and a spectral discretization.

Damped nonlinear wave. On $[0, 2\pi)^2$, both displacement and velocity are observed:

$$du = v\,dt, \qquad dv = [\Delta u - 0.1v - u - 0.25u^3]dt + \sigma Q^{1/2} dW_t, \qquad \sigma \in \{0, 0.2, 0.4, 0.8\}.$$

Stochastic forcing acts on velocity. Independent, zero-mean initial fields are obtained by spectrally filtering Gaussian draws and rescaling to displacement/velocity RMS $0.6/0.6$, with $(\alpha_{\text{init}}, k_{\text{init}}) = (0.05, 12)$.

Stochastic forcing and discretization. For physical wavenumber $\mathbf{k}$, the Fourier multiplier of $Q^{1/2}$ is proportional to $\exp(-\alpha|\mathbf{k}|^2/2)\mathbf{1}_{\{0<|\mathbf{k}|\le k_c\}}$, normalized to unit expected spatial mean square per unit time before multiplication by σ. The pairs (α, k_c) are $(0.1, 8)$, $(0.02, 16)$, and $(0.05, 8)$ for Kolmogorov, reaction–diffusion, and wave. Initial-field filters use the same exponential-cutoff form with their specified initial spectral parameters and per-draw RMS rescaling. All systems integrate 256^2 states in float64 and store amplitude-preserving Fourier restrictions to 128^2 in float32. Kolmogorov and reaction–diffusion use pseudo-spectral IMEX-RK3 drift updates. Wave uses symmetric nonlinear kicks around exact damped linear Fourier-mode evolution. The stochastic-forcing increment is applied after each complete solver step. The solver/output timesteps are respectively $0.001/0.2$, $0.002/0.1$, and $0.005/0.1$.

D.2 DNS DATA-GENERATION AND VALIDATION PIPELINE

All systems share fine-grid evolution, coarse-grid observation, and conditional branching, with state preparation adapted to developed driven flow for Kolmogorov and transient evolution for reaction–diffusion and wave.

State preparation. Kolmogorov chains start from independently filtered Gaussian vorticity draws, rescaled to RMS 2, with spectral parameter 0.08 and cutoff 12. Under each of the four forcing conditions, chains evolve for 60 physical time units before recording to reduce initialization transients. Burn-in frames are excluded. Reaction–diffusion and wave record directly from the prescribed initial states in Appendix D.1, preserving the transient response that defines these tasks.

DNS evolution and observations. The DNS remains on the fine grid throughout. Fourier restriction produces coarse observations at the intervals specified in Appendix D.1. Native trajectories retain the same forcing amplitude during preparation, history, and future continuation. Stochastic variability is generated by the forcing in the governing dynamics. Observations are obtained directly by restricting the DNS fields. Ordinary trajectories contain 64 frames, and each reference history contains ten consecutive observations.

Conditional-ensemble construction. The model observes only the coarse history H. DNS members branch from the saved fine-grid state at the final observed time under independent future stochastic-forcing paths. Physical coefficients and boundary conditions remain unchanged. Each reference ensemble varies future forcing at a common recorded fine-grid anchor, with the unresolved fine-grid state held fixed. Initial-state sampling, history forcing, and member-specific future forcing use distinct random-number streams. Kolmogorov additionally separates its burn-in stream. Native $\sigma = 0$ references contain one deterministic continuation. Matched deterministic controls instead reuse a stochastic history and anchor while disabling only future stochastic forcing.

Data partitioning and source independence. Ordinary trajectories are partitioned by source-chain or initial-condition identifier, with all overlapping windows retained in one split. Kolmogorov uses one independently initialized chain per source anchor. Reaction–diffusion and wave pair forcing amplitudes using the same initial fields within each partition, while training, validation, and test use disjoint initial-condition sets. This pairing provides matched initial states for studying forcing-dependent evolution. Reference allocations distinguish conditioning histories from repeated future members. Increasing ensemble size refines the statistics within each history.

Statistical and release checks. Kolmogorov sampling diagnostics. For each primary regime $\sigma \in \{0, \sigma_{\text{mid}}\}$, 16 independently initialized chains run to time 120. Diagnostics sampled every 0.2 over $[60, 120]$ give 301 observations per chain. We monitor vorticity RMS, kinetic energy, enstrophy $\frac{1}{2}\langle\omega^2\rangle_\Omega$, and palinstrophy $\frac{1}{2}\langle|\nabla\omega|^2\rangle_\Omega$ to assess both field amplitudes and fine-scale gradients.

Relative drift is the absolute fitted change in the across-chain mean divided by its interval mean. Half-window difference is the absolute difference between the two half-window means relative to their average. Classical split-$\widehat{R}$ compares between-chain and within-chain variation. The configured screening limits are 15% drift, 10% half-window difference, and split-$\widehat{R} < 2$. All monitored quantities satisfy these configured screening limits (Table 7): drift is at most 5.83%, half-window difference is at most 2.83%, and split-$\widehat{R}$ ranges from 1.27 to 1.65.

Temporal dependence is quantified by $g = 1 + 2\sum_{j=1}^{j_+} \widehat{\rho}(j)$, summing empirical lag correlations and stopping before the first nonpositive value. For each observable, the largest factor across chains, $g_{\max}$, defines $\tau_{\text{int}}^{\max} = 0.2 g_{\max}$ and $N_{\text{eff}} = 16 \times 301/g_{\max}$. These effective counts describe the diagnostic time series. Dataset anchors come from independent source chains.

Table 7: **Kolmogorov history-generation diagnostics.** Sixteen chains per stochastic-forcing regime, analyzed over physical times 60–120. Drift and half-window difference are relative percentages. $\tau_{\text{int}}^{\max}$ is in physical time units. The effective count N_{eff} uses the largest within-chain autocorrelation estimate for each observable.

Regime	Observable	Drift (%)	Half diff. (%)	$\tau_{\text{int}}^{\max}$	$\widehat{R}$	N_{eff}
$\sigma = 0$	Vorticity RMS	0.61	0.79	15.73	1.35	61.2
	Kinetic energy	5.83	2.83	19.29	1.65	49.9
	Enstrophy	1.16	1.34	15.92	1.33	60.5
	Palinstrophy	2.01	1.88	13.23	1.27	72.8
$\sigma = \sigma_{\text{mid}}$	Vorticity RMS	2.05	0.81	19.08	1.36	50.5
	Kinetic energy	2.18	1.16	23.34	1.60	41.3
	Enstrophy	5.41	1.07	18.61	1.33	51.8
	Palinstrophy	4.28	2.22	16.47	1.29	58.5

Equation-specific physical checks, numerical refinement, and conditional-ensemble diagnostics are summarized in Appendix D.4.

Released-data consistency. Verification checks file and content hashes, identifiers, split membership, and matched-control histories and anchors. The reaction–diffusion and wave audits each include $1{,}728$ exact paired-initial comparisons and 16 exact matched-history checks. These checks verify complete two-channel states and disjoint initial-condition sets across the training, validation, and test trajectory partitions. Kolmogorov metadata records $3{,}928$ distinct source anchors and chains across 210 source-state pool shards, all using the 60-unit burn-in. For one first-frame sample per model-facing Kolmogorov shard, restriction of the saved fine-grid source state reproduces the stored observation with zero measured discrepancy.

D.3 Dataset splits and reference allocations

Ordinary trajectories supply the training, validation, and test corpus. Additional DNS references supply repeated futures for conditional evaluation.

Trajectory corpus and splits. Table 8 summarizes the ordinary trajectories across all four forcing strengths. For every system, the training split includes $1{,}024$ trajectories without stochastic forcing and $1{,}024$ with mid-strength stochastic forcing. Each low- or high-strength training pool contains 256 trajectories for Kolmogorov and 128 each for reaction–diffusion and wave. Source-based partitioning follows Appendix D.2. Reference continuations are excluded from these counts.

Table 8: **Trajectory corpus.** Counts include all four forcing strengths. Each trajectory has 64 frames. The extra DNS reference branches are excluded from these counts.

System	Train	Validation	Test	Total
Kolmogorov	2,560	384	768	3,712
Reaction–diffusion	2,304	320	384	3,008
Damped wave	2,304	320	384	3,008

Primary evaluation references. We describe each reference allocation by $N_H \times M_{\mathrm{DNS}} \times L$: the number of conditioning histories, the DNS members per history, and the number of future frames per member. The principal stochastic comparison uses the wide mid-strength references in Table 9. Deterministic evaluations on this three-system benchmark use native $\sigma = 0$ references. Matched controls disable only future stochastic forcing after mid-strength stochastic histories and serve as diagnostic interventions.

Table 9: **Wide reference allocations and matched controls.** Entries are $N_H \times M_{\mathrm{DNS}} \times L$. Matched controls reuse the stochastic reference histories. The listed N_H counts those shared histories.

Reference	Kolmogorov	Reaction–diffusion	Damped wave
Stochastic ($\sigma = \sigma_{\mathrm{mid}}$)	$48 \times 128 \times 20$	$32 \times 128 \times 20$	$32 \times 128 \times 20$
Native deterministic ($\sigma = 0$)	$48 \times 1 \times 20$	$32 \times 1 \times 20$	$32 \times 1 \times 20$
Matched deterministic control	$48 \times 1 \times 20$	$32 \times 1 \times 20$	$32 \times 1 \times 20$

Extended reference allocations. Table 10 extends forcing coverage through low- and high-strength wide references. At mid strength, deep references provide more members per history for finite-ensemble assessment, while long references extend forecasting to 50 frames. The wide, deep, and long allocations define separate evaluation cohorts with their own history, member, and horizon counts.

Table 10: **Extended reference allocations.** Entries are $N_H \times M_{\mathrm{DNS}} \times L$. The first row applies separately to the low and high stochastic-forcing amplitudes. Deep and long references use σ_{mid}.

Reference	Kolmogorov	Reaction–diffusion	Damped wave
Wide, low/high σ	$48 \times 128 \times 20$	$32 \times 128 \times 20$	$32 \times 128 \times 20$
Deep, σ_{mid}	$8 \times 512 \times 20$	$4 \times 256 \times 20$	$4 \times 256 \times 20$
Long, σ_{mid}	$16 \times 128 \times 50$	$8 \times 64 \times 50$	$8 \times 64 \times 50$

D.4 Numerical and physical checks

Numerical consistency. Temporal and spatial comparisons share the same initial state and coupled realizations of the stochastic forcing. Coarse Brownian increments are sums of fine-step increments, and spatial comparisons retain the same forced Fourier modes. This separates discretization effects from differences in the random driving. Controlled tests show decreasing solution discrepancies with smaller timesteps and finer grids over the tested evolution intervals. Precision checks complement these comparisons.

Physical consistency. For Kolmogorov flow, velocity is reconstructed from vorticity using the incompressible Fourier relation. The developed-flow statistics are assessed by the sampling diagnostics in Appendix D.2. Reaction–diffusion preserves its homogeneous deterministic equilibrium and exhibits nontrivial transient evolution from perturbed initial states. For wave dynamics, Fourier modes agree with their analytical evolution in the linear, undamped, unforced limit. Separate nonlinear-wave tests with stochastic forcing disabled ($\sigma = 0$) check energy conservation without damping and the balance between energy loss and accumulated dissipation with damping.

Conditional variability. At a fixed fine-grid state, $\sigma = 0$ defines a deterministic future. Stochastically forced continuations exhibit nonzero ensemble spread. For reaction–diffusion and wave, variability also reaches the channel without direct stochastic forcing through the coupled dynamics. Within-history comparisons across ensemble sizes assess the sampling variability of conditional statistics separately from discretization effects. Together, these checks support the physical and numerical consistency of the reference trajectories in the tested configurations.

