CoMNO: Commutator Memory Neural Operators for Long-Horizon PDE Prediction

Anonymous authors
Paper under double-blind review

Abstract

Solving nonlinear partial differential equations (PDEs) over long time horizons is computationally expensive using traditional numerical methods. While neural operators provide an efficient alternative, they suffer from the accumulation of approximation errors along the predicted trajectory, typically caused by the failure to capture fine-scale details. In this work, we design a Commutator Memory Neural Operator that identifies fine-scale dynamical details by commuting two operations and leverages these details to enhance the neural operator's capability in capturing fine-scale features over long time horizons. Specifically, we identify a performance gap between two operations: downsampling the high-resolution field before rollout prediction versus performing rollout prediction before downsampling. We observe that fine-scale dynamical details are often encoded in this gap. By commuting these two operations, the fine-scale details removed during downsampling can be captured. Furthermore, we employ a memory module to store these fine-scale features along the trajectory and retrieve them during rollout, generating spatially adaptive corrections that enhance long-horizon fidelity. Empirical results across five systems demonstrate that CoMNO improves prediction accuracy and long-horizon rollout stability. The complete model implementation is provided in the supplementary material.

1 Introduction

High-fidelity simulation of nonlinear partial differential equations (PDEs) is central to climate modeling, turbulence analysis, and engineering design Lam et al. (2023); Kochkov et al. (2021); Pfaff et al. (2021). Conventional numerical solvers require fine grids and repeated time integration, making long-horizon simulation computationally expensive. To address this efficiency bottleneck, neural operators offer an efficient alternative by learning mappings between physical fields Li et al. (2021); Kovachki et al. (2023).

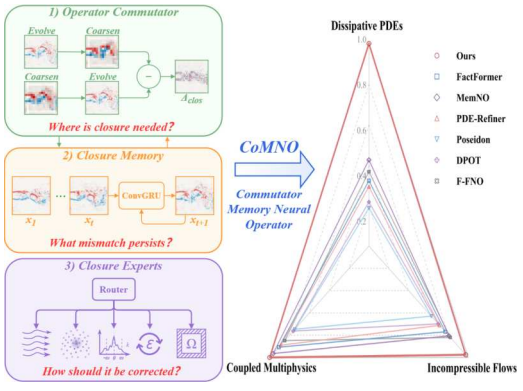

Figure 1: CoMNO's core mechanisms and gains in long-horizon prediction across diverse PDE systems.

Despite remarkable progress, neural operators typically suffer from performance deterioration during long-time rollout prediction, due to the accumulation of approximation errors along the predicted trajectory Lippe et al. (2023); Koehler et al. (2024); Wang et al. (2026b). This difficulty can be viewed as a finite-resolution closure problem Gupta & Lermusiaux (2021); Buitrago Ruiz et al. (2025). To formalize this, let u_t denote the high-resolution field, Φ the neural operator that maps it forward in time, and D_0 a downsampling operator that extracts the coarse observation $x_t = D_0 u_t$. Consider the "predict-then-downsample" pipeline: the neural operator first predicts the next high-resolution state $\Phi(u_t)$, which is then downsampled to yield $D_0\Phi(u_t)$. This composite step can be expressed as:

$$x_{t+1} = D_0\Phi(u_t) = x_t + \Delta_{\text{res}}(x_t) + \mathcal{R}(x_t, \mathcal{H}_t), \tag{1}$$

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.

where the right-hand side is interpreted as an update from the current coarse state x_t. The term $\Delta_{\mathrm{res}}(x_t)$ denotes the one-step state increment from x_t predicted solely from the current coarse-scale state, while the term $\mathcal{R}(x_t, \mathcal{H}_t)$ captures the residual influence of fine-scale structures that were discarded by downsampling but continue to affect the future coarse evolution. This residual correction is necessary because the coarse state x_t does not uniquely determine the future. The history $\mathcal{H}_t$ thus supplies the missing information needed to infer this fine-scale influence.

The above formulation suggests two key requirements for addressing this problem: first, one must identify whether the fine-scale information lost during downsampling actually matters for the future coarse rollout prediction; second, if it does, one must retain and utilize this information over time to enhance prediction fidelity. Existing approaches tackle this challenge from various angles. Multiscale methods refine spatial representation to preserve finer structures, memory-based models augment temporal modeling to capture long-range dependencies, and iterative correction schemes refine predictions through repeated feedback Lippe et al. (2023); Buitrago Ruiz et al. (2025). However, these methods typically infer missing dynamics from coarse states or latent features without explicitly tracing what information disappears during downsampling and how its removal changes future rollout predictions.

In this work, we find that the fine-scale influence underlying $\mathcal{R}(x_t, \mathcal{H}_t)$ can be identified by commuting rollout prediction and downsampling. For each scale s, let D_s denote the downsampling operator, and apply the same neural operator Φ in both paths. We define the commutator residual at time t and scale s as

$$c_t^{(s)} = D_s\Phi(u_t) - \Phi(D_s u_t), \tag{2}$$

The first path evolves u_t before removing scale s, whereas the second removes that scale before evolution. Their discrepancy therefore measures how the discarded fine-scale structures alter the predicted coarse state, providing direct evidence for their residual contribution to future coarse dynamics.

To compensate for fine-scale effects lost during coarse-resolution rollout, we introduce the Commutator Memory Neural Operator (CoMNO). Its key idea is that commutator residuals reveal where unresolved structures alter the predicted coarse evolution and how strongly they do so. A single residual captures this influence only at the current step, although its effect may persist and reappear later. CoMNO therefore stores residual evidence along the trajectory, integrates it across scales and time, and uses the accumulated history to determine where and how strongly future predictions should be corrected. In this way, CoMNO turns transient cross-scale discrepancies into spatially adaptive corrections, reducing the repeated loss of unresolved dynamics and stabilizing long-horizon rollout.

Our contributions are summarized as follows:

1. We identify fine-scale dynamical details through the prediction gap between downsampling before rollout prediction and performing rollout prediction before downsampling.
2. We introduce CoMNO, which stores these details in recurrent Closure Memory and retrieves them to generate spatially adaptive corrections throughout autoregressive rollout.
3. Across three APEBench systems and two RealPDEBench systems Koehler et al. (2024); Hu et al. (2026), CoMNO consistently improves long-horizon accuracy, rollout stability, spectral fidelity, and physical consistency over strong neural PDE baselines.

2 Related Work

Neural Operators and Long-Horizon PDE Prediction. Neural operators learn mappings between function spaces using integral kernels, spectral transforms, multiscale convolutions, attention, and graph or coordinate representations Lu et al. (2021); Kovachki et al. (2023); Li et al. (2021); Tran et al. (2023); Rahman et al. (2023); Raonić et al. (2023); Cao (2021); Li et al. (2023); Hao et al. (2024); Herde et al. (2024); Hao et al. (2023); Serrano et al. (2023); Pfaff et al. (2021); Brandstetter et al. (2022). Despite strong short-horizon accuracy Takamoto et al. (2022), autoregressive rollout accumulates distribution shift, phase error, and spectral distortion List et al. (2025); Lippe et al. (2023); Kohl et al. (2026). Multi-step training, physics-informed objectives, refinement, and probabilistic prediction improve stability List et al. (2025); Li et al. (2024); Lippe et al. (2023), yet do not localize unresolved corrections. CoMNO derives an online spatial signal for this purpose.

Closure Modeling and Operator Non-Commutation. Large-eddy simulation represents unresolved scales through subgrid closures Zhao et al. (2023); Sun et al. (2024); Johnson (2022), while data-driven methods learn subgrid stresses, fluxes, or coarse-grid corrections Maulik et al. (2019); Beck et al. (2019); Kochkov et al. (2021); Srinivasan et al. (2024); van Gastelen et al. (2024); Agdestein & Sanderse (2025). Coarsening and nonlinear evolution generally do not commute, yielding state-dependent residuals across space and resolution Sanderse et al. (2025); van Gastelen et al. (2024); Agdestein & Sanderse (2025); Yalla et al. (2021), and coarse chaotic states may admit nonunique deterministic closures Wang et al. (2026a). Prior work mainly uses non-commutation for analysis or supervised closure learning. CoMNO evaluates it online to allocate corrections where cross-resolution inconsistency appears.

Non-Markovian Reduction and Memory-Augmented Dynamics. The Mori-Zwanzig formalism decomposes projected dynamics into Markovian, history-dependent, and unresolved fluctuation terms Gilani et al. (2021); Lin & Lu (2021); Falkena et al. (2019). Finite-memory closures and reduced models capture persistent unresolved effects Parish & Duraisamy (2017a;b); Lin et al. (2021); Tian et al. (2021); Wang et al. (2020), while recurrent and reservoir models retain trajectory history Vlachas et al. (2018); Pathak et al. (2018); Vlachas et al. (2020), including MemNO Buitrago Ruiz et al. (2025). These methods update memory from states without measuring scale mismatch. CoMNO retains commutator evidence in Closure Memory to guide spatially adaptive corrections through Closure Experts.

3 METHODOLOGY

3.1 PROBLEM FORMULATION AND OVERVIEW

CoMNO learns the history-dependent closure in Eq. (1) by comparing predictions across resolutions and retaining the resulting evidence over time. Following the Introduction, the Current State is the coarse observation $x_t = D_0 u_t$, and the Coarse History provides the temporal context stored in h_t. Figure 2 shows the computation: the Operator Commutator measures how coarsening changes a prediction, Closure Memory retains these differences along the trajectory, and Closure Experts use this information to correct the Neural Operator's base increment.

3.2 OPERATOR COMMUTATOR

The Operator Commutator measures where and how spatial coarsening changes the forecast, providing evidence for the closure correction. It applies the comparison in Eq. (2) at observable resolutions: the current field x_t is compared with its downsampled version $D_s x_t$. Here, D_s performs spatial downsampling on the grid of its input, and the same factor is applied to memory.

Both paths use the shared prediction map $\Phi^0(x; h)$. It advances the field using the Neural Operator's base increment and the closure conditioned on memory. The superscript 0 indicates that the current commutator evidence is set to zero for this preliminary prediction. We denote the online discrepancy by $\tilde{c}_t^{(s)}$ to distinguish it from the high-resolution quantity in the Introduction:

$$\tilde{c}_t^{(s)} = \underbrace{D_s \Phi^0(x_t; h_t)}_{\text{Path A: evolve, then downsample}} - \underbrace{\Phi^0(D_s x_t; D_s h_t)}_{\text{Path B: downsample, then evolve}} . \tag{3}$$

Path A lets structures represented on the observation grid affect the prediction before downsampling. Path B removes these structures first, then predicts from the coarser field and memory. The two paths share parameters, advance over the same time interval, and produce states on the same grid. Subtracting their outputs measures how jointly coarsening the field and memory changes the forecast.

The spatial pattern locates the affected regions. For each state component, a positive discrepancy means that Path A predicts a larger value than Path B at that location. Its magnitude measures the size of this difference. Repeating the comparison across scales tests how the forecast changes as different amounts of spatial information are removed. CoMNO learns through rollout training how these observable differences inform corrections on the observation grid.

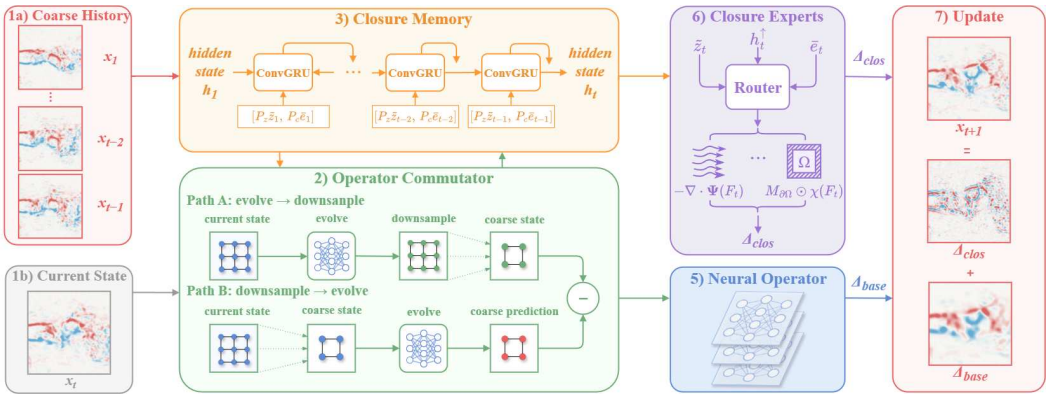

Figure 2: Overview of CoMNO: Operator Commutator evidence and Closure Memory jointly guide Closure Experts to produce spatially adaptive corrections.

The same comparison is applied to the provisional memory outputs to measure how coarsening changes the information carried forward. A scale difference can affect both the immediate prediction and the context used for later predictions. Combining state and memory discrepancies captures both effects in the evidence.

State and memory discrepancies are normalized by the magnitudes of their respective Path A outputs. This expresses each mismatch relative to its output scale while preserving its sign and spatial variation. The discrepancies are then lifted to the observation grid, concatenated along channels, and averaged across scales to form $\bar{e}_t$. Lifting aligns the evidence with the locations to be corrected. Concatenation keeps state and memory differences in separate channels. The final prediction and memory update use this evidence from the same starting pair (x_t, h_t).

3.3 Closure Memory

Closure Memory retains commutator evidence so that differences observed at earlier steps can guide later corrections. A single commutator records the effect of coarsening at the current step. Its history shows how this effect develops as the field evolves, providing context for interpreting each new discrepancy.