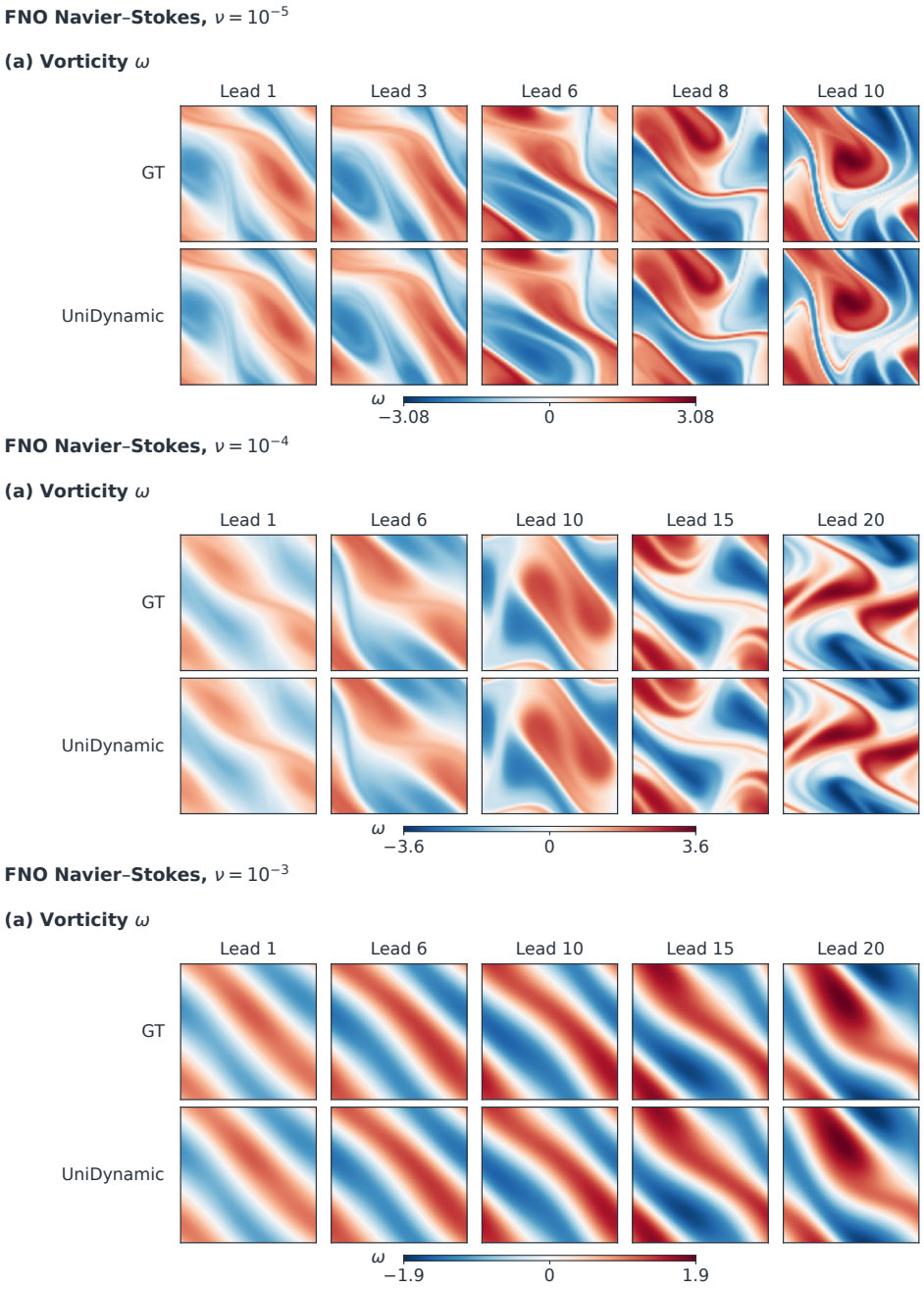

Figure 8: **FNO Navier–Stokes.** Top to bottom: $\nu = 10^{-5}$, 10^{-4}, and 10^{-3}. Within each block, GT is above UNIDYNAMIC.

E QUALITATIVE VISUALIZATIONS

E.1 DETERMINISTIC AUTOREGRESSIVE ROLLOUTS

Figures 8–16 show the first test trajectory and all physical channels for each of the eleven deterministic subsets. Forecasts are autoregressive from ten observed frames, using one Euler step per forecast frame. Each variable has a fixed color scale shared by ground truth (GT) and UNIDYNAMIC across forecast leads.

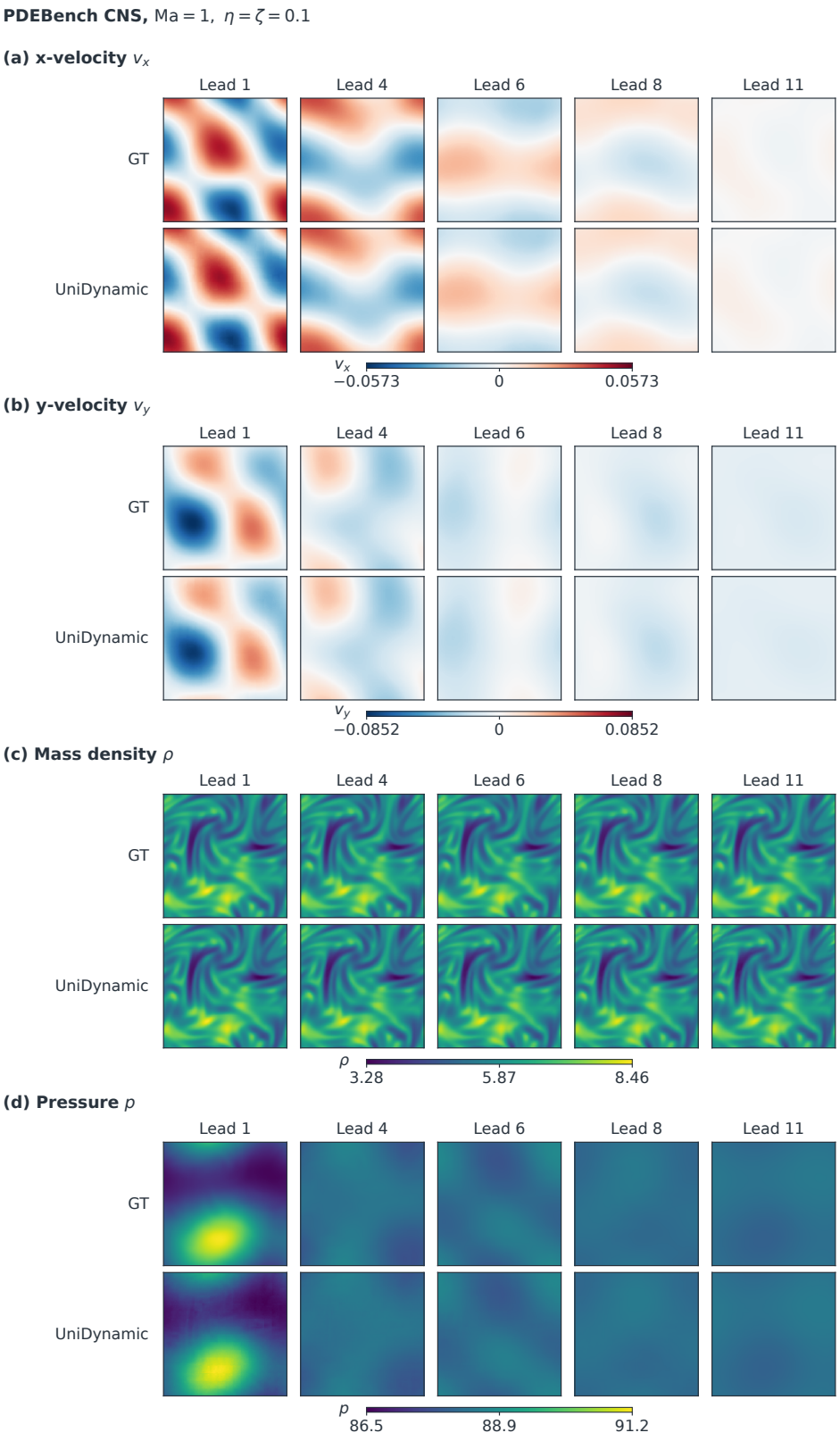

Figure 9: **PDEBench CNS** ($\mathrm{Ma} = 1$, $\eta = \zeta = 0.1$)**.** GT (top) and UNIDYNAMIC (bottom) for every labeled physical variable at the indicated rollout leads.

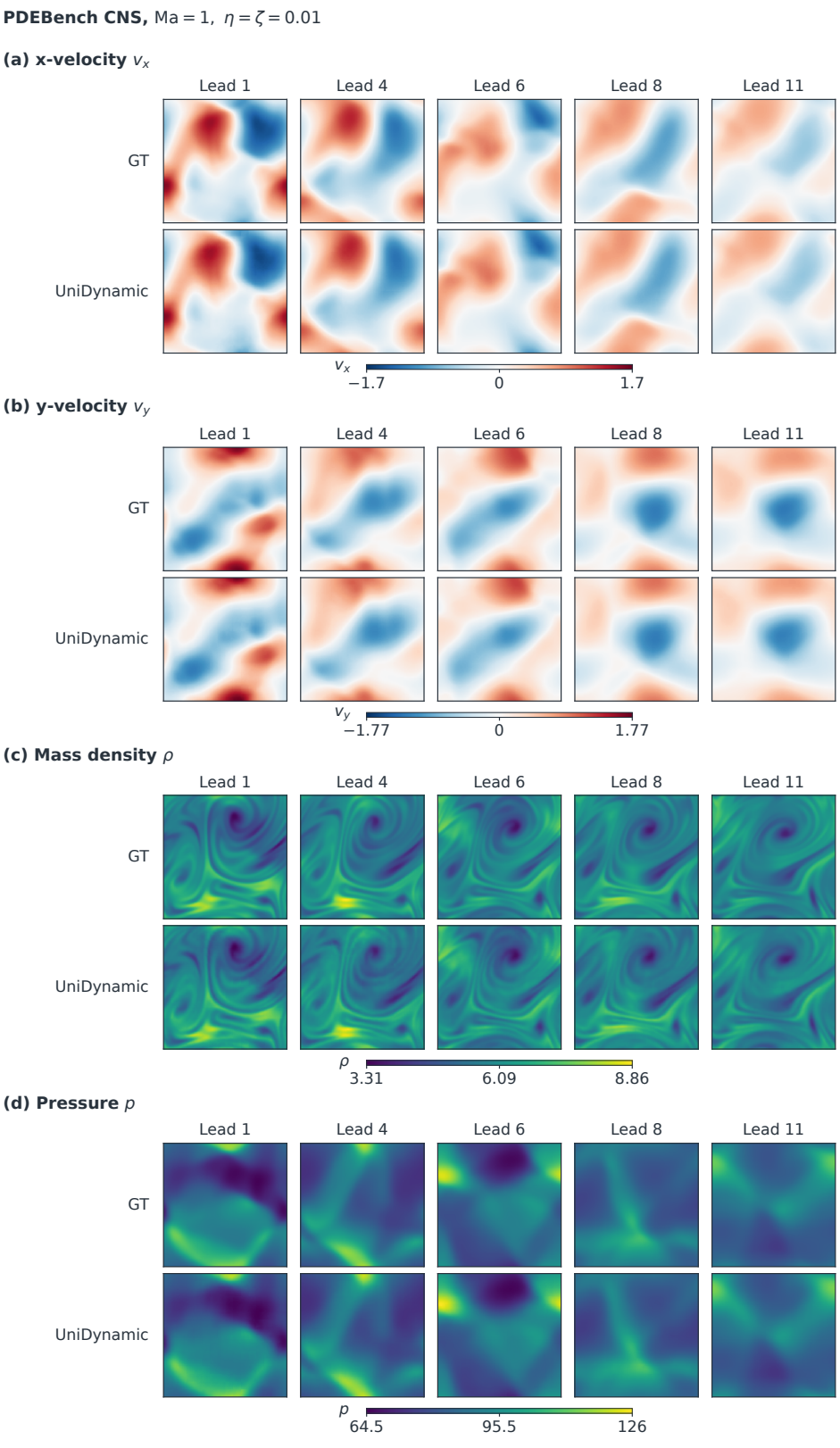

Figure 10: **PDEBench CNS** ($\mathrm{Ma} = 1$, $\eta = \zeta = 0.01$)**.** GT (top) and UNIDYNAMIC (bottom) for every labeled physical variable at the indicated rollout leads.

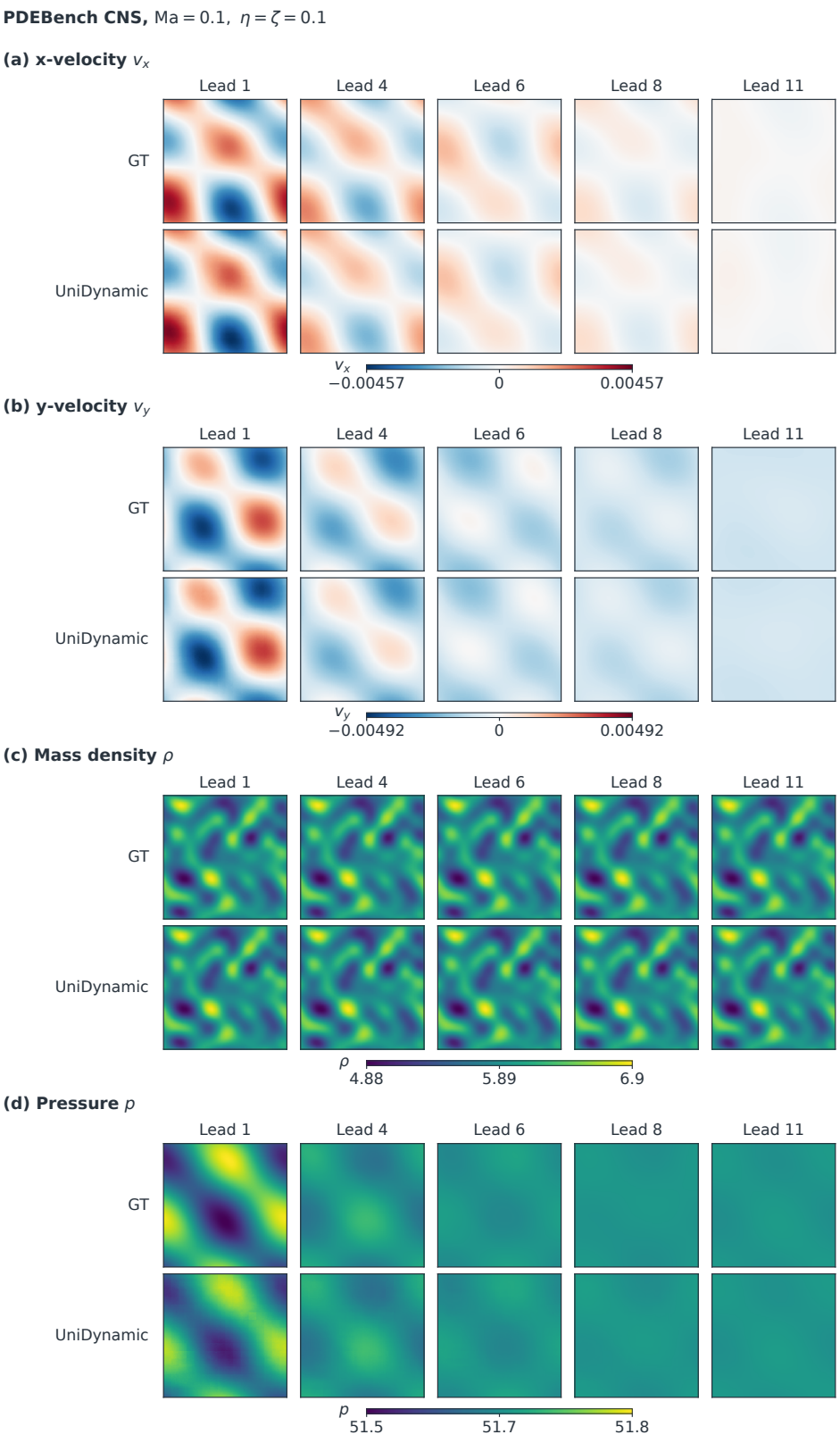

Figure 11: **PDEBench CNS** ($\mathrm{Ma} = 0.1$, $\eta = \zeta = 0.1$)**.** GT (top) and UNIDYNAMIC (bottom) for every labeled physical variable at the indicated rollout leads.

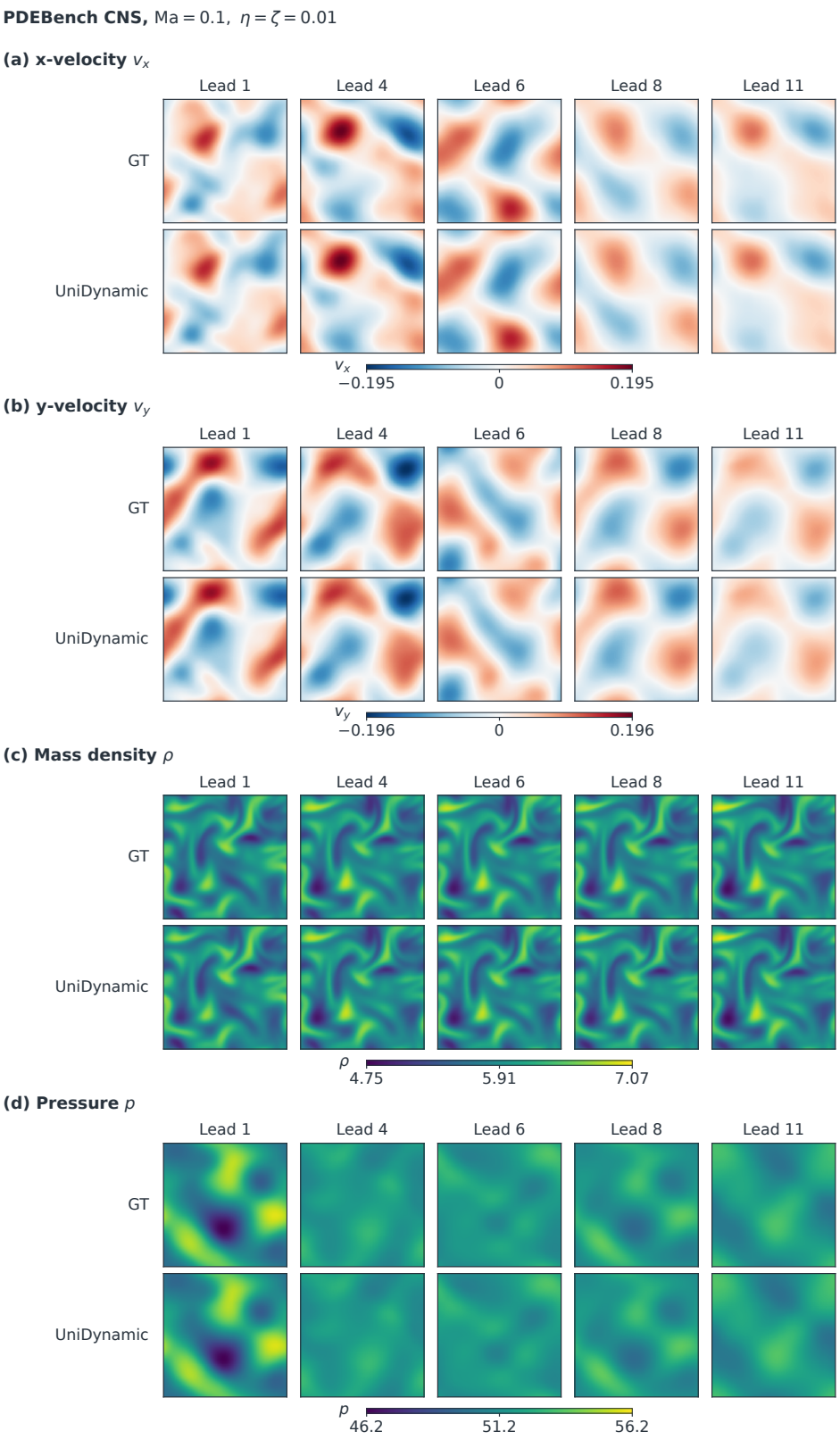

Figure 12: **PDEBench CNS** ($\mathrm{Ma} = 0.1$**,** $\eta = \zeta = 0.01$**).** GT (top) and UNIDYNAMIC (bottom) for every labeled physical variable at the indicated rollout leads.

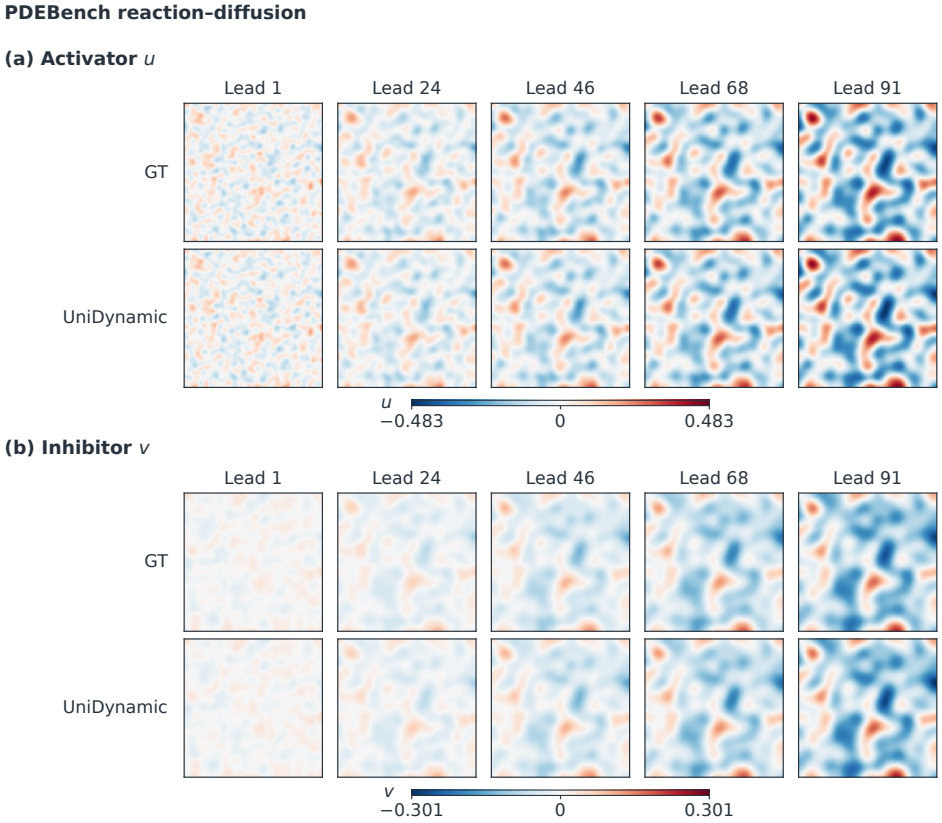

Figure 13: **PDEBench reaction–diffusion.** GT (top) and UNIDYNAMIC (bottom) for every labeled physical variable at the indicated rollout leads.

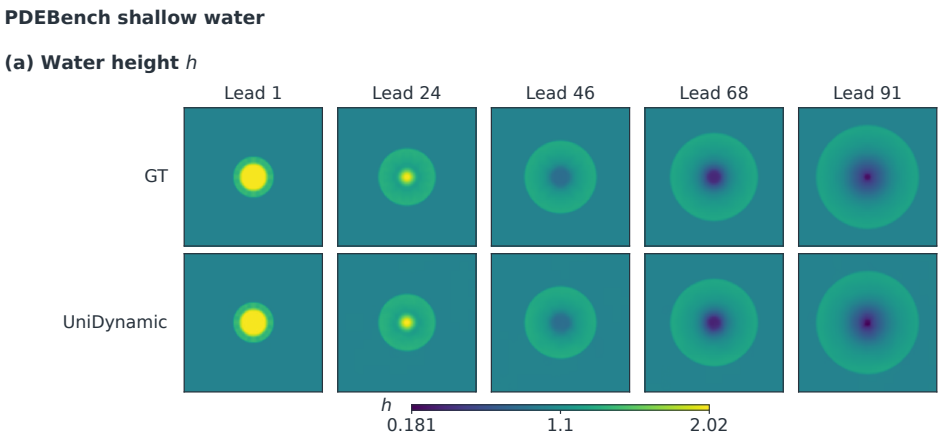

Figure 14: **PDEBench shallow water.** GT (top) and UNIDYNAMIC (bottom) for every labeled physical variable at the indicated rollout leads.