The Coarse History is processed in temporal order to initialize this context. Local features describe the current field and its spatial derivatives, capturing local variation and curvature. Conditioning them on h_t and $\bar{e}_t$ produces the features $\tilde{z}_t$ shown in Figure 2.

The ConvGRU updates memory as

$$h_{t+1} = \mathrm{ConvGRU}\left([P_z\tilde{z}_t, P_c\bar{e}_t], h_t\right), \tag{4}$$

where P_z and P_c project features and evidence onto the memory grid, and brackets denote channel concatenation. The projected features describe the local structure, the projected evidence records its current scale discrepancy, and h_t supplies the preceding context. Concatenation presents both current inputs to the gates, which determine how to combine them with the existing memory. Convolutional recurrence preserves the spatial layout while updating this temporal information.

Retaining field features together with the evidence allows memory to associate past discrepancies with the structures that produced them. Similar current fields can therefore receive different corrections when their preceding trajectories differ. The current correction uses h_t and $\bar{e}_t$; h_{t+1} carries the updated context to the next step.

3.4 Neural Operator and Closure Experts

The Neural Operator predicts the base increment $\Delta_{\mathrm{base}}(x_t)$ from the current field. It implements $\Delta_{\mathrm{res}}(x_t)$ in the Introduction. Closure Experts provide the additional correction needed to approximate the history-dependent term $\mathcal{R}(x_t, \mathcal{H}_t)$. The Router and Closure Experts receive

$$F_t = [\tilde{z}_t, h_t^{\uparrow}, \bar{e}_t], \tag{5}$$

where $h_t^{\uparrow}$ denotes memory lifted to the observation grid. Concatenation aligns local field features, past context, and current evidence at each spatial location. The field features describe the structure being corrected, memory describes its recent evolution, and the evidence describes its sensitivity to coarsening. The Router uses these inputs to select expert contributions. The experts use them to determine the direction and magnitude of their corrections.

The Router selects how strongly each expert contributes. It applies a softmax across experts at each location and combines their outputs as

$$\Delta_{\mathrm{clos}}(\mathbf{r}) = \sum_k \alpha_{t,k}(\mathbf{r}) E_k(F_t)(\mathbf{r}), \tag{6}$$

where E_k is expert k, $\alpha_{t,k}$ is its routing weight, and $\mathbf{r}$ is a spatial location. The weights are nonnegative and sum to one across experts. Each expert supplies a candidate correction, and its weight controls its contribution to the final sum. Spatially varying weights let different regions use different expert combinations within the same prediction step.

Figure 2 shows a flux contribution $-\nabla \cdot \Psi(F_t)$ and a boundary contribution $M_{\partial\Omega} \odot \chi(F_t)$. Here, Ω is the spatial domain and Ψ predicts a flux. Its divergence measures net outward transport, so the negative divergence gives the associated local state correction. The mask $M_{\partial\Omega}$ restricts the correction predicted by χ to the boundary region; $\odot$ denotes elementwise multiplication.

3.5 Update

The Update advances the current field by combining the base increment with the closure correction:

$$\hat{x}_{t+1} = x_t + \Delta_{\mathrm{base}}(x_t) + \Delta_{\mathrm{clos}}. \tag{7}$$

This additive form follows the closure formulation in the Introduction. The base increment describes evolution inferred from the current field, and the closure correction incorporates commutator evidence and trajectory history. The prediction becomes the Current State for the next rollout step, together with the updated memory h_{t+1}. Each correction therefore also changes the field from which later commutator evidence is computed.

3.6 Training Objective

The training objective links each correction to both current prediction accuracy and its effect on later steps. All modules are optimized jointly through autoregressive rollouts. The objective combines prediction, commutator regularization, and auxiliary regularization:

$$\mathcal{L} = \mathcal{L}_{\mathrm{pred}} + \mathcal{L}_{\mathrm{comm}} + \mathcal{L}_{\mathrm{reg}}. \tag{8}$$

Each term includes its corresponding weights. $\mathcal{L}_{\mathrm{pred}}$ measures rollout prediction error in physical space and the mismatch between predicted and reference Fourier amplitudes. The physical-space term trains field accuracy, while the Fourier term penalizes differences in the strength of spatial frequencies. $\mathcal{L}_{\mathrm{comm}}$ penalizes the squared, unnormalized state and memory discrepancies of the provisional paths across scales. This soft penalty balances cross-scale agreement against rollout accuracy, while the normalized discrepancies serve as evidence for the final update. $\mathcal{L}_{\mathrm{reg}}$ regularizes memory magnitude and temporal variation, closure-increment magnitude, expert diversity, and routing entropy. Appendix A provides theoretical analysis of CoMNO. Appendix B reports training-loss sensitivity. Appendix D details implementation and computational cost.

4 Experiments

4.1 Experimental Setup

Data and protocols. We evaluate CoMNO on five benchmarks from APEBench Koehler et al. (2024) and RealPDEBench Hu et al. (2026), covering diverse nonlinear dynamics, flow geometries, and fluid–structure coupling. APEBench includes 2D Burgers, Kuramoto–Sivashinsky (KS), and Kolmogorov flow, using native pseudo-spectral trajectories generated at the evaluated resolution. RealPDEBench includes Cylinder Flow and Fluid–Structure Interaction (FSI), with numerical and

real measurement data. Across both suites, models receive four input frames and are trained to predict 20 future steps. APEBench evaluation uses autoregressive rollouts of up to 197 steps, with the 100-step figures displaying the corresponding trajectory prefixes. RealPDEBench evaluation extends predictions to 100 steps by recursively chaining five 20-step forecast blocks. Normalization statistics are estimated separately for each modality using only its training split.

We compare CoMNO with six representative neural surrogates spanning complementary modeling approaches: spectral F-FNO Tran et al. (2023), factorized-attention FactFormer Li et al. (2023), memory-augmented MemNO Buitrago Ruiz et al. (2025), diffusion-based PDE-Refiner Lippe et al. (2023), and pretrained DPOT Hao et al. (2024) and Poseidon Herde et al. (2024). All baselines use official implementations and share the same data splits, channel conventions, normalization procedures, and training and evaluation protocols.

Metrics. Let $\hat{Y}, Y \in \mathbb{R}^{T\times C\times H\times W}$ denote predicted and reference trajectories on the evaluation grid, where C counts the evaluated channels. All metrics use inverse-normalized fields and are averaged over test trajectories. Official RealPDEBench evaluation excludes trailing zero-filled channels for unmeasured variables.

Rel. ℓ_2 measures full-trajectory error as $\|\hat{Y} - Y\|_2/\|Y\|_2$, with norms over time, channels, and space. For stepwise evaluation, we define

$$e_t = \frac{\|\hat{Y}_t - Y_t\|_2}{\|Y_t\|_2 + \epsilon}, \tag{9}$$

where $\epsilon = 10^{-6}$ and norms jointly cover channels and spatial locations. Trajectory nRMSE averages e_t over the forecast window, while Final Step nRMSE reports e_T. Mean Corr. averages the stepwise Pearson correlations $\mathrm{cov}(\hat{Y}_t, Y_t)/(\sigma_{\hat{Y}_t}\sigma_{Y_t} + 10^{-8})$, with covariance and population standard deviations computed over the same flattened channel and spatial entries.

HF Error measures relative complex Fourier error within the radial frequency band Ω_{high} spanning one-third to two-thirds of the Nyquist frequency:

$$\text{HF Error} = \frac{1}{T}\sum_{t=1}^{T} \frac{\left(\sum_{c=1}^{C}\sum_{k\in\Omega_{\mathrm{high}}} w_k \left|\mathcal{F}(\hat{Y}_{t,c})_k - \mathcal{F}(Y_{t,c})_k\right|^2\right)^{1/2}}{\left(\sum_{c=1}^{C}\sum_{k\in\Omega_{\mathrm{high}}} w_k \left|\mathcal{F}(Y_{t,c})_k\right|^2\right)^{1/2} + \epsilon}. \tag{10}$$

Here, $\mathcal{F}$ is the orthonormal spatial 2D real FFT; the Nyquist frequency is $1/2$ cycles per grid point. The Parseval weight w_k equals one on the zero-frequency and, when present, Nyquist columns of the reduced FFT axis, and two elsewhere.

We further report four physical-consistency diagnostics on the same evaluation grid. PDE Step Residual (StepRes) for Burgers and Vorticity Relative ℓ_2 Error (VortRel) for Cylinder are defined as

$$\mathrm{StepRes} = \frac{1}{T-1}\sum_{t=1}^{T-1} \frac{\|\hat{Y}_{t+1} - \Phi_{\mathrm{ref}}(\hat{Y}_t)\|_{\mathrm{rms}}}{\max\{\|\Phi_{\mathrm{ref}}(\hat{Y}_t)\|_{\mathrm{rms}}, \epsilon_{\mathrm{phys}}\}}, \qquad \mathrm{VortRel} = \frac{1}{T}\sum_{t=1}^{T} \frac{\|\hat{\omega}_t - \omega_t\|_2}{\|\omega_t\|_2 + \epsilon}. \tag{11}$$

Here, Φ_{ref} is the APEBench reference time-stepper; $\|\cdot\|_{\mathrm{rms}}$ denotes RMS over components and space, with ϵ_{phys} set to 10^{-12}. Cylinder vorticity ω_t is the spatial curl of the velocity channels in Y_t, with $\hat{\omega}_t$ computed analogously from $\hat{Y}_t$. Norms cover all relevant components and spatial locations. StepRes measures consistency with reference numerical evolution, whereas VortRel measures error in the spatial vorticity field.

Energy Curve Relative Error for Burgers divides the temporal RMS of the prediction-reference energy difference by the reference energy curve's RMS, floored at ϵ_{phys}. Energy averages squared field values over components and space. Enstrophy Relative Error for Cylinder averages stepwise absolute differences in spatial mean squared vorticity, divided by corresponding reference values plus ϵ. These errors measure fidelity to reference energy and enstrophy evolution, without assuming either quantity is conserved.

Table 1: Main comparison on long-horizon PDE rollouts for Burgers from APEBench Koehler et al. (2024) and Cylinder Flow from RealPDEBench Hu et al. (2026).

Model	Rel. $\ell_2 \downarrow$	Trajectory nRMSE $\downarrow$	Final Step nRMSE $\downarrow$	HF Error $\downarrow$	Mean Corr. $\uparrow$
Burgers Equation from APEBench Koehler et al. (2024)					
DPOT Hao et al. (2024)	0.3959 ± 0.0108	0.3959 ± 0.0115	0.6571 ± 0.0186	25.2052 ± 0.6931	0.9051 ± 0.0247
FactFormer Li et al. (2023)	0.4442 ± 0.0027	0.4442 ± 0.0024	0.6559 ± 0.0079	4.1411 ± 0.0093	0.8814 ± 0.0009
F-FNO Tran et al. (2023)	0.0969 ± 0.0028	0.0969 ± 0.0027	0.1555 ± 0.0043	0.4637 ± 0.0118	0.9946 ± 0.0193
MemNO Buitrago Ruiz et al. (2025)	0.2139 ± 0.0061	0.1548 ± 0.0045	0.4655 ± 0.0021	8.1969 ± 0.0795	0.9728 ± 0.0004
PDE-Refiner Lippe et al. (2023)	0.4152 ± 0.0019	0.0698 ± 0.0020	0.8458 ± 0.0043	13.8931 ± 0.0012	0.8947 ± 0.0007
Poseidon Herde et al. (2024)	0.8926 ± 0.0051	0.6306 ± 0.0081	0.4520 ± 0.0029	1.4260 ± 0.0037	0.8935 ± 0.0040
Ours	**0.0568 ± 0.0016**	**0.0568 ± 0.0017**	**0.1412 ± 0.0040**	**0.4288 ± 0.0012**	**0.9978 ± 0.0002**
Cylinder Flow from RealPDEBench Hu et al. (2026)					
DPOT Hao et al. (2024)	0.1389 ± 0.0040	0.1861 ± 0.0053	0.2066 ± 0.0058	0.9526 ± 0.0071	0.9795 ± 0.0006
FactFormer Li et al. (2023)	0.1242 ± 0.0035	0.2046 ± 0.0058	0.1730 ± 0.0049	0.8092 ± 0.0031	0.9564 ± 0.0004
F-FNO Tran et al. (2023)	0.1302 ± 0.0037	0.1302 ± 0.0036	0.1815 ± 0.0052	0.8115 ± 0.0034	0.9845 ± 0.0004
MemNO Buitrago Ruiz et al. (2025)	0.1282 ± 0.0036	0.1282 ± 0.0037	0.1807 ± 0.0051	0.8546 ± 0.0045	0.9854 ± 0.0002
PDE-Refiner Lippe et al. (2023)	0.1248 ± 0.0036	0.1248 ± 0.0035	0.1829 ± 0.0053	0.8150 ± 0.0032	0.9855 ± 0.0009
Poseidon Herde et al. (2024)	0.1370 ± 0.0039	0.1370 ± 0.0038	0.2020 ± 0.0057	0.9568 ± 0.0076	0.9802 ± 0.0007
Ours	**0.1151 ± 0.0034**	**0.1161 ± 0.0033**	**0.1618 ± 0.0048**	**0.8006 ± 0.0028**	**0.9874 ± 0.0004**

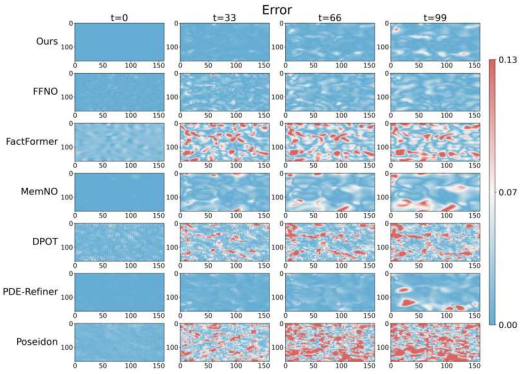

Figure 3: Error evolution over long-horizon rollouts on APEBench Burgers Koehler et al. (2024).