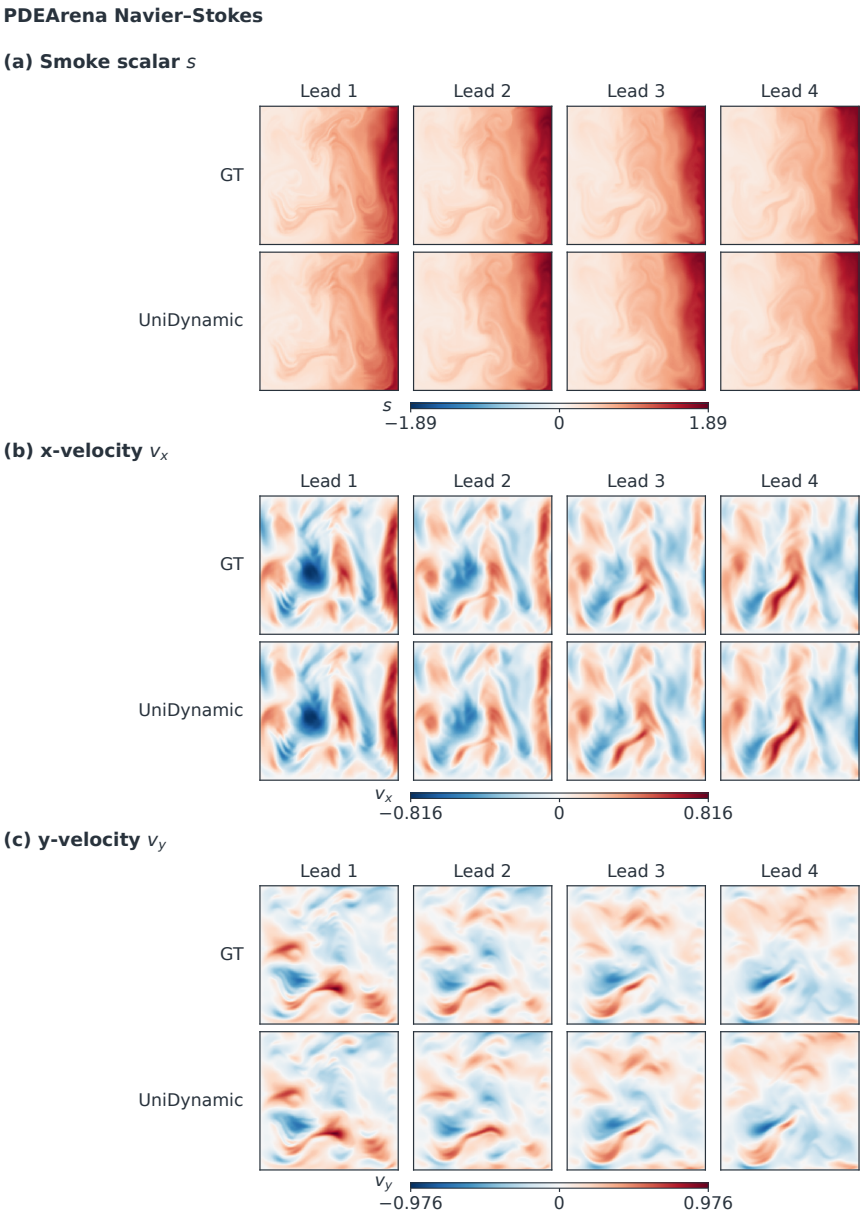

Figure 15: **PDEArena Navier–Stokes.** GT (top) and UNIDYNAMIC (bottom) for every labeled physical variable at the indicated rollout leads.

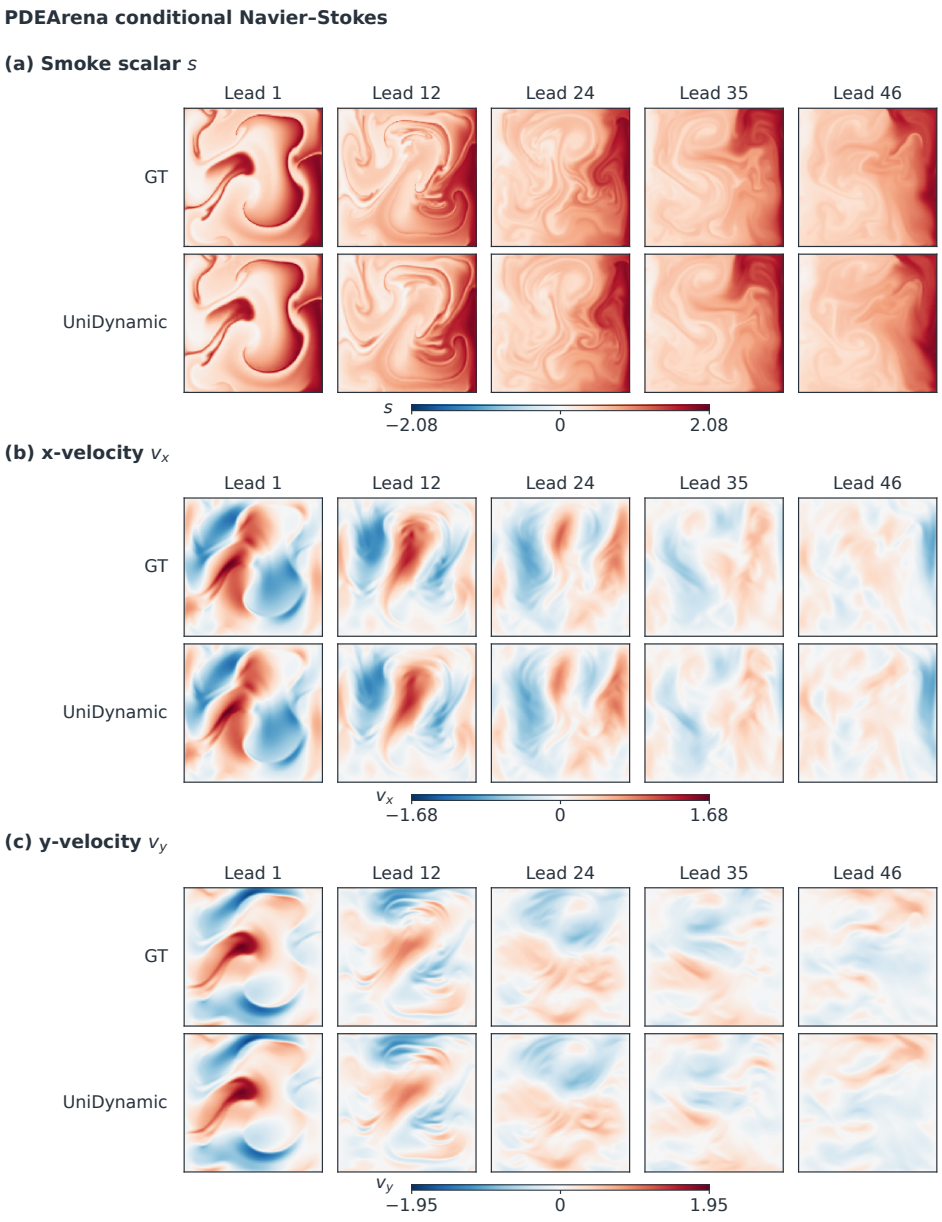

Figure 16: **PDEArena conditional Navier–Stokes.** GT (top) and UniDynamic (bottom) for every labeled physical variable at the indicated rollout leads.

E.2 STOCHASTIC CONDITIONAL SAMPLES

Figures 17–19 show three DNS reference futures and three UNIDYNAMIC samples conditioned on the same ten-frame history for each stochastic system. We use the first reference history and the first three members of each ensemble, with 32 Euler steps per forecast frame. Rows show forecast leads 1, 5, 10, 15, and 20. Each column follows a fixed member's trajectory, with predictions fed back autoregressively. DNS futures and generated samples are independent conditional draws, not paired realizations. All physical channels are shown, with a shared color scale for each variable across the displayed members and forecast leads.

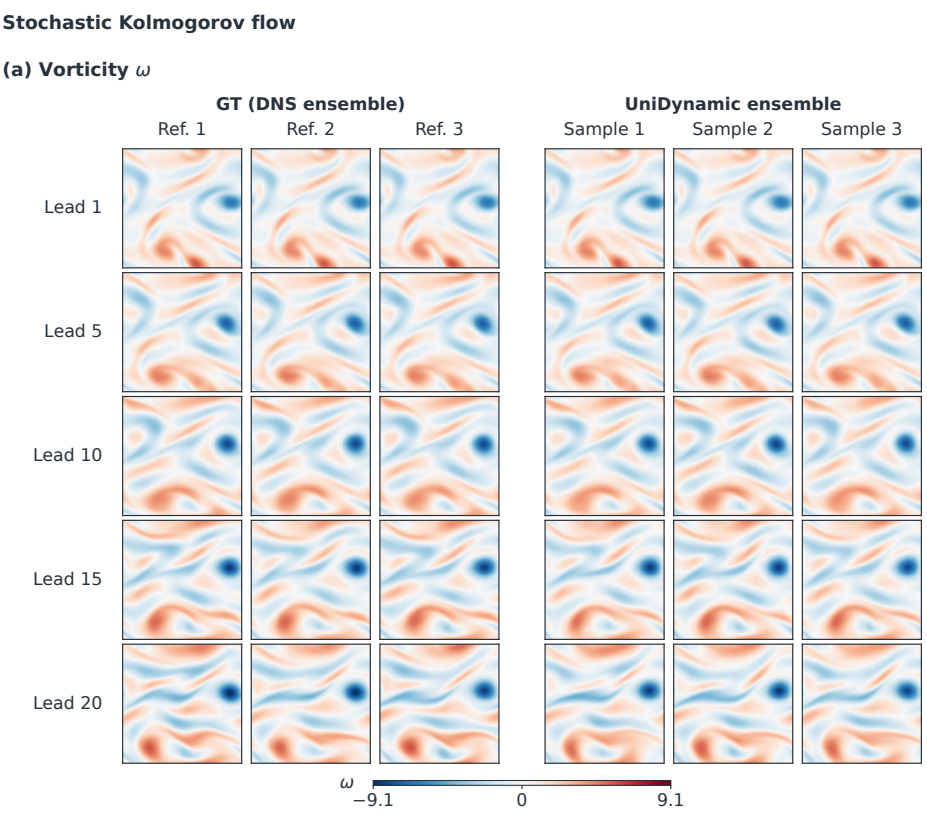

Figure 17: **Stochastic Kolmogorov flow.** Three DNS futures (left) and three UNIDYNAMIC samples (right) from the same history, with all physical channels. Columns retain member identity across leads. The ensembles are not paired.

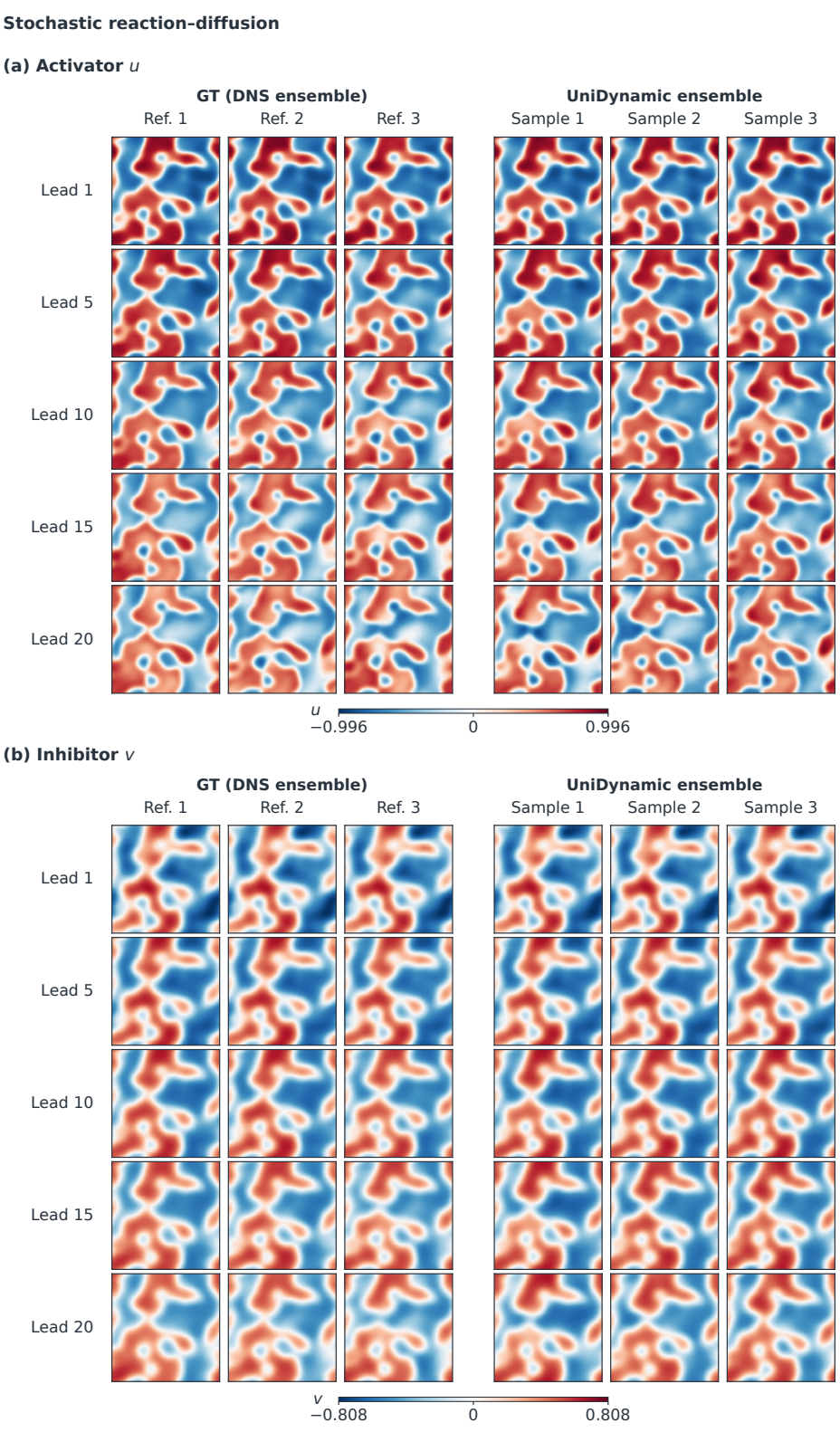

Figure 18: **Stochastic reaction–diffusion.** Three DNS futures (left) and three UNIDYNAMIC samples (right) from the same history, with all physical channels. Columns retain member identity across leads. The ensembles are not paired.

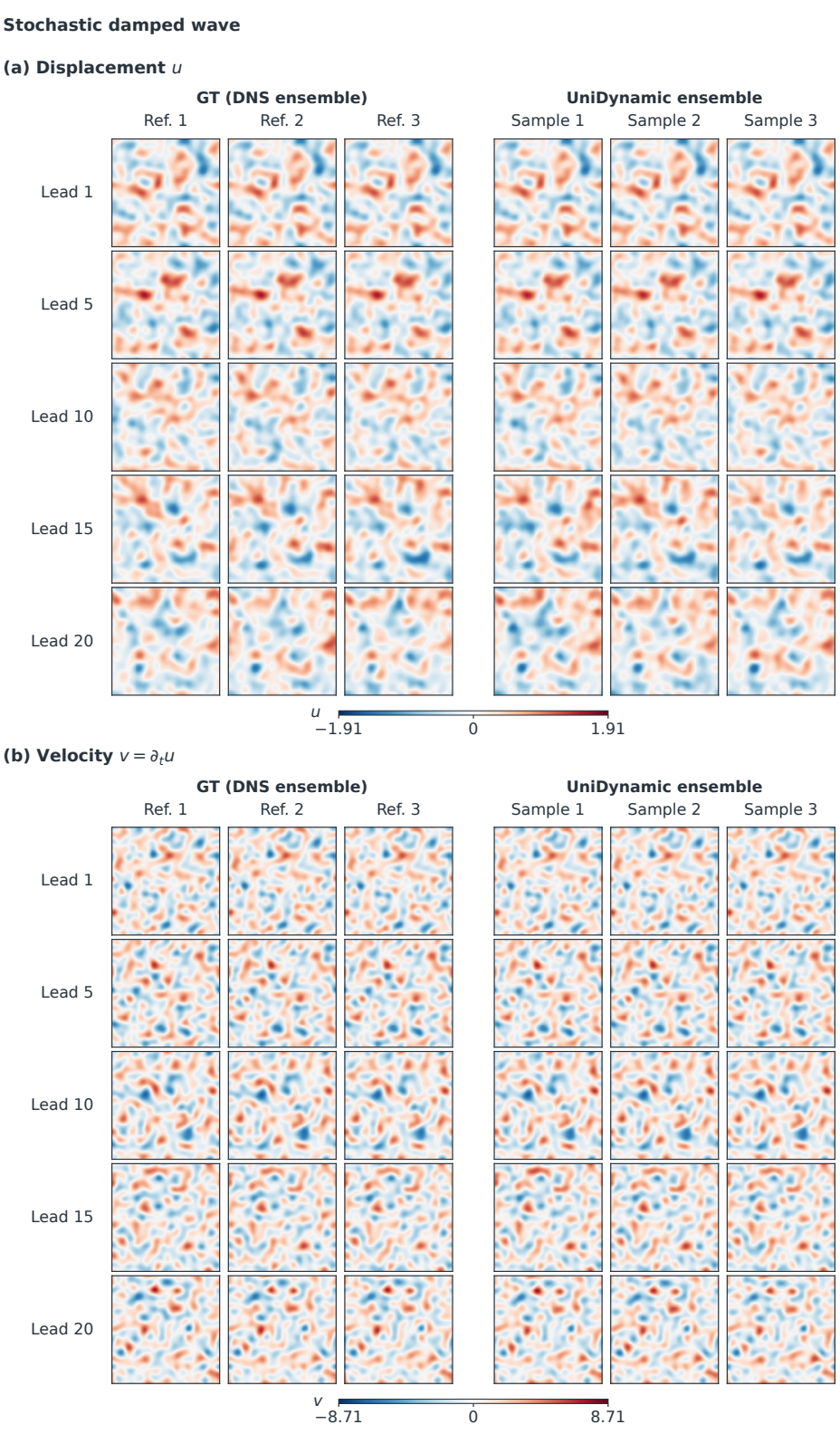

Figure 19: **Stochastic damped wave.** Three DNS futures (left) and three UniDynamic samples (right) from the same history, with all physical channels. Columns retain member identity across leads. The ensembles are not paired.

E.3 OUT-OF-DISTRIBUTION FINE-TUNING

Figures 20 and 21 show task-specific UNIDYNAMIC models fine-tuned on 128 target trajectories for the unseen Wave-Layer and stochastic Burgers equations. We display the first fixed test case or reference history for each task, using ten observed frames to initialize autoregressive prediction. Wave-Layer uses one Euler step per forecast frame over an 11-frame horizon. Burgers uses 32 Euler steps per frame over a 20-frame horizon. For Burgers, the first three DNS and generated members are displayed. Each column follows one member's trajectory. The two ensembles are independent conditional draws, not paired realizations. All physical output channels are shown, with one shared color scale per variable across GT, predictions, and forecast leads.

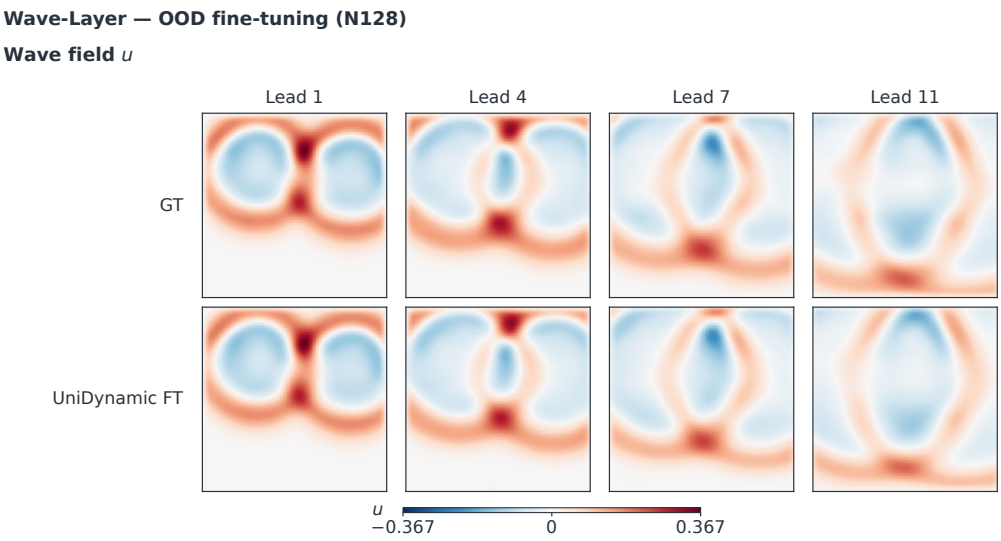

Figure 20: **OOD fine-tuning on Wave-Layer.** Ground truth (top) and fine-tuned UNIDYNAMIC predictions (bottom) of the wave field u at forecast leads 1, 4, 7, and 11.

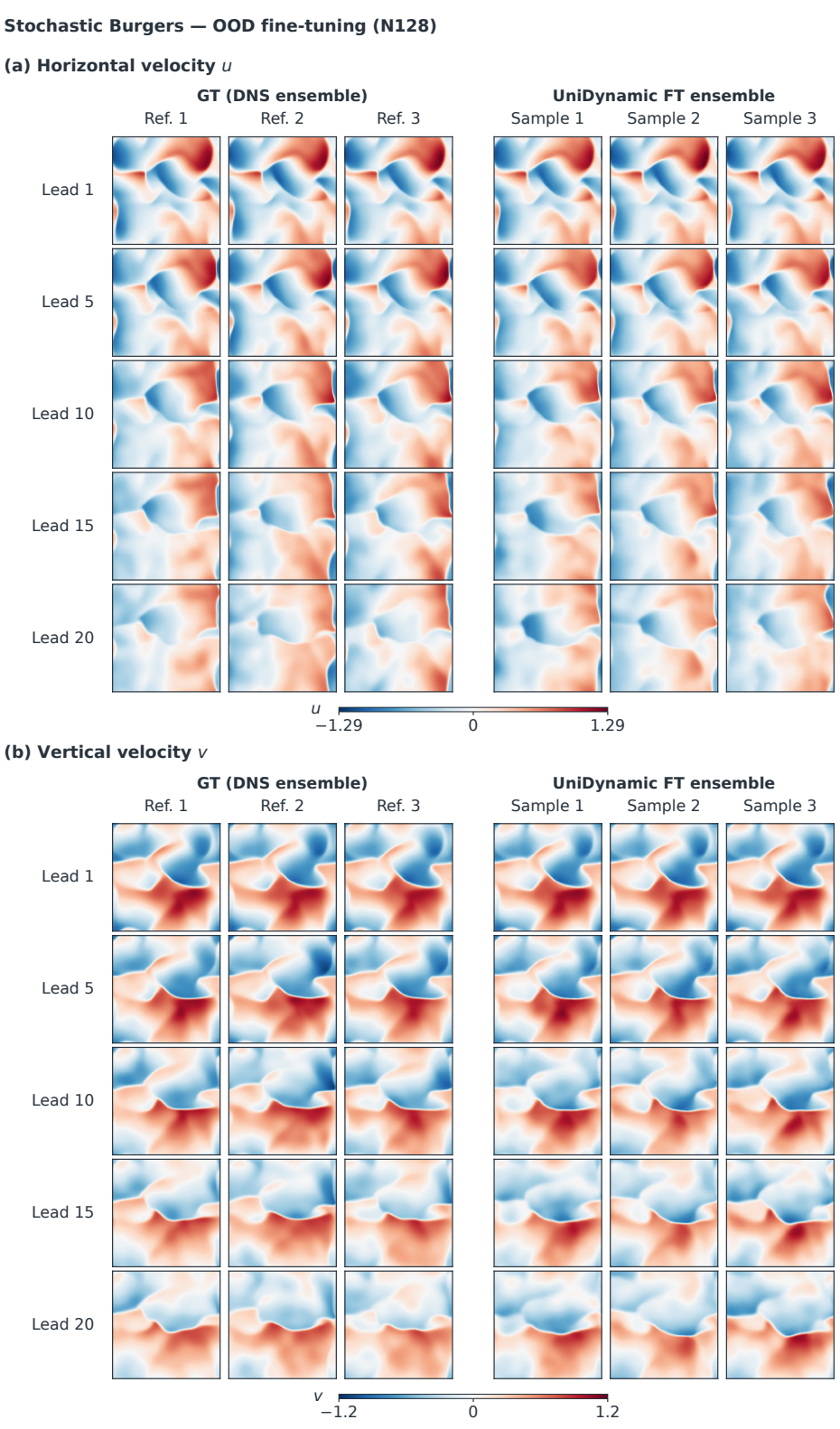

Figure 21: **OOD fine-tuning on stochastic Burgers.** Three DNS futures (left) and three fine-tuned UniDynamic samples (right), showing velocities u and v. Columns track fixed members. The two ensembles are not paired.

F Additional Long-Horizon Experiments

F.1 Evaluation and quantitative results

We extend stochastic forecasting to 50 autoregressive frames using the same frozen checkpoints and method-specific samplers as in the main stochastic comparison, without further training. Each rollout starts from ten observed frames. The long-horizon reference set contains 16 Kolmogorov histories and eight histories each for reaction–diffusion and damped waves, with 128, 64, and 64 DNS futures per history, respectively. All five methods generate 128 members per history. All leads, including 1–20, are evaluated on these same long-horizon histories.

We use the metrics defined in Section 5.3, averaging physical channels and then histories within each system. Table 11 reports lead-50 errors and the mean ED_n over leads 21–50. The three-system macro average weights the systems equally.