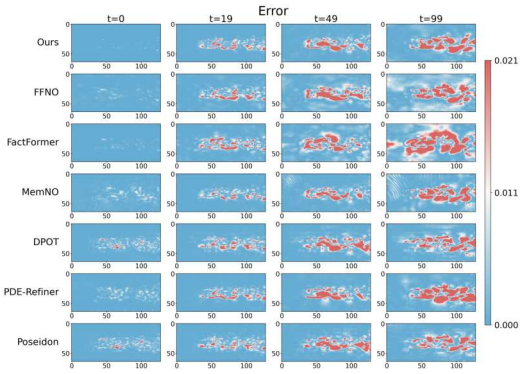

Figure 4: Error evolution over long-horizon rollouts on RealPDEBench Cylinder Flow Hu et al. (2026).

4.2 Main Comparison

Table 1 compares CoMNO with representative neural operators on APEBench Burgers Koehler et al. (2024) and RealPDEBench Cylinder Flow Hu et al. (2026). CoMNO achieves the best reported mean on every metric across both benchmarks. Its advantages span trajectory-level and terminal accuracy, accompanied by improvements in high-frequency fidelity and spatial correlation. These complementary results demonstrate that CoMNO improves long-horizon prediction accuracy while more faithfully capturing the spectral content and spatial structure of the predicted fields.

Figures 3 and 4 further illustrate the evolution of spatial errors during autoregressive rollouts. In particular, CoMNO exhibits predominantly low error at the displayed Burgers prediction times, whereas several baselines develop pronounced deviations in later snapshots, consistent with the quantitative comparison. Additional quantitative results and visualizations are provided in Appendix C.1.

4.3 Functional Role of the Commutator Signal

We intervene on frozen checkpoints via $\bar{e}_t \mapsto \lambda \bar{e}_t$, comparing signal removal ($\lambda = 0$) with the unmodified predictor ($\lambda = 1$). All learned parameters remain fixed, and the signal is recomputed along each resulting trajectory. Let $\varepsilon_{i,m}^{(\lambda)}$ denote the relative ℓ_2 error of $\hat{x}_t^{(\lambda)}$ against x_t under metric m. We report the mean paired gain

$$\overline{G}_m = \frac{1}{N} \sum_{i=1}^{N} \left(\varepsilon_{i,m}^{(0)} - \varepsilon_{i,m}^{(1)} \right), \tag{12}$$

Table 2: Mean paired error reductions from restoring the commutator signal; positive values indicate improvement.

Dataset	One-step gain ↑	Full-rollout gain ↑	Late-horizon gain ↑	HF gain ↑
Burgers	0.0001	0.0045	0.0098	0.0105
KS	0.0001	0.0101	0.0118	0.0233
Kolmogorov	0.0002	0.0178	0.0291	0.0345
Cylinder	0.0003	0.0487	0.0679	0.0503
FSI	0.0001	0.0084	0.0137	0.0195

Table 3: Relative error change (%) against the best baseline per setting and metric; negative values indicate improvement.

Dataset	Factor	Mean nRMSE	Final nRMSE	Spectral Error
Burgers	$4\times$	-92.1	$+2.2$	-39.7
Burgers	$8\times$	-96.8	$+1.6$	-7.5
KS	$4\times$	-23.5	-0.6	-56.2
KS	$8\times$	$+0.4$	-1.8	-1.2
Kolmogorov	$4\times$	$+2.8$	-10.8	$+3.5$
Kolmogorov	$8\times$	-4.2	-8.3	$+1.8$

where i indexes independent trajectories or physical trajectory groups. The metrics evaluate the first prediction, average error over the full rollout and its final quarter, and full-rollout error restricted to Fourier modes above the median radial wavenumber (HF).

Table 2 reveals a consistent pattern across systems: restoring the signal yields small one-step gains, larger full-rollout gains, and late-horizon gains that consistently exceed the full-rollout average. The signal's predictive contribution is therefore more fully expressed during closed-loop evolution, where its influence propagates through autoregressive feedback. Positive HF gains across all systems further show that the benefit extends to fine-scale spatial structure. These interventions at fixed parameters establish the commutator's functional utility within the trained predictor and support its role as an informative input for online closure correction. Appendix C.2 shows that the commutator improves terminal prediction, evolves with the trajectory, and benefits from calibrated signal strength.

4.4 Performance under Spatial Coarsening

We evaluate CoMNO under $4\times$ and $8\times$ spatial coarsening to test closure correction with reduced spatial information. Table 3 shows improvements over the best baseline in most comparisons, spanning trajectory accuracy, terminal accuracy, and spectral fidelity, while the remaining error increases stay within 3.5%. These benefits persist at both coarsening levels, although their magnitude and distribution vary across systems and metrics. Together with the commutator interventions, the results support CoMNO's central design: extracting closure evidence from observable cross-scale discrepancies and retaining its temporal context to guide predictions when fine-scale structures are unresolved. Appendix C.3 reports the full coarsening results under $4\times$ and $8\times$ spatial coarsening.

4.5 Rollout Stability

Figure 5 evaluates autoregressive stability on Burgers and Cylinder flow using relative ℓ_2 error across rollout steps and final nRMSE at the last evaluated step. CoMNO achieves the lowest relative ℓ_2 error at every displayed horizon on both systems, maintaining its advantage as predictions are recursively reused as inputs. On Burgers, it remains clearly separated from most baselines at later steps. On Cylinder flow, all methods exhibit increasing error, with CoMNO consistently tracing the lower envelope. The insets corroborate this advantage through the lowest final nRMSE at step 197 for Burgers and step 100 for Cylinder. Together, these results demonstrate sustained predictive accuracy under extended autoregressive rollout. This persistence is consistent with CoMNO's use of commutator evidence and Closure Memory to inform successive closure corrections as the trajectory evolves. Full statistics and extended-horizon results are reported in Appendix C.4.

4.6 Physical Consistency

Figure 7 complements the rollout evaluation with four physical diagnostics. CoMNO achieves the lowest reported values on both axes in each system. On Burgers, it combines the lowest StepRes with the most accurate energy evolution. On Cylinder flow, it jointly reduces vorticity and enstrophy errors, improving agreement with the reference in both the spatial flow structure and its aggregate rotational intensity. The strongest competing baseline differs between the two systems, yet CoMNO retains its advantage on both diagnostic pairs. These gains show that CoMNO's sustained rollout

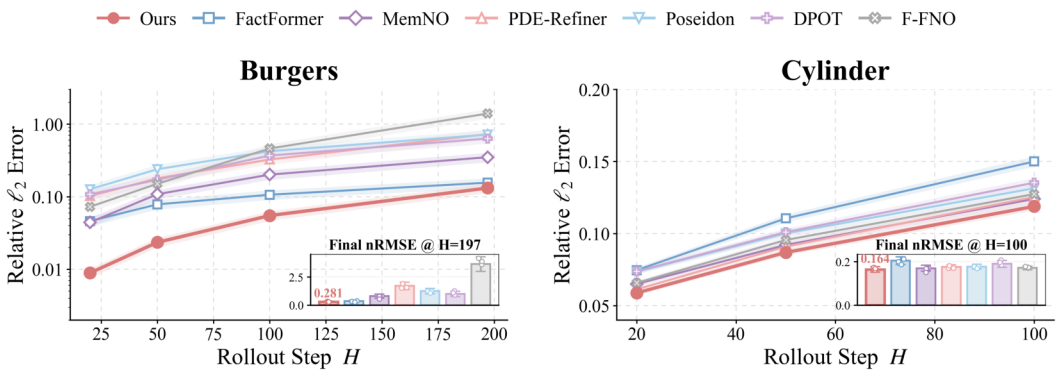

Figure 5: Rollout error across steps on Burgers and Cylinder.

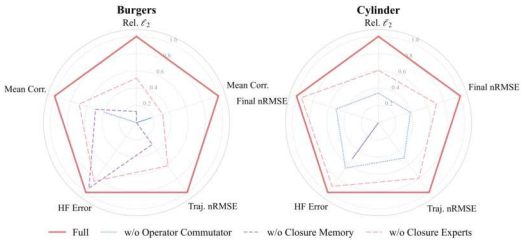

Figure 6: Ablation on Burgers and Cylinder.

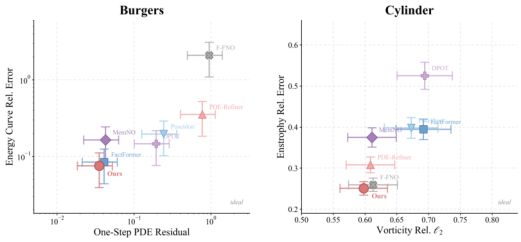

Figure 7: Physical consistency on Burgers and Cylinder.

accuracy is accompanied by more faithful energy and vorticity evolution, with improvements extending from spatial fields to global physical observables. Together, the results support improved physical fidelity alongside lower prediction error. Full statistics and system-specific analyses are provided in Appendix C.5.

4.7 Ablation Study

Figure 6 evaluates CoMNO by individually removing Operator Commutator, Closure Memory, and Closure Experts. The full model performs best across all six metrics on both Burgers and Cylinder flow. Each removal degrades performance across multiple aspects of prediction, affecting state accuracy, rollout error, spectral fidelity, and spatial correlation. The losses in both trajectory and final-step accuracy show that the contribution of each module extends through the rollout to its endpoint. Although the effects vary across metrics and systems, retaining all three modules consistently yields the strongest overall performance. These results support the combined use of commutator evidence, temporal memory, and adaptive closure correction: each module contributes to the full model's performance, with benefits spanning temporal prediction and spatial fidelity. Complete quantitative results and further analyses are provided in Appendix C.6.

5 Conclusion

We presented CoMNO, which learns history-dependent closure by combining cross-scale evidence with trajectory memory. The Operator Commutator measures discrepancies between evolution and downsampling orders; Closure Memory retains their history; and Closure Experts generate spatially adaptive corrections. Across five PDE systems, CoMNO improves predictive accuracy, rollout stability, spectral fidelity, and physical consistency. Gains extend beyond the training horizon and persist in most comparisons under spatial coarsening. Frozen-checkpoint interventions reveal stronger commutator benefits at later horizons, while ablations support the core components. Together, these results demonstrate the value of integrating current cross-scale discrepancies with their temporal history for accurate long-horizon PDE prediction.

AI USE STATEMENT

AI tools assisted with research ideation and experimental design, drafting portions of the manuscript, and refining the writing. The authors reviewed and verified all AI-assisted content, including technical claims and experimental procedures, and take full responsibility for the final manuscript and reported results.

REFERENCES

Syver Døving Agdestein and Benjamin Sanderse. Discretize first, filter next: Learning divergence-consistent closure models for large-eddy simulation. *Journal of Computational Physics*, 522: 113577, 2025. doi: 10.1016/j.jcp.2024.113577.

Andrea Beck, David Flad, and Claus-Dieter Munz. Deep neural networks for data-driven LES closure models. *Journal of Computational Physics*, 398:108910, 2019. doi: 10.1016/j.jcp.2019.108910.

Johannes Brandstetter, Daniel E. Worrall, and Max Welling. Message passing neural PDE solvers. In *International Conference on Learning Representations*, 2022.

Ricardo Buitrago Ruiz, Tanya Marwah, Albert Gu, and Andrej Risteski. On the benefits of memory for modeling time-dependent PDEs. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/89379d5fc6eb34ff98488202fb52b9d0-Abstract-Conference.html.

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

Swinda K. J. Falkena, Courtney Quinn, Jan Sieber, Jason Frank, and Henk A. Dijkstra. Derivation of delay equation climate models using the Mori–Zwanzig formalism. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 475(2227):20190075, 2019. doi: 10.1098/rspa.2019.0075.

Faheem Gilani, Dimitrios Giannakis, and John Harlim. Kernel-based prediction of non-markovian time series. *Physica D: Nonlinear Phenomena*, 418:132829, 2021. doi: 10.1016/j.physd.2020.132829.

Abhinav Gupta and Pierre F. J. Lermusiaux. Neural closure models for dynamical systems. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 477(2252): 20201004, 2021. doi: 10.1098/rspa.2020.1004.

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. URL https://proceedings.mlr.press/v202/hao23c.html.

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, 2024. URL https://arxiv.org/abs/2403.03542.

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

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

Perry L. Johnson. A physics-inspired alternative to spatial filtering for large-eddy simulations of turbulent flows. *Journal of Fluid Mechanics*, 934:A30, 2022. doi: 10.1017/jfm.2021.1150.

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

Felix Koehler, Simon Niedermayr, Rüdiger Westermann, and Nils Thuerey. APEBench: A benchmark for autoregressive neural emulators of PDEs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 120252–120310, 2024. doi: 10.52202/079017-3822. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d9875ebcf74bccdc5076acab0dbee62c-Abstract-Datasets_and_Benchmarks_Track.html.