Table 11: **Long-horizon stochastic forecasting.** All metrics ↓. Bold marks the best value within each group. The second metric column averages ED_n over leads 21–50. The macro row group gives equal weight to the three systems.

Method	$\mathrm{ED}_n(50)$	$\overline{\mathrm{ED}}_n^{21:50}$	$E_\mu(50)$	$E_s(50)$	$E_{\mathrm{fluc}}(50)$
Kolmogorov flow					
DPOT-FM	0.0488	0.0680	0.3409	0.1809	0.2011
DPOT-CRPS	0.1364	0.1096	0.5866	0.3530	0.4422
GenCFD	0.0576	0.0786	0.3973	0.1764	0.1761
ACDM	0.3582	0.4889	0.9835	0.4056	0.4515
UniDynamic	**0.0245**	**0.0295**	**0.2425**	**0.1271**	**0.0994**
Reaction–diffusion					
DPOT-FM	0.0493	0.0364	0.2570	0.1514	0.1891
DPOT-CRPS	0.2394	0.1663	0.2944	0.8505	1.0103
GenCFD	0.0528	0.0435	0.2771	0.1912	0.2473
ACDM	0.1570	0.1314	0.3526	0.5324	0.7923
UniDynamic	**0.0233**	**0.0192**	**0.1625**	**0.1214**	**0.1086**
Damped wave					
DPOT-FM	0.0304	0.0219	0.1453	**0.1224**	**0.1158**
DPOT-CRPS	0.1238	0.0933	0.1359	0.7081	1.0289
GenCFD	0.2230	0.1371	0.4570	0.1479	0.4080
ACDM	0.6115	0.5151	0.8878	0.3521	1.1515
UniDynamic	**0.0142**	**0.0111**	**0.0848**	0.1284	0.1460
Three-system macro average					
DPOT-FM	0.0428	0.0421	0.2477	0.1516	0.1687
DPOT-CRPS	0.1665	0.1231	0.3390	0.6372	0.8271
GenCFD	0.1112	0.0864	0.3772	0.1719	0.2771
ACDM	0.3756	0.3785	0.7413	0.4300	0.7985
UniDynamic	**0.0206**	**0.0199**	**0.1633**	**0.1256**	**0.1180**

UniDynamic attains the lowest lead-50 ED_n and E_μ, and the lowest mean ED_n over leads 21–50, on all three systems. Relative to the strongest macro-average baseline, DPOT-FM, the lead-50 and leads-21–50 ED_n are reduced by 51.8% and 52.6%, respectively, using unrounded values. On damped waves, DPOT-FM has lower lead-50 E_s and E_{fluc}, while UniDynamic has lower marginal-distribution and conditional-mean errors.

F.2 Error trajectories over 50 forecast frames

Figure 22 shows all 50 forecast leads for the five methods on the complete long-horizon reference set. The dotted line marks the 20-frame evaluation horizon used in the main comparison. These curves show how the endpoint and late-horizon summaries in Table 11 arise over the rollout.

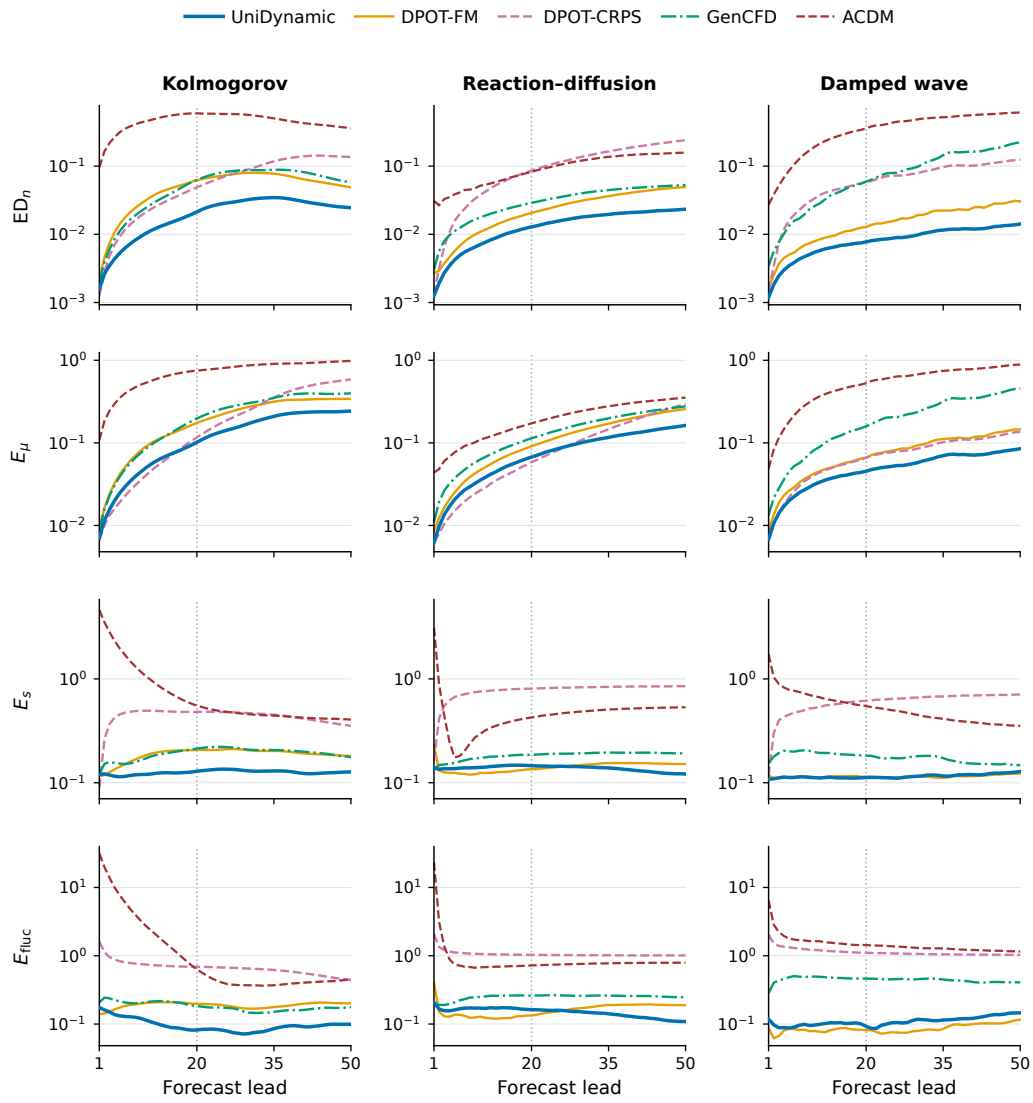

Figure 22: **Long-horizon forecasting errors.** Columns correspond to the three systems. Rows show ED_n, E_μ, E_s, and E_{fluc}. Curves average all available histories after physical-channel averaging. Error axes are logarithmic and forecast lead is linear. Every evaluated lead is shown without smoothing.

F.3 AUTOREGRESSIVE FIELD EVOLUTION

Figures 23–25 show the first fixed long-reference history for each system. Three DNS futures and three UNIDYNAMIC members are displayed at leads 1, 10, 20, 35, and 50, with all physical channels retained. Each column follows one member's trajectory. DNS and model members are independent conditional samples, not paired outcomes. Each variable uses a common color scale across the displayed fields.

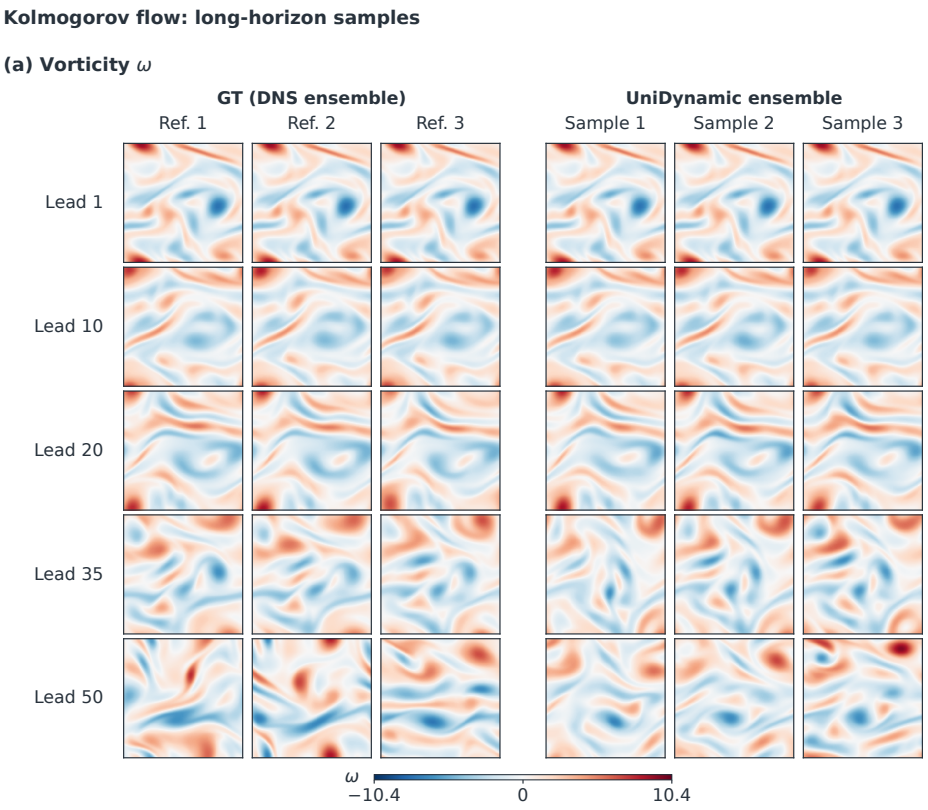

Figure 23: **Long-horizon Kolmogorov flow.** Three DNS futures (left) and three UNIDYNAMIC samples (right). Columns track fixed members over the indicated forecast leads.

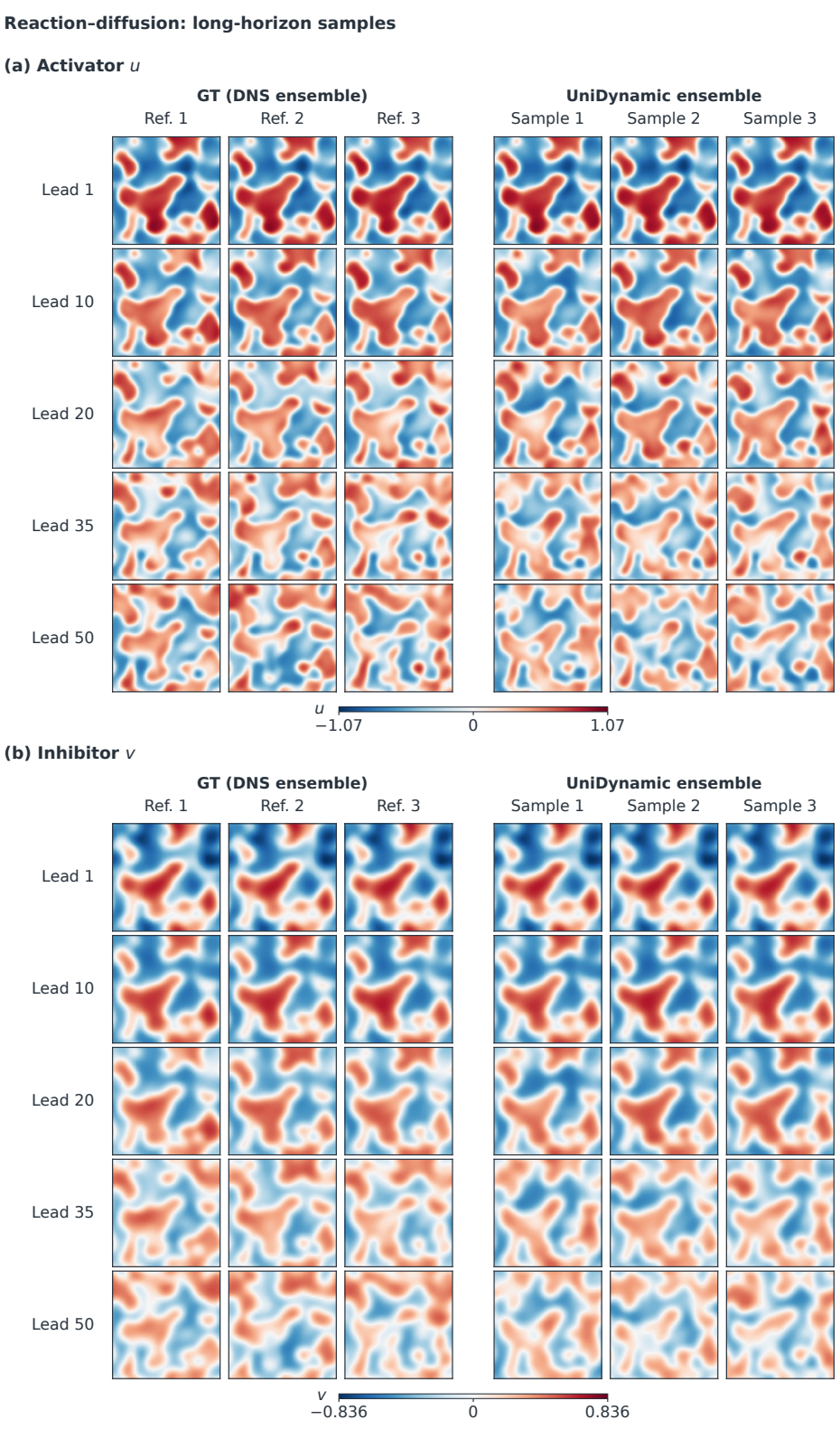

Figure 24: **Long-horizon reaction–diffusion.** Three DNS futures (left) and three UNIDYNAMIC samples (right), showing both physical channels. Columns track fixed members. The two ensembles are not paired.