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

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to PDEs. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

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

Zijie Li, Dule Shu, and Amir Barati Farimani. Scalable transformer for PDE surrogate modeling. In *Advances in Neural Information Processing Systems*, volume 36, pp. 28010–28039, 2023. doi: 10.52202/075280-1216.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=c8P9NQVtmnO.

Zongyi Li, Hongkai Zheng, Nikola Kovachki, David Jin, Haoxuan Chen, Burigede Liu, Kamyar Azizzadenesheli, and Anima Anandkumar. Physics-informed neural operator for learning partial differential equations. *ACM/IMS Journal of Data Science*, 1(3):1–27, 2024. doi: 10.1145/3648506.

Kevin K. Lin and Fei Lu. Data-driven model reduction, wiener projections, and the Koopman–Mori–Zwanzig formalism. *Journal of Computational Physics*, 424:109864, 2021. doi: 10.1016/j.jcp.2020.109864.

Yen Ting Lin, Yifeng Tian, Daniel Livescu, and Marian Anghel. Data-driven learning for the mori–zwanzig formalism: A generalization of the koopman learning framework. *SIAM Journal on Applied Dynamical Systems*, 20(4):2558–2601, 2021. doi: 10.1137/21M1401759.

Phillip Lippe, Bastiaan S. Veeling, Paris Perdikaris, Richard E. Turner, and Johannes Brandstetter. PDE-Refiner: Achieving accurate long rollouts with neural PDE solvers. In *Advances in Neural Information Processing Systems*, volume 36, pp. 67398–67433, 2023. doi: 10.52202/075280-2946.

Bjoern List, Li-Wei Chen, Kartik Bali, and Nils Thuerey. Differentiability in unrolled training of neural physics simulators on transient dynamics. *Computer Methods in Applied Mechanics and Engineering*, 433:117441, 2025. doi: 10.1016/j.cma.2024.117441.

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*, 3:218–229, 2021. doi: 10.1038/s42256-021-00302-5.

Romit Maulik, Omer San, Adil Rasheed, and Prakash Vedula. Subgrid modelling for two-dimensional turbulence using neural networks. *Journal of Fluid Mechanics*, 858:122–144, 2019. doi: 10.1017/jfm.2018.770.

Eric J. Parish and Karthik Duraisamy. A dynamic subgrid scale model for large eddy simulations based on the mori–zwanzig formalism. *Journal of Computational Physics*, 349:154–175, 2017a. doi: 10.1016/j.jcp.2017.07.053.

Eric J. Parish and Karthik Duraisamy. Non-markovian closure models for large eddy simulations using the mori–zwanzig formalism. *Physical Review Fluids*, 2(1):014604, 2017b. doi: 10.1103/PhysRevFluids.2.014604.

Jaideep Pathak, Brian Hunt, Michelle Girvan, Zhixin Lu, and Edward Ott. Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach. *Physical Review Letters*, 120(2):024102, 2018. doi: 10.1103/PhysRevLett.120.024102.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter W. Battaglia. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations*, 2021.

Md Ashiqur Rahman, Zachary E. Ross, and Kamyar Azizzadenesheli. U-NO: U-shaped neural operators. *Transactions on Machine Learning Research*, 2023.

Bogdan Raonić, Roberto Molinaro, Tim De Ryck, Tobias Rohner, Francesca Bartolucci, Rima Alaifari, Siddhartha Mishra, and Emmanuel de Bézenac. Convolutional neural operators for robust and accurate learning of PDEs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 77187–77200, 2023.

Benjamin Sanderse, Panos Stinis, Romit Maulik, and Shady E. Ahmed. Scientific machine learning for closure models in multiscale problems: A review. *Foundations of Data Science*, 7(1):298–337, 2025. doi: 10.3934/fods.2024043.

Louis Serrano, Lise Le Boudec, Armand Kassaï Koupaï, Thomas X. Wang, Yuan Yin, Jean-Noël Vittaut, and Patrick Gallinari. Operator learning with neural fields: Tackling PDEs on general geometries. In *Advances in Neural Information Processing Systems*, volume 36, pp. 70581–70611, 2023. doi: 10.52202/075280-3093.

Kaushik Srinivasan, Mickaël D. Chekroun, and James C. McWilliams. Turbulence closure with small, local neural networks: Forced two-dimensional and beta-plane flows. *Journal of Advances in Modeling Earth Systems*, 16(4):e2023MS003795, 2024. doi: 10.1029/2023MS003795.

Yifan Sun, Haiying Qi, Guihua Zhang, and Yuxin Wu. A novel dynamic subgrid-scale model based on partial-averaged velocity for large eddy simulation. *Physics of Fluids*, 36(2):025114, 2024. doi: 10.1063/5.0183372.

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

Yifeng Tian, Yen Ting Lin, Marian Anghel, and Daniel Livescu. Data-driven learning of mori–zwanzig operators for isotropic turbulence. *Physics of Fluids*, 33(12):125118, 2021. doi: 10.1063/5.0070548.

Alasdair Tran, Alexander Mathews, Lexing Xie, and Cheng Soon Ong. Factorized fourier neural operators. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=tmIiMPl4IPa`.

Toby van Gastelen, Wouter Edeling, and Benjamin Sanderse. Energy-conserving neural network for turbulence closure modeling. *Journal of Computational Physics*, 508:113003, 2024. doi: 10.1016/j.jcp.2024.113003.

Pantelis R. Vlachas, Wonmin Byeon, Zhong Yi Wan, Themistoklis P. Sapsis, and Petros Koumoutsakos. Data-driven forecasting of high-dimensional chaotic systems with long short-term memory networks. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 474(2213):20170844, 2018. doi: 10.1098/rspa.2017.0844.

Pantelis R. Vlachas, Jaideep Pathak, Brian R. Hunt, Themistoklis P. Sapsis, Michelle Girvan, Edward Ott, and Petros Koumoutsakos. Backpropagation algorithms and reservoir computing in recurrent neural networks for the forecasting of complex spatiotemporal dynamics. *Neural Networks*, 126:191–217, 2020. doi: 10.1016/j.neunet.2020.02.016.

Chuwei Wang, Boris Bonev, Julius Berner, Zongyi Li, Di Zhou, Jiayun Wang, Thorsten Kurth, Jane Bae, and Anima Anandkumar. Coarse graining with neural operators for simulating chaotic systems, 2026a. URL https://arxiv.org/abs/2408.05177.

Qian Wang, Nicolò Ripamonti, and Jan S. Hesthaven. Recurrent neural network closure of parametric POD-galerkin reduced-order models based on the mori–zwanzig formalism. *Journal of Computational Physics*, 410:109402, 2020. doi: 10.1016/j.jcp.2020.109402.

Zhenzhong Wang, Xin Zhang, Jun Liao, and Min Jiang. Cross-field interface-aware neural operators for multiphase flow simulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 26769–26777, 2026b. doi: 10.1609/aaai.v40i31.39887.

Gopal R. Yalla, Todd A. Oliver, Sigfried W. Haering, Björn Engquist, and Robert D. Moser. Effects of resolution inhomogeneity in large-eddy simulation. *Physical Review Fluids*, 6(7):074604, 2021. doi: 10.1103/PhysRevFluids.6.074604.

Wei-wen Zhao, Fu-chang Zhou, Guo-qing Fan, and De-cheng Wan. Assessment of subgrid-scale models in wall-modeled large-eddy simulations of turbulent channel flows. *Journal of Hydrodynamics*, 35(3):407–416, 2023. doi: 10.1007/s42241-023-0039-6.

A Theory of Commutator-Guided Closure

We analyze how the Operator Commutator supplies a residual proxy, when the memory-conditioned correction improves prediction, and how its effect propagates through autoregressive rollout. Model parameters are fixed. We retain the notation of the main text and use $\|\cdot\|_2$ for the Euclidean norm over field components and spatial locations.

A.1 Calibration of the Online Commutator

The two provisional paths differ in the order of prediction and downsampling. Path A predicts on the observation grid before downsampling, whereas Path B first downsamples the state and memory. Both use Φ^0 with zero current evidence. Their discrepancy provides an estimate of the correction needed by Path B.

To make this interpretation precise, let $r_t^{(s)}$ denote the correction that makes the Path B prediction equal the restricted reference state $D_s x_{t+1}$. The commutator and this correction subtract the same Path B prediction. Their difference is therefore exactly the observation-grid provisional prediction error after restriction. For linear D_s, this gives

$$\left\|\tilde{c}_t^{(s)} - r_t^{(s)}\right\|_2 \le \|D_s\|_{2\to 2}\left\|\Phi^0(x_t; h_t) - x_{t+1}\right\|_2, \tag{13}$$

where $\|D_s\|_{2\to 2}$ is the induced operator norm. The bound connects the accuracy of the residual estimate directly to the provisional forecast error. In particular, an exact forecast on the observation grid makes the commutator equal the correction required on the restricted grid.

CoMNO normalizes, lifts, concatenates, and averages the state and memory discrepancies across scales to form $\bar{e}_t$. Starting again from the same pair (x_t, h_t), the final pass uses this processed evidence together with local features and retained history. Closure Experts thereby learn how cross-resolution disagreement informs the correction on the observation grid.

A.2 Memory-Conditioned Closure and the Value of Evidence

All fields in this subsection are square-integrable under the reference-trajectory distribution. The available history $\mathcal{H}_t$ includes x_t, and h_t is obtained causally from this history using the model's

initialization and memory updates. Following the main-text formulation, the target supplied to the learned closure is

$$R_t := x_{t+1} - x_t - \Delta_{\text{base}}(x_t). \tag{14}$$

Under the closure decomposition in the main text, Δ_{base} implements the resolved increment, so R_t is the history-dependent residual that Closure Experts approximate. Its conditional means with and without memory are

$$m_t^h := \mathbb{E}[R_t \mid x_t, h_t], \qquad m_t^x := \mathbb{E}[R_t \mid x_t]. \tag{15}$$

Proposition A.1 (Retained information and approximation error)**.** *For the final CoMNO prediction from the reference pair* (x_t, h_t)*,*

$$\begin{aligned} &\mathbb{E}\|R_t - m_t^x\|_2^2 - \mathbb{E}\|x_{t+1} - \hat{x}_{t+1}\|_2^2 \\ &\quad = \mathbb{E}\|m_t^h - m_t^x\|_2^2 - \mathbb{E}\|m_t^h - \Delta_{\text{clos}}\|_2^2. \end{aligned} \tag{16}$$

Proof. The model update gives $x_{t+1} - \hat{x}_{t+1} = R_t - \Delta_{\text{clos}}$. At fixed parameters, $\bar{e}_t$, F_t, and Δ_{clos} are functions of (x_t, h_t). Hence $R_t - m_t^h$ is orthogonal in mean square to both $m_t^h - m_t^x$ and $m_t^h - \Delta_{\text{clos}}$. Decompose $R_t - m_t^x$ and $R_t - \Delta_{\text{clos}}$ through m_t^h, expand their squared norms, and subtract. The common residual variance cancels. □

The left side compares the optimal state-only prediction risk with CoMNO's risk. The right side separates the predictive variation resolved by memory from the learned closure's approximation error. The quantity $m_t^h - m_t^x$ therefore measures the part of the history-dependent residual that becomes predictable once retained context is available. CoMNO improves on the optimal state-only forecast precisely when its closure approximation error is smaller than this retained-information benefit.

The Router's softmax weights make Δ_{clos} a convex combination of $E_k(F_t)$ at each location. Its squared error relative to m_t^h is therefore at most the weighted average of expert errors. Spatial summation and expectation bound the approximation term in Eq. (16).

Gain from the actual evidence-conditioned correction. At the same reference pair (x_t, h_t), let $\Delta_{\text{clos}}^{(\lambda)}$ be the correction produced with current evidence $\lambda\bar{e}_t$, for $\lambda \in \{0, 1\}$. Features, routing weights, and expert outputs are evaluated with that evidence, while the starting state and memory remain fixed. Write $\delta_{\text{clos}} := \Delta_{\text{clos}}^{(1)} - \Delta_{\text{clos}}^{(0)}$. Both passes approximate the same conditional mean m_t^h, and their mean-square prediction gain is exactly

$$\begin{aligned} &\mathbb{E}\|R_t - \Delta_{\text{clos}}^{(0)}\|_2^2 - \mathbb{E}\|R_t - \Delta_{\text{clos}}^{(1)}\|_2^2 \\ &\quad = 2\mathbb{E}\langle m_t^h - \Delta_{\text{clos}}^{(0)}, \delta_{\text{clos}}\rangle - \mathbb{E}\|\delta_{\text{clos}}\|_2^2. \end{aligned} \tag{17}$$

Expanding the squared norms and conditioning on (x_t, h_t) replaces R_t by m_t^h in the cross term. Thus, evidence improves the actual one-step mean-square prediction precisely when twice the expected alignment with the remaining mean residual exceeds the expected squared correction norm. Closure Memory supplies retained historical information; the current evidence guides the learned approximation of the resulting mean closure.

A.3 Closed-Loop Propagation and Prediction Gains

We now apply the frozen-checkpoint intervention $\bar{e}_t \mapsto \lambda\bar{e}_t$ throughout a rollout, with a common initial field and memory. As in the main text, each rollout updates both state and memory and recomputes evidence from its own inputs according to the fixed evaluation schedule. In this subsection, state fields are expressed after the inverse normalization used for evaluation.

Propagation through state and memory. Let $\hat{x}_T^{[j]}$ be the terminal prediction when evidence is removed for the first j updates and enabled thereafter. All these hybrid rollouts start from the same initialization. Their endpoints are $\hat{x}_T^{[0]} = \hat{x}_T^{(1)}$ and $\hat{x}_T^{[T]} = \hat{x}_T^{(0)}$, so telescoping gives

$$\delta_T := \hat{x}_T^{(1)} - \hat{x}_T^{(0)} = \sum_{j=0}^{T-1} \big(\hat{x}_T^{[j]} - \hat{x}_T^{[j+1]}\big). \tag{18}$$

Adjacent hybrids share the same state–memory input at update j and differ in whether evidence is supplied at that update. Each terminal contrast includes the subsequent nonlinear response of the state, memory, and recomputed evidence. A change to the immediate memory output can therefore affect later states even when the immediate state predictions coincide.

Connection to the reported error reductions. Define the terminal error field $d_T^{(\lambda)} := x_T - \hat{x}_T^{(\lambda)}$. The intervention gives $d_T^{(1)} = d_T^{(0)} - \delta_T$. Let $e_T^{(\lambda)}$ be the scalar stepwise relative error from the main text, with the same $\epsilon = 10^{-6}$. If at least one error field is nonzero, expansion and the difference-of-squares identity give

$$e_T^{(0)} - e_T^{(1)} = \frac{2\langle d_T^{(0)}, \delta_T \rangle - \|\delta_T\|_2^2}{(\|x_T\|_2 + \epsilon)(\|d_T^{(0)}\|_2 + \|d_T^{(1)}\|_2)}. \tag{19}$$

When both error fields vanish, the gain is zero. Otherwise, the prediction improves precisely when twice the alignment of the total response with the original error exceeds the squared response magnitude. Longer-horizon gains depend on how the recurrent updates change both this response and its alignment with the reference error.

Averaging these normalized differences over the prescribed times and trajectory groups gives the corresponding paired gains $\overline{G}_m$. For full-trajectory norms, stack the evaluated fields and use the metric's normalization. For the intervention HF gain, apply the above-median radial-frequency restriction defined in Section 4.3; the main HF Error retains the one-third-to-two-thirds Nyquist band defined in the evaluation metrics. The two spectral quantities are therefore distinct diagnostics. The measured gains include the recurrent effect of evidence carried forward by Closure Memory.

B Sensitivity Analysis

Training-loss multiplier. We vary a common multiplier β on the spectral, commutator, and regularization terms while keeping the physical-space prediction contribution fixed:

$$\mathcal{L}_\beta = \mathcal{L}_{\text{field}} + \beta\big(\lambda_{\text{spec}}\mathcal{L}_{\text{spec}} + \mathcal{L}_{\text{comm}} + \mathcal{L}_{\text{reg}}\big). \tag{20}$$

Here, $\mathcal{L}_{\text{field}}$ and $\lambda_{\text{spec}}\mathcal{L}_{\text{spec}}$ are the physical-space and Fourier-amplitude contributions to $\mathcal{L}_{\text{pred}}$, respectively. All component weights follow Appendix D; $\beta = 1$ recovers the main-text objective, while $\beta = 0$ retains only its physical-space prediction contribution. Each setting is trained separately using the same architecture, data splits, and checkpoint-selection rule. The inference-time signal multiplier λ is examined separately in Appendix C.2.

Results. Table 4 shows that the default setting achieves the lowest reported relative ℓ_2, terminal, and HF errors, together with the highest mean correlation. The setting $\beta = 0.5$ gives the lowest trajectory nRMSE. Setting $\beta = 0$ increases terminal and HF errors relative to the default, while $\beta = 2$ worsens all five metrics. The different optima for trajectory and terminal accuracy highlight the value of evaluating the loss balance across the forecast, its endpoint, and its spatial frequencies. The default leads four of the five metrics, supporting its use for jointly preserving predictive accuracy and spatial fidelity.

Table 4: Training-loss sensitivity on Burgers. β jointly scales the spectral, commutator, and regularization terms; $\beta = 1$ is the default.

β	Rel. $\ell_2 \downarrow$	Traj. nRMSE $\downarrow$	Final nRMSE $\downarrow$	HF Error $\downarrow$	Mean Corr. $\uparrow$
0	0.0649 ± 0.0018	0.0660 ± 0.0019	0.1845 ± 0.0052	0.5072 ± 0.0144	0.9917 ± 0.0029
0.1	0.0641 ± 0.0018	0.0657 ± 0.0019	0.1621 ± 0.0046	0.5307 ± 0.0151	0.9968 ± 0.0028
0.5	0.0573 ± 0.0016	**0.0553 ± 0.0016**	0.1430 ± 0.0041	0.5132 ± 0.0146	0.9977 ± 0.0028
1	**0.0568 ± 0.0016**	0.0568 ± 0.0017	**0.1412 ± 0.0040**	**0.4288 ± 0.0012**	**0.9978 ± 0.0002**
2	0.0758 ± 0.0022	0.0790 ± 0.0023	0.1968 ± 0.0056	0.7910 ± 0.0225	0.9959 ± 0.0029

Table 5: Prediction accuracy on Burgers. Entries report mean and standard deviation across runs.

Model	Rel. $\ell_2 \downarrow$	Traj. nRMSE $\downarrow$	Final nRMSE $\downarrow$	HF Error $\downarrow$	Mean Corr. $\uparrow$
DPOT	0.3959 ± 0.0108	0.3959 ± 0.0115	0.6571 ± 0.0186	25.2052 ± 0.6931	0.9051 ± 0.0247
FactFormer	0.4442 ± 0.0027	0.4442 ± 0.0024	0.6559 ± 0.0079	4.1411 ± 0.0093	0.8814 ± 0.0009
F-FNO	0.0969 ± 0.0028	0.0969 ± 0.0027	0.1555 ± 0.0043	0.4637 ± 0.0118	0.9946 ± 0.0193
MemNO	0.2139 ± 0.0061	0.1548 ± 0.0045	0.4655 ± 0.0021	8.1969 ± 0.0795	0.9728 ± 0.0004
PDE-Refiner	0.4152 ± 0.0019	0.0698 ± 0.0020	0.8458 ± 0.0043	13.8931 ± 0.0012	0.8947 ± 0.0007
Poseidon	0.8926 ± 0.0051	0.6306 ± 0.0081	0.4520 ± 0.0029	1.4260 ± 0.0037	0.8935 ± 0.0040
CoMNO	$\mathbf{0.0568 \pm 0.0016}$	$\mathbf{0.0568 \pm 0.0017}$	$\mathbf{0.1412 \pm 0.0040}$	$\mathbf{0.4288 \pm 0.0012}$	$\mathbf{0.9978 \pm 0.0002}$

Table 6: Prediction accuracy on Cylinder. Entries report mean and standard deviation across runs.

Model	Rel. $\ell_2 \downarrow$	Traj. nRMSE $\downarrow$	Final nRMSE $\downarrow$	HF Error $\downarrow$	Mean Corr. $\uparrow$
DPOT	0.1389 ± 0.0040	0.1861 ± 0.0053	0.2066 ± 0.0058	0.9526 ± 0.0071	0.9795 ± 0.0006
FactFormer	0.1242 ± 0.0035	0.2046 ± 0.0058	0.1730 ± 0.0049	0.8092 ± 0.0031	0.9564 ± 0.0004
F-FNO	0.1302 ± 0.0037	0.1302 ± 0.0036	0.1815 ± 0.0052	0.8115 ± 0.0034	0.9845 ± 0.0004
MemNO	0.1282 ± 0.0036	0.1282 ± 0.0037	0.1807 ± 0.0051	0.8546 ± 0.0045	0.9854 ± 0.0002
PDE-Refiner	0.1248 ± 0.0036	0.1248 ± 0.0035	0.1829 ± 0.0053	0.8150 ± 0.0032	0.9855 ± 0.0009
Poseidon	0.1370 ± 0.0039	0.1370 ± 0.0038	0.2020 ± 0.0057	0.9568 ± 0.0076	0.9802 ± 0.0007
CoMNO	$\mathbf{0.1151 \pm 0.0034}$	$\mathbf{0.1161 \pm 0.0033}$	$\mathbf{0.1618 \pm 0.0048}$	$\mathbf{0.8006 \pm 0.0028}$	$\mathbf{0.9874 \pm 0.0004}$

Table 7: Prediction accuracy on KS. Entries report mean and standard deviation across runs.

Model	Rel. $\ell_2 \downarrow$	Traj. nRMSE $\downarrow$	Final nRMSE $\downarrow$	HF Error $\downarrow$	Mean Corr. $\uparrow$
DPOT	0.8902 ± 0.0263	0.8902 ± 0.0193	1.2258 ± 0.0224	5.2418 ± 0.1303	0.5141 ± 0.0150
FactFormer	0.8176 ± 0.0137	0.8176 ± 0.0165	1.0127 ± 0.0286	1.0503 ± 0.0254	0.5062 ± 0.0080
F-FNO	0.8435 ± 0.0238	0.8435 ± 0.0220	1.2171 ± 0.0237	1.8693 ± 0.0375	0.5537 ± 0.0122
MemNO	0.3749 ± 0.0082	0.1067 ± 0.0022	0.8095 ± 0.0198	1.3892 ± 0.0390	0.9024 ± 0.0144
PDE-Refiner	0.7620 ± 0.0223	0.1140 ± 0.0023	1.2125 ± 0.0197	1.7985 ± 0.0456	0.6464 ± 0.0140
Poseidon	1.5387 ± 0.0306	0.8356 ± 0.0213	1.7151 ± 0.0309	13.5822 ± 0.3339	0.5183 ± 0.0089
CoMNO	$\mathbf{0.1204 \pm 0.0026}$	$\mathbf{0.0903 \pm 0.0016}$	$\mathbf{0.2836 \pm 0.0067}$	$\mathbf{0.2226 \pm 0.0036}$	$\mathbf{0.9925 \pm 0.0273}$

Table 8: Prediction accuracy on Kolmogorov flow. Entries report mean and standard deviation across runs.

Model	Rel. $\ell_2 \downarrow$	Traj. nRMSE $\downarrow$	Final nRMSE $\downarrow$	HF Error $\downarrow$	Mean Corr. $\uparrow$
DPOT	0.7910 ± 0.0143	0.7910 ± 0.0203	0.9136 ± 0.0183	12.0998 ± 0.2756	0.6113 ± 0.0109
FactFormer	0.8143 ± 0.0162	0.8143 ± 0.0130	1.0123 ± 0.0224	1.9624 ± 0.0312	0.5773 ± 0.0094
F-FNO	0.5784 ± 0.0108	0.5784 ± 0.0114	1.0111 ± 0.0226	1.4050 ± 0.0253	0.7794 ± 0.0127
MemNO	0.8083 ± 0.0215	0.5365 ± 0.0155	1.0724 ± 0.0309	5.6845 ± 0.1471	0.5766 ± 0.0157
PDE-Refiner	0.9183 ± 0.0145	0.8337 ± 0.0146	1.2231 ± 0.0236	13.9558 ± 0.3941	0.3836 ± 0.0079
Poseidon	1.2515 ± 0.0268	0.8961 ± 0.0269	1.6569 ± 0.0283	4.2207 ± 0.1245	0.3730 ± 0.0059
CoMNO	$\mathbf{0.4585 \pm 0.0126}$	$\mathbf{0.4585 \pm 0.0136}$	$\mathbf{0.8851 \pm 0.0149}$	$\mathbf{1.2676 \pm 0.0250}$	$\mathbf{0.8518 \pm 0.0219}$

Table 9: Prediction accuracy on FSI. Entries report mean and standard deviation across runs.