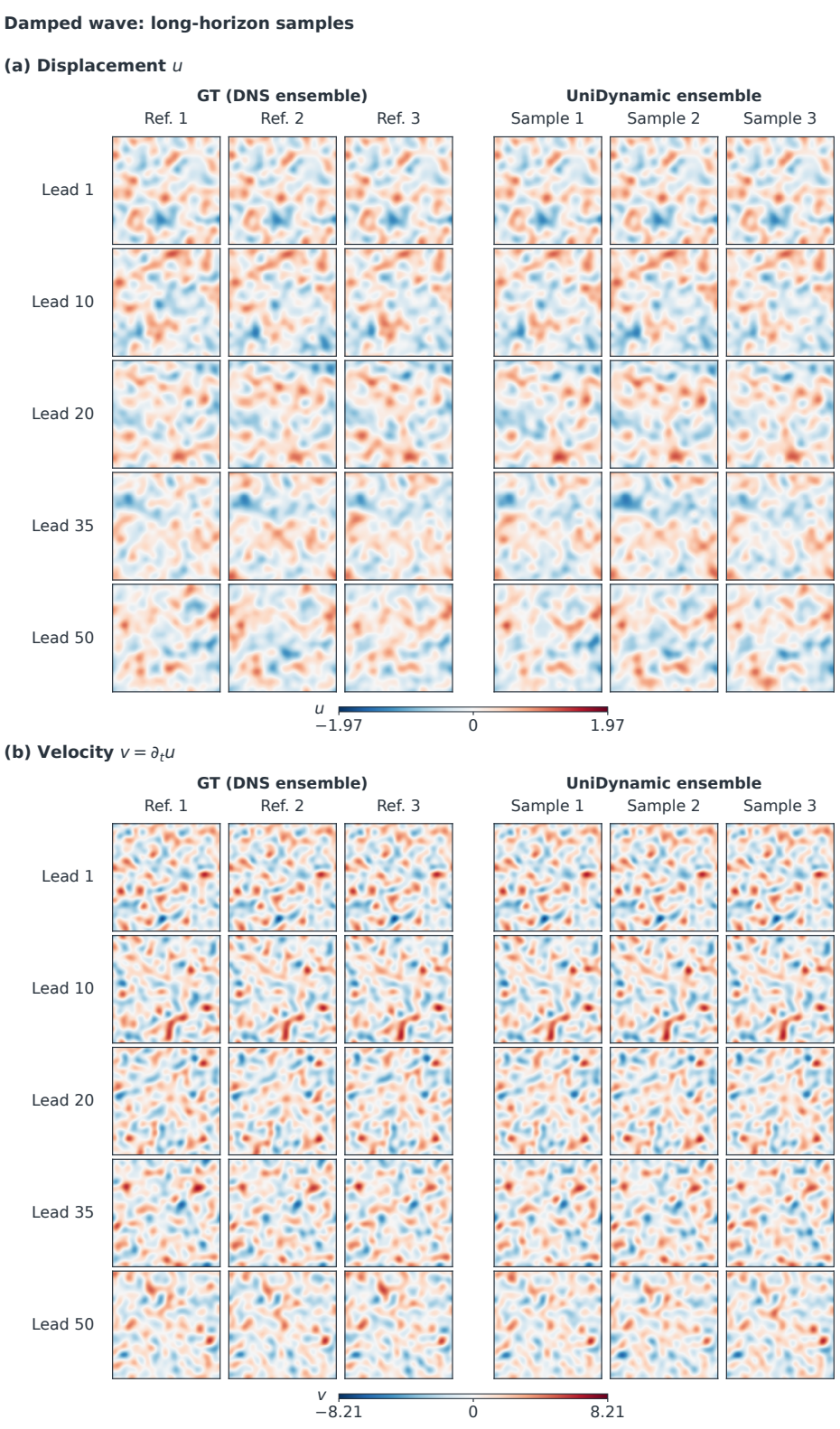

Figure 25: **Long-horizon damped wave.** Three DNS futures (left) and three UniDynamic samples (right), showing displacement and velocity. Columns track fixed members. The two ensembles are not paired.

G Additional Deep-Ensemble Experiments

G.1 Evaluation and quantitative results

We evaluate conditional generation against the benchmark's deep reference ensembles. For UniDynamic, DPOT-FM, DPOT-CRPS, GenCFD, and ACDM, we reuse the frozen checkpoints and method-specific samplers from the main stochastic comparison, without further training. The deep set contains eight Kolmogorov histories and four histories each for reaction–diffusion and damped waves, with 512, 256, and 256 DNS futures per history, respectively. Each rollout starts from ten observed frames. All methods generate 128 members over 20 autoregressive forecast frames. Thus, deep provides a larger DNS ensemble per history than the 128-member wide references used in the main comparison.

We use the metrics defined in Section 5.3, averaging physical channels and then histories within each system. Table 12 reports lead-20 errors and mean ED_n over leads 1–20. The three-system macro average weights the systems equally.

Table 12: **Deep-ensemble stochastic forecasting.** All metrics ↓. Bold marks the lowest value within each group. Every model generates 128 members per history. DNS reference ensembles contain 512, 256, and 256 members for the three systems, respectively. The macro group weights the systems equally.

Method	$\mathrm{ED}_n(20)$	$\overline{\mathrm{ED}}_n^{1:20}$	$E_\mu(20)$	$E_s(20)$	$E_{\mathrm{fluc}}(20)$
Kolmogorov flow					
DPOT-FM	0.0535	0.0237	0.1595	0.1756	0.1732
DPOT-CRPS	0.0355	0.0181	0.0896	0.4519	0.6759
GenCFD	0.0337	0.0163	0.1397	0.1708	0.1648
ACDM	0.4866	0.3612	0.6508	0.5185	0.5921
UniDynamic	**0.0083**	**0.0050**	**0.0623**	**0.1038**	**0.0948**
Reaction–diffusion					
DPOT-FM	0.0172	0.0086	0.0893	0.1029	0.1072
DPOT-CRPS	0.0800	0.0377	**0.0460**	0.8037	1.0322
GenCFD	0.0252	0.0156	0.1079	0.1742	0.2662
ACDM	0.0970	0.0629	0.1985	0.4171	0.7233
UniDynamic	**0.0076**	**0.0040**	0.0542	**0.0919**	**0.0824**
Damped wave					
DPOT-FM	0.0090	0.0051	0.0595	**0.0815**	**0.0640**
DPOT-CRPS	0.0529	0.0301	0.0553	0.6123	1.1104
GenCFD	0.0558	0.0263	0.1546	0.1713	0.4719
ACDM	0.3735	0.1988	0.5478	0.5310	1.3982
UniDynamic	**0.0045**	**0.0029**	**0.0361**	0.0836	0.0916
Three-system macro average					
DPOT-FM	0.0266	0.0125	0.1028	0.1200	0.1148
DPOT-CRPS	0.0561	0.0286	0.0637	0.6226	0.9395
GenCFD	0.0382	0.0194	0.1341	0.1721	0.3009
ACDM	0.3190	0.2076	0.4657	0.4889	0.9045
UniDynamic	**0.0068**	**0.0040**	**0.0509**	**0.0931**	**0.0896**

UniDynamic achieves the lowest lead-20 and rollout-averaged ED_n on all three systems. Relative to DPOT-FM, the strongest macro-average baseline on these metrics, the reductions are 74.4% and 68.0%, respectively. DPOT-CRPS has lower lead-20 E_μ on reaction–diffusion. DPOT-FM has lower lead-20 E_s and E_{fluc} on damped waves.

G.2 Error trajectories over 20 forecast frames

Figure 26 shows the four metrics at every forecast lead on the complete deep reference set. The per-frame sampling budgets remain 32 for UniDynamic and DPOT-FM, 1 for DPOT-CRPS, 128 for GenCFD, and 20 for ACDM.

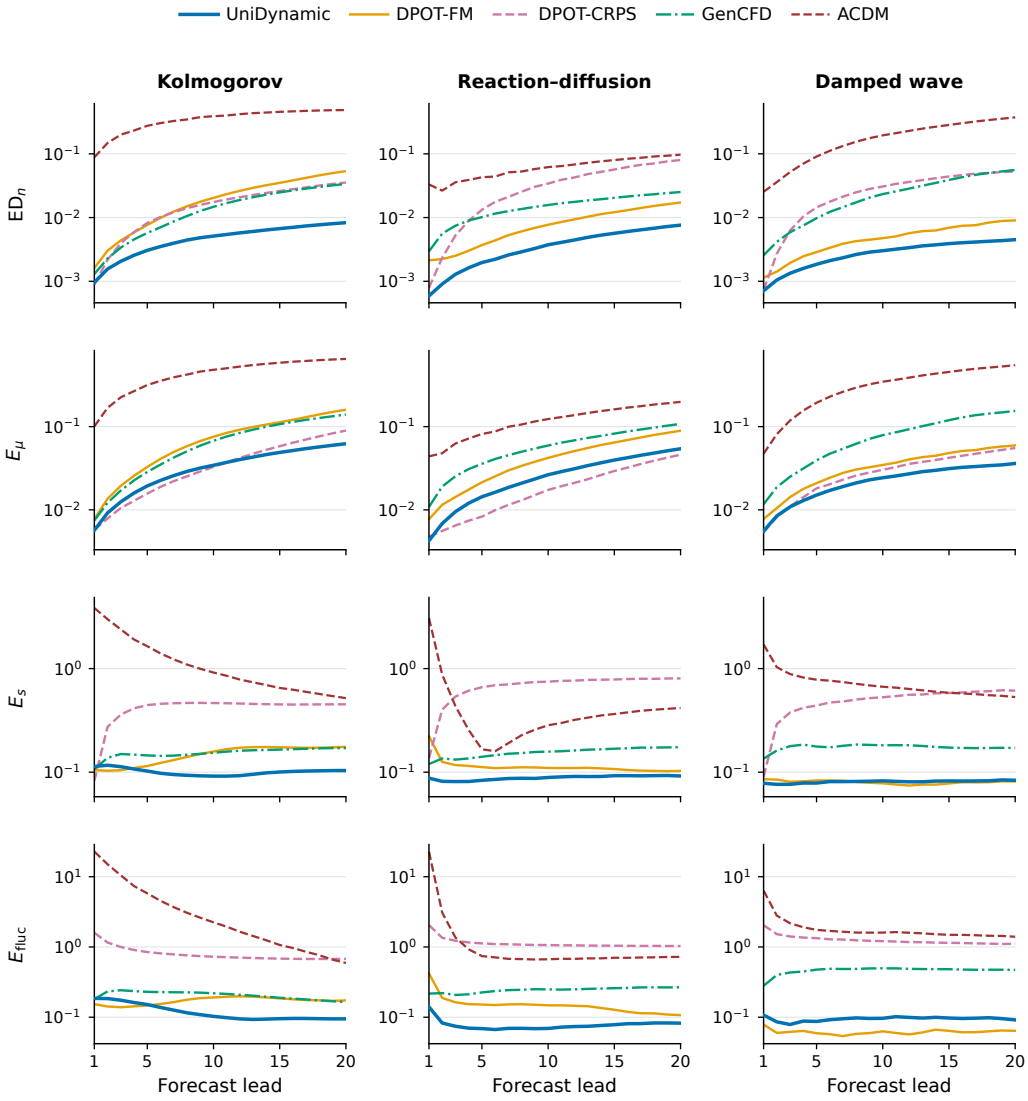

Figure 26: **Deep-ensemble forecasting errors.** Columns correspond to the three systems. Rows show ED_n, E_μ, E_s, and E_{fluc}. Curves average all deep histories after physical-channel averaging. Forecast lead is linear and error axes are logarithmic. Every evaluated lead is shown without smoothing.