Model	Rel. $\ell_2 \downarrow$	Traj. nRMSE $\downarrow$	Final nRMSE $\downarrow$	HF Error $\downarrow$	Mean Corr. $\uparrow$
DPOT	0.1964 ± 0.0044	0.2228 ± 0.0061	0.4167 ± 0.0019	0.9353 ± 0.0027	0.9634 ± 0.0004
FactFormer	0.1547 ± 0.0033	0.1641 ± 0.0027	0.2508 ± 0.0073	0.6827 ± 0.0072	0.9822 ± 0.0009
F-FNO	0.1793 ± 0.0046	0.1793 ± 0.0028	0.2991 ± 0.0049	0.7158 ± 0.0008	0.9757 ± 0.0005
MemNO	0.1522 ± 0.0036	0.1522 ± 0.0040	0.2376 ± 0.0068	0.6791 ± 0.0081	0.9831 ± 0.0005
PDE-Refiner	0.1698 ± 0.0047	0.1698 ± 0.0050	0.2755 ± 0.0075	0.7236 ± 0.0086	0.9790 ± 0.0006
Poseidon	0.2228 ± 0.0053	0.2236 ± 0.0042	0.4359 ± 0.0088	0.8964 ± 0.0080	0.9751 ± 0.0003
CoMNO	$\mathbf{0.1487 \pm 0.0028}$	$\mathbf{0.1487 \pm 0.0040}$	$\mathbf{0.2305 \pm 0.0063}$	$\mathbf{0.6765 \pm 0.0024}$	$\mathbf{0.9836 \pm 0.0005}$

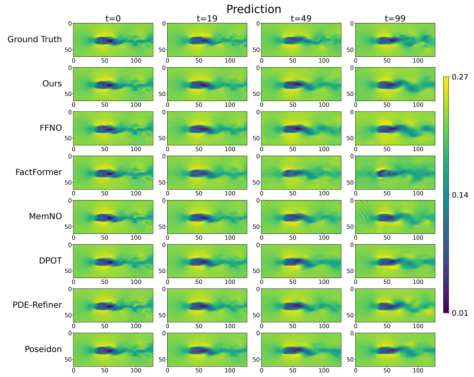

Figure 8: Cylinder u predictions

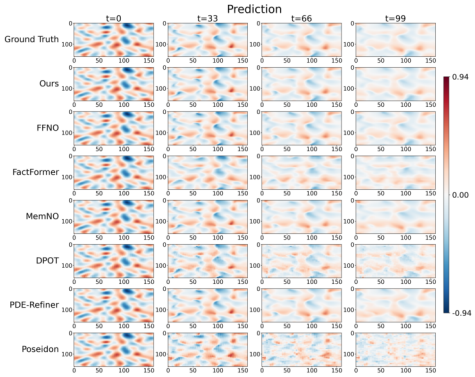

Figure 9: Burgers predictions

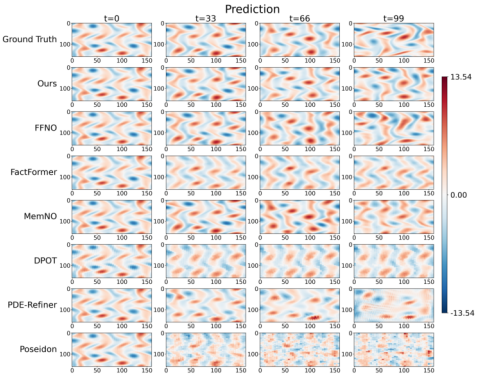

Figure 10: Kolmogorov flow predictions

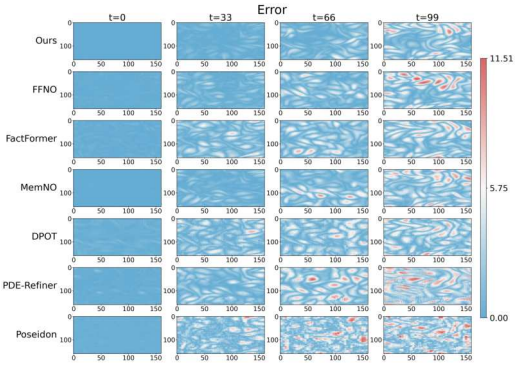

Figure 11: Kolmogorov flow spatial errors

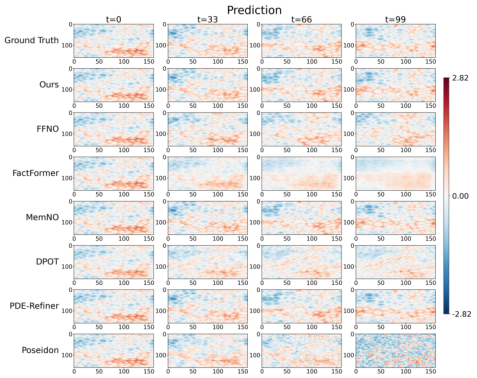

Figure 12: KS predictions

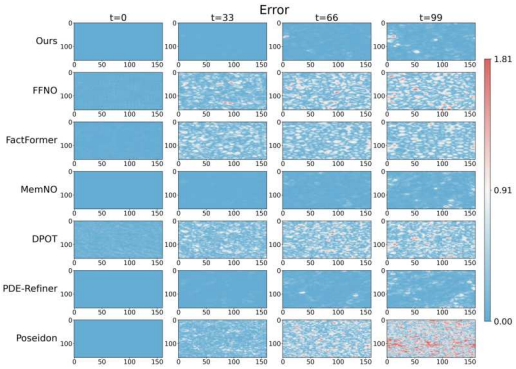

Figure 13: KS spatial errors

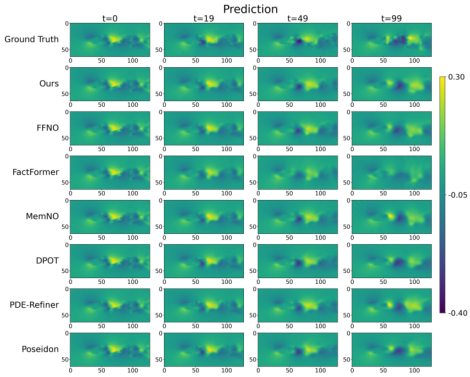

Figure 14: Cylinder v predictions

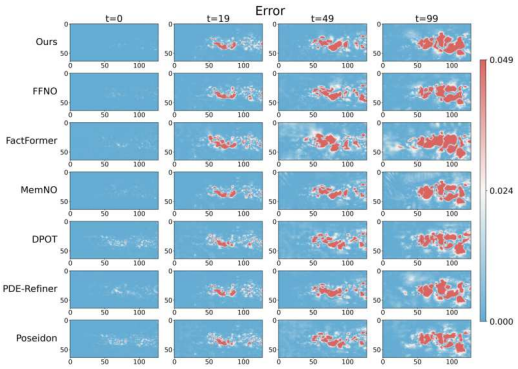

Figure 15: Cylinder v spatial errors

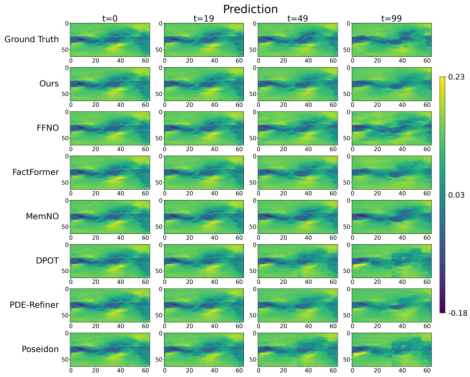

Figure 16: FSI u predictions

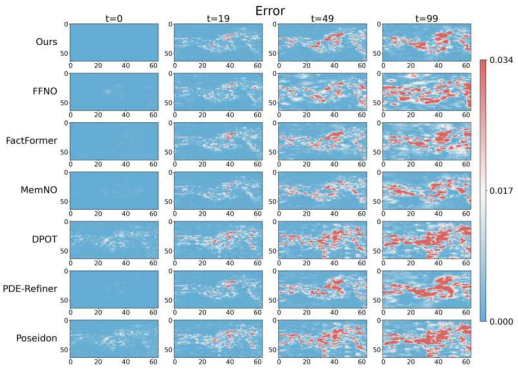

Figure 17: FSI u spatial errors

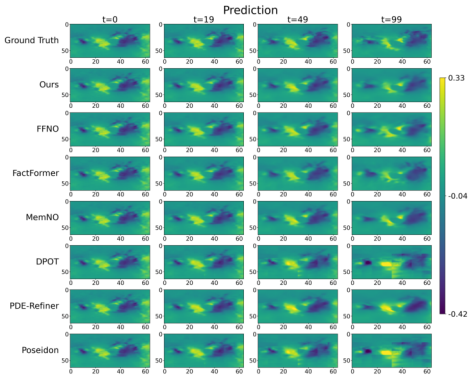

Figure 18: FSI v predictions

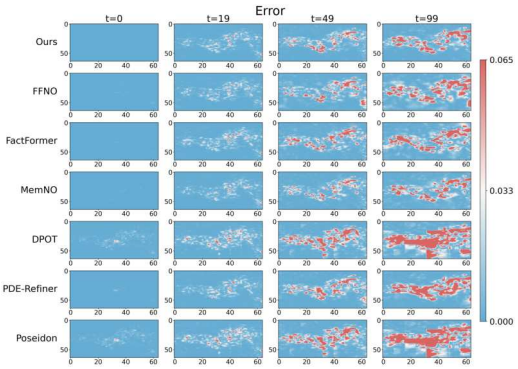

Figure 19: FSI v spatial errors

C Supplementary Results for the Main Experiments

C.1 Main Comparison

Prediction across systems. Tables 5–9 reproduce the five-system main comparison under the evaluation protocol specified in the main text. Horizon-specific recursive extensions are reported separately in Appendix C.4. CoMNO leads the reported means for field, trajectory, terminal, and HF errors and spatial correlation. KS shows a clear joint improvement in rollout accuracy and fine-scale prediction. On Kolmogorov flow, the lower errors accompany better spatial agreement. FSI shows smaller but consistent gains over the strongest baseline. Agreement across full-trajectory and stepwise errors connects the improvement to both the complete forecast and individual predicted states.

Lower HF Error and higher correlation further connect these gains to the spatial structure of the predicted fields.

Table 10: Mean paired reduction in final-step nRMSE from restoring the commutator signal. Positive values indicate improvement.

System	Burgers	KS	Kolmogorov	Cylinder	FSI
Terminal gain ↑	0.0028	0.0098	0.0266	0.0697	0.0053

Table 11: Temporal growth exponents fitted over $k \in \{1, 2, 4, 8, 16\}$. Burgers and KS use step units; the other systems use the listed physical time increments. N_{train} denotes the training-trajectory count.

System	Time increment	N_{train}	State commutator slope	State increment slope	Normalized evidence slope
Burgers	Step	30	0.8927	0.9126	0.8511
KS	Step	30	0.9528	0.9579	0.8247
Kolmogorov	0.1	30	0.7350	0.7847	0.4079
Cylinder	0.005	84	1.0473	0.9667	0.7003
FSI	0.002	42	1.0080	0.9967	0.7937

Table 12: Full-rollout mean stepwise relative ℓ_2 error under inference-time signal scaling. $\lambda = 0$ removes the signal; $\lambda = 1$ gives the unmodified predictor. Boldface marks the lowest displayed error in each row.

System	$\lambda = 0$	$\lambda = 0.25$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 2$	$\lambda = 4$
Burgers	0.0589	0.0542	**0.0540**	0.0544	0.0585	0.0711
KS	0.5692	0.5636	0.5612	**0.5591**	0.5748	0.5920
Kolmogorov	0.5689	0.5559	**0.5512**	**0.5512**	0.5796	0.5920
Cylinder	0.1942	0.1683	0.1534	**0.1455**	0.1504	0.1700
FSI	0.2304	0.2202	**0.2178**	0.2220	0.2499	0.3622

C.2 Functional Role of the Commutator Signal

We extend the frozen-checkpoint intervention in Section 4.3 with terminal prediction, temporal growth, and amplitude sensitivity diagnostics. For the prediction interventions, all learned parameters remain fixed, and $\bar{e}_t \mapsto \lambda \bar{e}_t$ is applied during inference. Evidence is recomputed along each resulting trajectory using the same schedule as the unmodified predictor. The interventions share the main-text test cases and forecast horizons. Errors are first summarized within each trajectory; windows from the same physical recording are averaged within their trajectory group before averaging across independent groups. Paired gains compare the same evaluation units with and without the commutator signal.

Terminal prediction. Table 10 applies the main-text paired gain definition to final-step nRMSE. Each entry is the mean error with the signal removed minus that with the signal restored. All five gains are positive, with the largest absolute reduction on Cylinder. The benefit therefore extends to the final predicted state, complementing the full-rollout and late-horizon results in the main text.

Temporal growth of the evidence. Starting from matched field and memory states, we repeatedly apply the shared provisional transition with zero current evidence on the observation grid and its coarsened versions. At each horizon, we compare evolution followed by downsampling with downsampling followed by evolution. The comparison uses the same normalization, lifting, and evidence construction as Operator Commutator in Section 3.2. For each diagnostic magnitude m, we fit

$$\log m(k\Delta t) = a \log(k\Delta t) + b, \tag{21}$$

by least squares over $k \in \{1, 2, 4, 8, 16\}$. Here, Δt is the time increment per provisional step, a is the fitted growth exponent, and b is the intercept. Table 11 reports a for the raw state commutator, state increment, and normalized evidence. Burgers and KS use step units; the remaining systems use their listed physical time increments. The raw state commutator and state increment have similar fitted exponents. Normalized evidence has a positive exponent smaller than both on every system, indicating slower temporal growth while retaining sensitivity to the evolving trajectory over the tested horizons.

Sensitivity to signal amplitude. Table 12 uses the main-text full-rollout diagnostic: stepwise relative ℓ_2 errors are averaged over the forecast window and then over independent evaluation units. The default amplitude achieves the lowest displayed error on KS and Cylinder and ties with half amplitude on Kolmogorov flow. Half amplitude gives the lowest displayed errors on Burgers and FSI. All three tested amplitudes between 0.25 and 1 improve over signal removal on every system. Increasing the amplitude beyond 1 raises error relative to the default on all five systems, although Cylinder remains better than signal removal at every tested nonzero amplitude. These results support the predictive value of the commutator signal and show that its contribution depends on amplitude, with moderate scaling providing the most consistent gains.

Table 13: Burgers under $4\times$ spatial coarsening. Metrics follow the main-text definitions; boldface marks the best mean in each column.

Model	Rel. $\ell_2 \downarrow$	Traj. nRMSE $\downarrow$	Final nRMSE $\downarrow$	HF Error $\downarrow$	Mean Corr. $\uparrow$
DPOT	1.8691 ± 0.0459	1.6318 ± 0.0411	5.6546 ± 0.0301	4.5972 ± 0.0026	0.6446 ± 0.0061
FactFormer	0.1022 ± 0.0027	0.1108 ± 0.0026	$\mathbf{0.1730 \pm 0.0046}$	0.2621 ± 0.0071	$\mathbf{0.9932 \pm 0.0009}$
F-FNO	0.1598 ± 0.0047	0.1368 ± 0.0033	0.6986 ± 0.0169	0.1501 ± 0.0038	0.9782 ± 0.0017
MemNO	0.0921 ± 0.0026	0.0950 ± 0.0027	0.3645 ± 0.0096	0.1270 ± 0.0033	0.9025 ± 0.0028
PDE-Refiner	0.2297 ± 0.0056	0.1868 ± 0.0056	0.9345 ± 0.0246	0.1825 ± 0.0038	0.9632 ± 0.0047
Poseidon	1.9680 ± 0.0454	1.7388 ± 0.0369	5.3749 ± 0.1513	4.9805 ± 0.1311	0.6223 ± 0.0134
CoMNO	$\mathbf{0.0369 \pm 0.0009}$	$\mathbf{0.0075 \pm 0.0009}$	0.1768 ± 0.0017	$\mathbf{0.0766 \pm 0.0025}$	0.9901 ± 0.0008

Table 14: KS under $4\times$ spatial coarsening. Metrics follow the main-text definitions; boldface marks the best mean in each column.

Model	Rel. $\ell_2 \downarrow$	Traj. nRMSE $\downarrow$	Final nRMSE $\downarrow$	HF Error $\downarrow$	Mean Corr. $\uparrow$
DPOT	1.0642 ± 0.0315	1.0178 ± 0.0266	1.4467 ± 0.0401	1.1944 ± 0.0303	0.4203 ± 0.0104
FactFormer	0.2644 ± 0.0076	0.2072 ± 0.0049	0.5552 ± 0.0166	0.2652 ± 0.0068	0.9643 ± 0.0065
F-FNO	0.4105 ± 0.0098	0.3396 ± 0.0096	0.9282 ± 0.0220	0.4248 ± 0.0109	0.9107 ± 0.0014
MemNO	0.2727 ± 0.0066	0.2146 ± 0.0047	0.5772 ± 0.0172	0.2722 ± 0.0078	0.9620 ± 0.0015
PDE-Refiner	0.3101 ± 0.0084	0.2456 ± 0.0062	0.6352 ± 0.0185	0.3124 ± 0.0080	0.9508 ± 0.0069
Poseidon	1.0137 ± 0.0258	1.1506 ± 0.0337	1.2768 ± 0.0302	1.1360 ± 0.0324	0.2754 ± 0.0017
CoMNO	$\mathbf{0.2539 \pm 0.0058}$	$\mathbf{0.1585 \pm 0.0060}$	$\mathbf{0.5519 \pm 0.0116}$	$\mathbf{0.1162 \pm 0.0061}$	$\mathbf{0.9662 \pm 0.0042}$

Table 15: Kolmogorov flow under $4\times$ spatial coarsening. Metrics follow the main-text definitions; boldface marks the best mean in each column.

Model	Rel. $\ell_2 \downarrow$	Traj. nRMSE $\downarrow$	Final nRMSE $\downarrow$	HF Error $\downarrow$	Mean Corr. $\uparrow$
DPOT	0.8340 ± 0.0083	0.8235 ± 0.0097	1.0202 ± 0.0006	1.1445 ± 0.0024	0.5513 ± 0.0165
FactFormer	0.8107 ± 0.0095	0.7500 ± 0.0004	1.1085 ± 0.0092	1.0817 ± 0.0007	0.6616 ± 0.0077
F-FNO	0.7661 ± 0.0089	0.7030 ± 0.0085	1.1302 ± 0.0012	1.0826 ± 0.0063	0.7068 ± 0.0062
MemNO	0.6828 ± 0.0071	$\mathbf{0.6060 \pm 0.0056}$	1.0607 ± 0.0067	$\mathbf{0.9691 \pm 0.0056}$	0.7068 ± 0.0050
PDE-Refiner	0.8022 ± 0.0018	0.7245 ± 0.0098	1.1724 ± 0.0089	1.0424 ± 0.0005	0.6799 ± 0.0072
Poseidon	0.8874 ± 0.0014	0.8546 ± 0.0044	0.9508 ± 0.0074	1.1794 ± 0.0032	0.5837 ± 0.0035
CoMNO	$\mathbf{0.5523 \pm 0.0041}$	0.6230 ± 0.0025	$\mathbf{0.8481 \pm 0.0062}$	1.0030 ± 0.0023	$\mathbf{0.8340 \pm 0.0037}$

Table 16: Burgers under $8\times$ spatial coarsening. Metrics follow the main-text definitions; boldface marks the best mean in each column.

Model	Rel. $\ell_2 \downarrow$	Traj. nRMSE $\downarrow$	Final nRMSE $\downarrow$	HF Error $\downarrow$	Mean Corr. $\uparrow$
DPOT	0.5241 ± 0.0051	0.5623 ± 0.0029	0.8472 ± 0.0022	0.7066 ± 0.0010	0.7943 ± 0.0015
FactFormer	0.0791 ± 0.0022	0.0819 ± 0.0024	$\mathbf{0.2175 \pm 0.0055}$	0.0814 ± 0.0018	0.9959 ± 0.0034
F-FNO	0.2612 ± 0.0078	0.2406 ± 0.0054	0.8523 ± 0.0051	0.2071 ± 0.0047	0.9495 ± 0.0279
MemNO	0.1186 ± 0.0025	0.1177 ± 0.0028	0.3095 ± 0.0067	0.1198 ± 0.0025	0.9885 ± 0.0216
PDE-Refiner	0.2614 ± 0.0066	0.2247 ± 0.0049	1.0682 ± 0.0238	0.1300 ± 0.0031	0.9555 ± 0.0073
Poseidon	0.5251 ± 0.0117	0.5634 ± 0.0154	0.8497 ± 0.0029	0.7071 ± 0.0003	0.7935 ± 0.0025
CoMNO	$\mathbf{0.0733 \pm 0.0015}$	$\mathbf{0.0026 \pm 0.0023}$	0.2210 ± 0.0032	$\mathbf{0.0753 \pm 0.0004}$	$\mathbf{0.9965 \pm 0.0021}$

Table 17: KS under $8\times$ spatial coarsening. Metrics follow the main-text definitions; boldface marks the best mean in each column.

Model	Rel. $\ell_2 \downarrow$	Traj. nRMSE $\downarrow$	Final nRMSE $\downarrow$	HF Error $\downarrow$	Mean Corr. $\uparrow$
DPOT	1.0016 ± 0.0247	0.9849 ± 0.0081	1.1943 ± 0.0286	1.2075 ± 0.0336	0.4025 ± 0.0096
FactFormer	0.8185 ± 0.0036	$\mathbf{0.7359 \pm 0.0093}$	1.1958 ± 0.0064	0.8782 ± 0.0012	0.6521 ± 0.0080
F-FNO	1.3084 ± 0.0292	1.1810 ± 0.0313	1.9977 ± 0.0494	1.2119 ± 0.0010	0.4421 ± 0.0113
MemNO	1.3343 ± 0.0014	1.1852 ± 0.0336	1.7395 ± 0.0017	1.5925 ± 0.0045	0.3159 ± 0.0092
PDE-Refiner	0.9359 ± 0.0029	0.8854 ± 0.0202	1.1956 ± 0.0014	1.0869 ± 0.0006	0.5288 ± 0.0049
Poseidon	0.9971 ± 0.0288	0.9812 ± 0.0230	1.1016 ± 0.0282	1.1881 ± 0.0344	0.4375 ± 0.0099
CoMNO	$\mathbf{0.7911 \pm 0.0009}$	0.7388 ± 0.0084	$\mathbf{1.0818 \pm 0.0087}$	$\mathbf{0.8677 \pm 0.0024}$	$\mathbf{0.6924 \pm 0.0206}$

Table 18: Kolmogorov flow under $8\times$ spatial coarsening. Metrics follow the main-text definitions; boldface marks the best mean in each column.

Model	Rel. $\ell_2 \downarrow$	Traj. nRMSE $\downarrow$	Final nRMSE $\downarrow$	HF Error $\downarrow$	Mean Corr. $\uparrow$
DPOT	0.9186 ± 0.0201	0.9009 ± 0.0029	1.0320 ± 0.0071	0.8337 ± 0.0046	0.5181 ± 0.0121
FactFormer	0.6587 ± 0.0187	0.5774 ± 0.0033	1.0420 ± 0.0268	0.5419 ± 0.0144	0.7800 ± 0.0072
F-FNO	0.7432 ± 0.0200	0.6496 ± 0.0180	1.1729 ± 0.0021	0.5667 ± 0.0057	0.7510 ± 0.0201
MemNO	0.6075 ± 0.0059	0.5299 ± 0.0155	1.0175 ± 0.0087	0.4965 ± 0.0045	0.8149 ± 0.0037
PDE-Refiner	0.6076 ± 0.0130	0.5251 ± 0.0018	1.0547 ± 0.0223	$\mathbf{0.4007 \pm 0.0018}$	0.8146 ± 0.0200
Poseidon	0.9390 ± 0.0266	0.9298 ± 0.0056	1.0765 ± 0.0054	0.8403 ± 0.0220	0.5031 ± 0.0012
CoMNO	$\mathbf{0.6012 \pm 0.0179}$	$\mathbf{0.5030 \pm 0.0135}$	$\mathbf{0.9330 \pm 0.0087}$	0.4079 ± 0.0039	$\mathbf{0.8199 \pm 0.0012}$

C.3 Performance under Spatial Coarsening

Protocol and metrics. We coarsen the observed histories of Burgers, KS, and Kolmogorov flow by factors of $4\times$ and $8\times$, while retaining future reference fields on the original grid. Each method is trained separately for each factor using the same data splits, temporal windows, normalization procedures, and evaluation protocol. Tables 13 to 18 report five prediction metrics using the main-text definitions. Entries are means and standard deviations across five random seeds. In Table 3, Mean nRMSE and Spectral Error correspond to Traj. nRMSE and HF Error, respectively. For each setting and metric, the relative error change is computed from the reported means as

$$\Delta_m = 100 \left(\frac{\overline{E}_{\mathrm{CoMNO},m}}{\min_{b \in \mathcal{B}} \overline{E}_{b,m}} - 1 \right), \tag{22}$$

where $\mathcal{B}$ contains the six baselines and m indexes one of the three summarized error metrics. The best baseline is selected independently for each setting and metric; negative values indicate improvement.

Results. CoMNO achieves the lowest relative ℓ_2 error in all six settings. Across the three metrics summarized in Table 3, it improves 12 of 18 comparisons, with the remaining error increases bounded by 3.5%. On Burgers, CoMNO improves trajectory and spectral accuracy at both coarsening factors, while FactFormer retains a small advantage in terminal accuracy. On KS, CoMNO leads

all three summarized metrics at $4\times$ and retains terminal and spectral gains at $8\times$, where FactFormer has the lowest trajectory nRMSE. On Kolmogorov flow, CoMNO improves terminal accuracy at both factors and trajectory accuracy at $8\times$; MemNO leads trajectory and spectral accuracy at $4\times$, and PDE-Refiner leads spectral accuracy at $8\times$. These results show that CoMNO's predictive benefits persist with reduced spatial information, with gains distributed across trajectory, terminal, and spectral fidelity.

Table 19: Rollout accuracy on Burgers. Relative ℓ_2 uses the first H predictions; Final Step nRMSE uses step 197.

Model	Rel. $\ell_2 \downarrow$				Final Step nRMSE $\downarrow$
	$H = 20$	$H = 50$	$H = 100$	$H = 197$	$H = 197$
DPOT	0.1072 ± 0.0031	0.1712 ± 0.0049	0.3679 ± 0.0005	0.6317 ± 0.0081	0.9955 ± 0.0084
FactFormer	0.0461 ± 0.0013	0.0783 ± 0.0022	0.1063 ± 0.0030	0.1557 ± 0.0045	0.3433 ± 0.0098
F-FNO	0.0725 ± 0.0021	0.1506 ± 0.0043	0.4601 ± 0.0031	1.3937 ± 0.0099	3.6467 ± 0.0346
MemNO	0.0442 ± 0.0013	0.1082 ± 0.0031	0.2011 ± 0.0058	0.3494 ± 0.0053	0.7902 ± 0.0026
PDE-Refiner	0.1017 ± 0.0029	0.1791 ± 0.0051	0.3249 ± 0.0093	0.7272 ± 0.0008	1.7102 ± 0.0091
Poseidon	0.1264 ± 0.0036	0.2386 ± 0.0068	0.4233 ± 0.0021	0.7154 ± 0.0005	1.2365 ± 0.0054
CoMNO	**0.0089 ± 0.0003**	**0.0235 ± 0.0007**	**0.0549 ± 0.0016**	**0.1315 ± 0.0038**	**0.2812 ± 0.0080**

Table 20: Rollout accuracy on KS. Relative ℓ_2 uses the first H predictions; Final Step nRMSE uses step 197.

Model	Rel. $\ell_2 \downarrow$				Final Step nRMSE $\downarrow$
	$H = 20$	$H = 50$	$H = 100$	$H = 197$	$H = 197$
DPOT	0.1159 ± 0.0032	0.4763 ± 0.0028	0.8799 ± 0.0031	1.1432 ± 0.0006	1.3979 ± 0.0085
FactFormer	0.2476 ± 0.0067	0.5825 ± 0.0059	0.8192 ± 0.0023	0.9395 ± 0.0056	**1.0761 ± 0.0094**
F-FNO	**0.0357 ± 0.0010**	0.1021 ± 0.0028	0.3118 ± 0.0087	0.8598 ± 0.0039	1.6042 ± 0.0036
MemNO	0.0534 ± 0.0015	0.1898 ± 0.0052	0.4489 ± 0.0021	0.8584 ± 0.0034	1.3305 ± 0.0059
PDE-Refiner	0.2638 ± 0.0071	0.4556 ± 0.0022	0.7727 ± 0.0010	1.0484 ± 0.0082	1.3378 ± 0.0064
Poseidon	0.1387 ± 0.0038	0.5443 ± 0.0147	0.7654 ± 0.0206	1.1764 ± 0.0019	1.4233 ± 0.0092
CoMNO	0.0486 ± 0.0013	**0.0938 ± 0.0025**	**0.2687 ± 0.0072**	**0.8506 ± 0.0029**	1.1310 ± 0.0031

Table 21: Rollout accuracy on Kolmogorov flow. Relative ℓ_2 uses the first H predictions; Final Step nRMSE uses step 197.