G.3 CONDITIONAL-ENSEMBLE MEAN AND SPREAD FIELDS

Figures 27–29 show the first fixed deep history for each system at leads 1, 10, and 20. The mean and standard-deviation fields use all 128 generated members and all 512 or 256 DNS futures, with every physical channel retained. Standard deviations use the unbiased sample-variance convention. For each variable and statistic, a common color scale is used across DNS, UNIDYNAMIC, and the displayed leads.

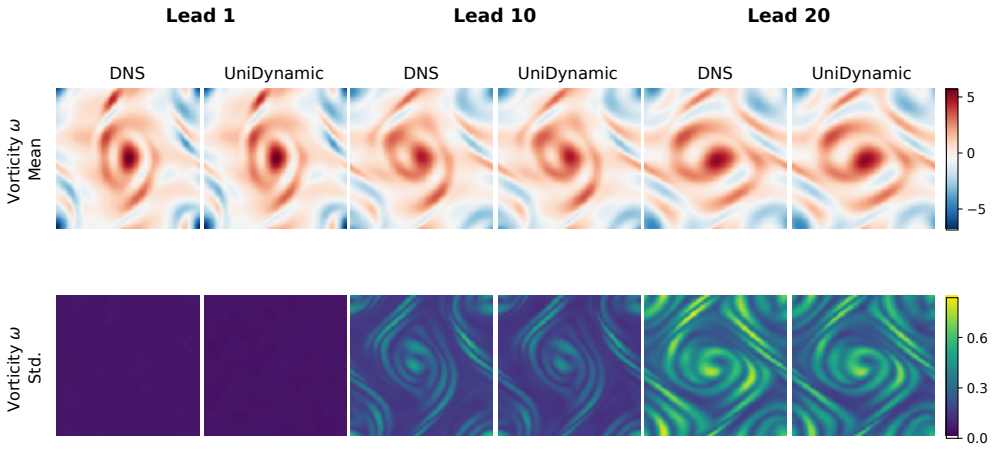

Figure 27: **Deep-ensemble Kolmogorov flow.** Ensemble mean and standard deviation of vorticity. Each lead pairs the 512-member DNS summary with the 128-member UNIDYNAMIC summary conditioned on the same history.

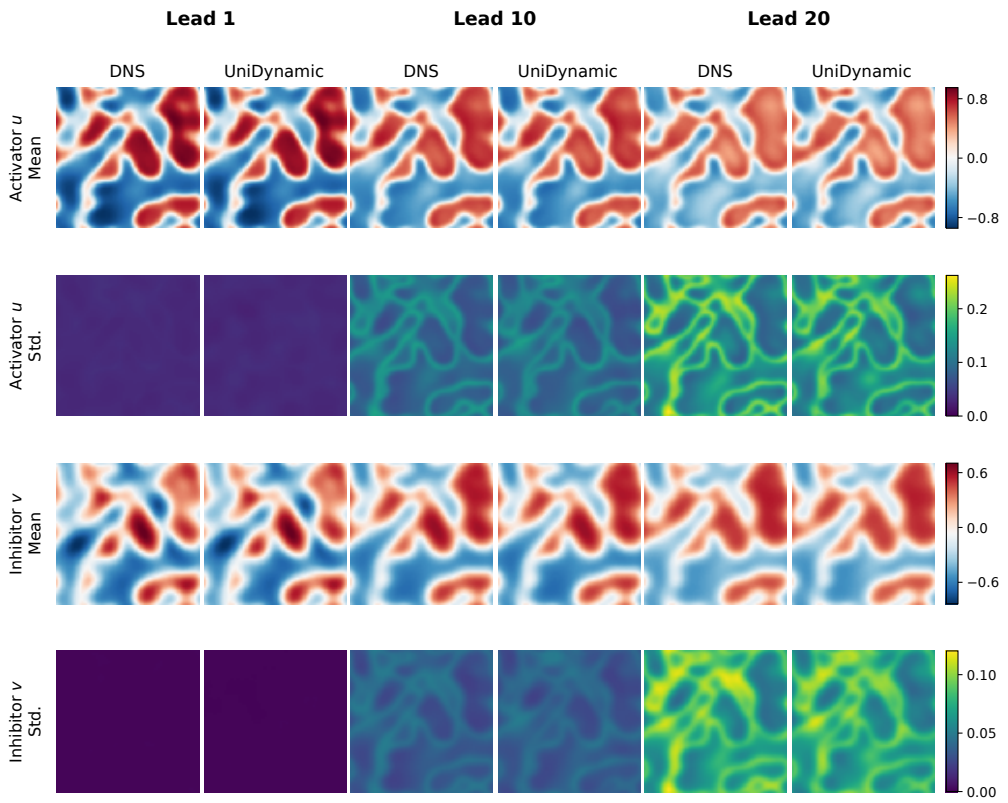

Figure 28: **Deep-ensemble reaction–diffusion.** Ensemble means and standard deviations of the activator and inhibitor fields, using 256 DNS futures and 128 UniDynamic samples. Both ensembles share the observed history.

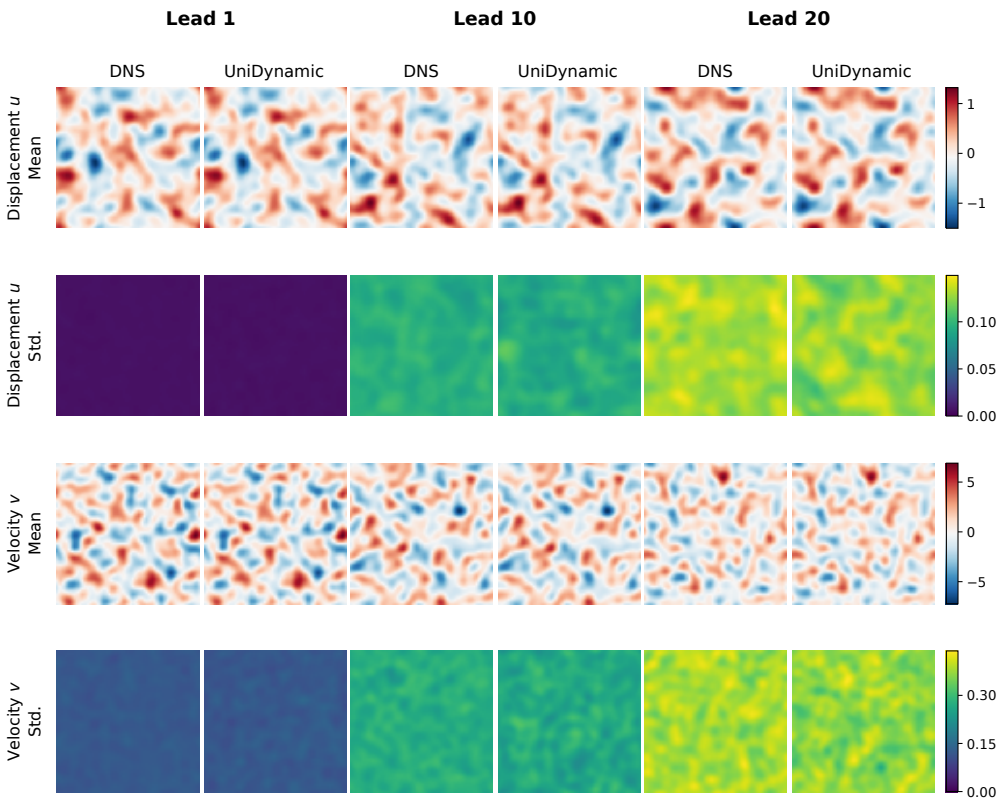

Figure 29: **Deep-ensemble damped wave.** Ensemble means and standard deviations of displacement and velocity, using 256 DNS futures and 128 UNIDYNAMIC samples. Both ensembles share the observed history.

G.4 SENSITIVITY TO DNS REFERENCE-ENSEMBLE SIZE

We recompute the lead-20 metrics using nested subsets of 64, 128, and 256 DNS futures per history, and all 512 futures for Kolmogorov. The deep histories and all 128 generated members remain fixed. This isolates sensitivity to the empirical DNS reference size without additional model inference. UNIDYNAMIC retains the lowest ED_n among the five methods at every evaluated reference size on all three systems.

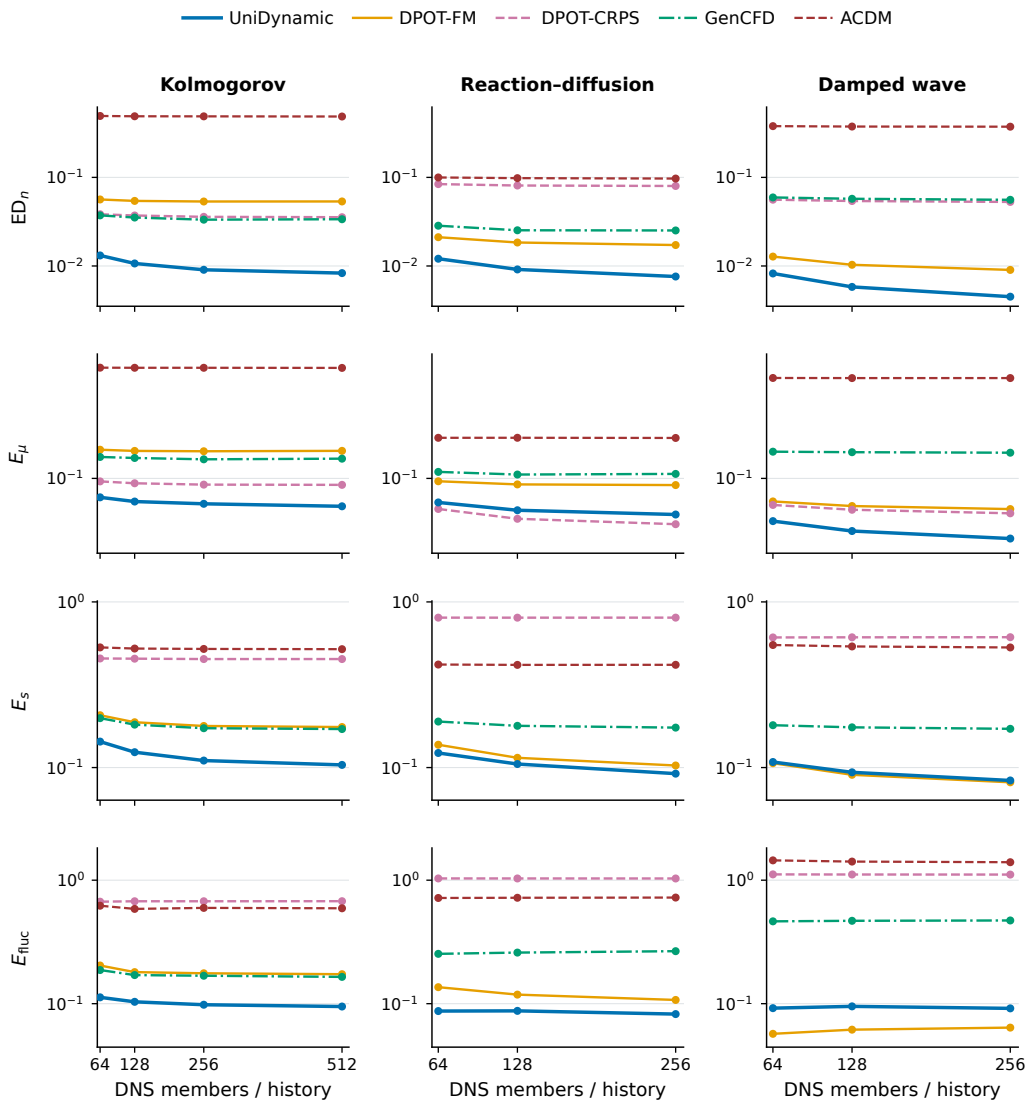

Figure 30: **Same-history DNS reference-size sensitivity.** Lead-20 errors as a function of the number of DNS futures per history. Curves use nested prefixes of the same deep reference ensemble and hold all generated samples fixed. Reference size is linear and error axes are logarithmic.