Model	Rel. $\ell_2 \downarrow$				Final Step nRMSE $\downarrow$
	$H = 20$	$H = 50$	$H = 100$	$H = 197$	$H = 197$
DPOT	0.4546 ± 0.0024	0.7116 ± 0.0092	0.8576 ± 0.0031	0.9995 ± 0.0002	1.1965 ± 0.0025
FactFormer	0.4763 ± 0.0029	0.6968 ± 0.0089	0.8076 ± 0.0018	**0.8955 ± 0.0002**	1.0062 ± 0.0075
F-FNO	0.2020 ± 0.0055	0.4352 ± 0.0017	0.7602 ± 0.0005	1.0692 ± 0.0089	1.4734 ± 0.0095
MemNO	0.3295 ± 0.0089	0.6584 ± 0.0077	0.8547 ± 0.0030	1.0103 ± 0.0073	1.2073 ± 0.0029
PDE-Refiner	0.3691 ± 0.0042	0.6988 ± 0.0088	0.9162 ± 0.0047	1.0740 ± 0.0088	1.2081 ± 0.0026
Poseidon	0.5038 ± 0.0035	0.6749 ± 0.0082	0.8693 ± 0.0035	1.0546 ± 0.0085	1.1934 ± 0.0022
CoMNO	**0.1243 ± 0.0034**	**0.2879 ± 0.0078**	**0.5689 ± 0.0053**	0.9651 ± 0.0061	**0.9992 ± 0.0070**

Table 22: Rollout accuracy on Cylinder. Relative ℓ_2 uses the first H predictions; Final Step nRMSE uses step 100.

Model	Rel. $\ell_2 \downarrow$			Final Step nRMSE $\downarrow$
	$H = 20$	$H = 50$	$H = 100$	$H = 100$
DPOT	0.0738 ± 0.0021	0.1010 ± 0.0029	0.1353 ± 0.0039	0.1896 ± 0.0054
FactFormer	0.0746 ± 0.0021	0.1105 ± 0.0032	0.1500 ± 0.0043	0.2047 ± 0.0059
F-FNO	0.0658 ± 0.0019	0.0955 ± 0.0027	0.1274 ± 0.0036	0.1707 ± 0.0049
MemNO	0.0650 ± 0.0019	0.0921 ± 0.0026	0.1240 ± 0.0035	0.1694 ± 0.0048
PDE-Refiner	0.0613 ± 0.0018	0.0909 ± 0.0026	0.1254 ± 0.0036	0.1752 ± 0.0050
Poseidon	0.0734 ± 0.0021	0.1001 ± 0.0029	0.1314 ± 0.0038	0.1769 ± 0.0051
CoMNO	**0.0588 ± 0.0017**	**0.0869 ± 0.0025**	**0.1188 ± 0.0034**	**0.1640 ± 0.0047**

Table 23: Rollout accuracy on FSI. Relative ℓ_2 uses the first H predictions; Final Step nRMSE uses step 100.

Model	Rel. $\ell_2 \downarrow$			Final Step nRMSE $\downarrow$
	$H = 20$	$H = 50$	$H = 100$	$H = 100$
DPOT	0.1021 ± 0.0028	0.1550 ± 0.0042	0.2425 ± 0.0066	0.4233 ± 0.0116
FactFormer	0.0707 ± 0.0019	0.1260 ± 0.0034	0.1970 ± 0.0053	0.2943 ± 0.0080
F-FNO	0.0690 ± 0.0019	0.1227 ± 0.0033	0.1967 ± 0.0053	0.3021 ± 0.0082
MemNO	0.0631 ± 0.0017	0.1091 ± 0.0029	0.1645 ± 0.0044	0.2397 ± 0.0065
PDE-Refiner	0.0676 ± 0.0018	0.1195 ± 0.0032	0.1852 ± 0.0050	0.2786 ± 0.0075
Poseidon	0.1035 ± 0.0028	0.1593 ± 0.0043	0.2467 ± 0.0067	0.4357 ± 0.0118
CoMNO	**0.0610 ± 0.0016**	**0.1080 ± 0.0029**	**0.1607 ± 0.0043**	**0.2318 ± 0.0063**

Table 24: Full-rollout diagnostics on Burgers (197 steps). Late Step nRMSE uses the final half of the forecast; Energy Curve Error measures relative error in energy evolution.

Model	Trajectory nRMSE $\downarrow$	Final Step nRMSE $\downarrow$	Late Step nRMSE $\downarrow$	Mean Corr. $\uparrow$	Energy Curve Error $\downarrow$
DPOT	0.7323 ± 0.0081	0.9955 ± 0.0084	0.9867 ± 0.0012	0.6244 ± 0.0031	0.6728 ± 0.0072
FactFormer	0.1830 ± 0.0046	0.3433 ± 0.0098	0.2786 ± 0.0055	0.9854 ± 0.0025	0.4130 ± 0.0080
F-FNO	1.5670 ± 0.0029	3.6467 ± 0.0346	3.0536 ± 0.0087	0.5999 ± 0.0025	11.6846 ± 0.0904
MemNO	0.4121 ± 0.0099	0.7902 ± 0.0026	0.6757 ± 0.0064	0.9158 ± 0.0041	0.7061 ± 0.0035
PDE-Refiner	0.8474 ± 0.0022	1.7102 ± 0.0091	1.5007 ± 0.0031	0.6686 ± 0.0036	1.7539 ± 0.0084
Poseidon	0.7684 ± 0.0069	1.2365 ± 0.0054	1.3627 ± 0.0071	0.6837 ± 0.0082	0.6421 ± 0.0057
CoMNO	**0.1805 ± 0.0047**	**0.2812 ± 0.0080**	**0.2657 ± 0.0063**	**0.9868 ± 0.0074**	**0.3924 ± 0.0088**

Table 25: Full-rollout diagnostics on KS (197 steps). Late Step nRMSE uses the final half of the forecast; Energy Curve Error measures relative error in energy evolution.

Model	Trajectory nRMSE $\downarrow$	Final Step nRMSE $\downarrow$	Late Step nRMSE $\downarrow$	Mean Corr. $\uparrow$	Energy Curve Error $\downarrow$
DPOT	1.0595 ± 0.0021	1.3979 ± 0.0085	1.3777 ± 0.0070	0.3390 ± 0.0074	0.1303 ± 0.0032
FactFormer	0.9011 ± 0.0003	**1.0761 ± 0.0094**	**1.0547 ± 0.0044**	0.3625 ± 0.0078	0.7322 ± 0.0055
F-FNO	0.6890 ± 0.0066	1.6042 ± 0.0036	1.3032 ± 0.0016	**0.6872 ± 0.0086**	0.9159 ± 0.0007
MemNO	0.7438 ± 0.0056	1.3305 ± 0.0059	1.2119 ± 0.0086	0.6533 ± 0.0029	0.1908 ± 0.0037
PDE-Refiner	0.9837 ± 0.0097	1.3378 ± 0.0064	1.2963 ± 0.0039	0.4393 ± 0.0004	0.1233 ± 0.0024
Poseidon	1.0235 ± 0.0030	1.4233 ± 0.0092	1.2553 ± 0.0098	0.4297 ± 0.0019	0.1684 ± 0.0045
CoMNO	**0.6630 ± 0.0027**	1.1310 ± 0.0031	1.0957 ± 0.0075	0.6732 ± 0.0060	**0.0941 ± 0.0020**

Table 26: Full-rollout diagnostics on Kolmogorov flow (197 steps). Late Step nRMSE uses the final half of the forecast; Energy Curve Error measures relative error in energy evolution.

Model	Trajectory nRMSE ↓	Final Step nRMSE ↓	Late Step nRMSE ↓	Mean Corr. ↑	Energy Curve Error ↓
DPOT	0.9768 ± 0.0027	1.1965 ± 0.0025	1.1565 ± 0.0047	0.3991 ± 0.0007	0.3246 ± 0.0074
FactFormer	0.9816 ± 0.0005	1.0062 ± 0.0075	**0.9933 ± 0.0054**	0.5200 ± 0.0008	0.3585 ± 0.0078
F-FNO	0.9966 ± 0.0079	1.4734 ± 0.0095	1.3657 ± 0.0013	0.5210 ± 0.0023	0.7592 ± 0.0053
MemNO	0.9773 ± 0.0005	1.2073 ± 0.0029	1.1768 ± 0.0086	0.4079 ± 0.0086	0.3163 ± 0.0066
PDE-Refiner	1.0418 ± 0.0005	1.2081 ± 0.0026	1.2322 ± 0.0060	0.2517 ± 0.0068	0.4942 ± 0.0032
Poseidon	0.9822 ± 0.0093	1.1934 ± 0.0022	1.1652 ± 0.0092	0.4228 ± 0.0088	0.3657 ± 0.0074
CoMNO	**0.7886 ± 0.0016**	**0.9992 ± 0.0070**	1.0294 ± 0.0039	**0.6222 ± 0.0062**	**0.2164 ± 0.0057**

Table 27: Full-rollout diagnostics on Cylinder (100 steps). Late Step nRMSE uses the final half of the forecast; Energy Curve Error measures relative error in energy evolution.

Model	Trajectory nRMSE ↓	Final Step nRMSE ↓	Late Step nRMSE ↓	Mean Corr. ↑	Energy Curve Error ↓
DPOT	0.1292 ± 0.0031	0.1896 ± 0.0054	0.1693 ± 0.0042	0.9840 ± 0.0063	0.2070 ± 0.0050
FactFormer	0.1422 ± 0.0036	0.2047 ± 0.0059	0.1883 ± 0.0040	0.9831 ± 0.0012	0.2099 ± 0.0041
F-FNO	0.1216 ± 0.0026	0.1707 ± 0.0049	0.1587 ± 0.0034	0.9871 ± 0.0044	0.2249 ± 0.0050
MemNO	0.1181 ± 0.0026	0.1694 ± 0.0048	0.1554 ± 0.0033	0.9877 ± 0.0007	0.2261 ± 0.0056
PDE-Refiner	0.1189 ± 0.0029	0.1752 ± 0.0050	0.1590 ± 0.0041	0.9873 ± 0.0002	0.0807 ± 0.0018
Poseidon	0.1267 ± 0.0035	0.1769 ± 0.0051	0.1687 ± 0.0041	0.9826 ± 0.0047	0.2081 ± 0.0055
CoMNO	**0.1127 ± 0.0029**	**0.1640 ± 0.0047**	**0.1503 ± 0.0029**	**0.9888 ± 0.0006**	**0.0497 ± 0.0011**

Table 28: Full-rollout diagnostics on FSI (100 steps). Late Step nRMSE uses the final half of the forecast; Energy Curve Error measures relative error in energy evolution.

Model	Trajectory nRMSE ↓	Final Step nRMSE ↓	Late Step nRMSE ↓	Mean Corr. ↑	Energy Curve Error ↓
DPOT	0.2241 ± 0.0056	0.4233 ± 0.0116	0.3238 ± 0.0074	0.9628 ± 0.0002	0.0754 ± 0.0021
FactFormer	0.1818 ± 0.0049	0.2943 ± 0.0080	0.2636 ± 0.0062	0.9752 ± 0.0001	**0.0301 ± 0.0008**
F-FNO	0.1805 ± 0.0048	0.3021 ± 0.0082	0.2662 ± 0.0066	0.9753 ± 0.0239	0.0328 ± 0.0007
MemNO	0.1530 ± 0.0040	0.2397 ± 0.0065	0.2169 ± 0.0056	0.9828 ± 0.0034	0.0519 ± 0.0010
PDE-Refiner	0.1712 ± 0.0038	0.2786 ± 0.0075	0.2474 ± 0.0068	0.9786 ± 0.0063	0.0398 ± 0.0011
Poseidon	0.2457 ± 0.0053	0.4357 ± 0.0118	0.3485 ± 0.0094	0.9583 ± 0.0064	0.0867 ± 0.0017
CoMNO	**0.1493 ± 0.0035**	**0.2318 ± 0.0063**	**0.2105 ± 0.0054**	**0.9833 ± 0.0005**	0.0314 ± 0.0007

C.4 Rollout Stability

Recursive evaluation. We extend prediction beyond the supervised 20-step window using the four observed input frames specified in the main text. APEBench rollouts continue autoregressively for 197 steps; RealPDEBench rollouts reach 100 steps by recursively chaining five 20-step forecast blocks. Predicted states provide subsequent inputs, and CoMNO carries its recurrent memory along the rollout. Tables 19 to 23 report relative ℓ_2 error over the first H predicted steps and Final Step nRMSE at the longest evaluated horizon. The 100-step APEBench plots display prefixes of these extended rollouts. All metrics use inverse-normalized fields and the channel conventions defined in the main text. Entries report means and standard deviations across the same five random seeds; boldface marks the best mean in each column.

Complementary error measures. Prefix relative ℓ_2 aggregates squared errors and reference magnitudes over time, channels, and space before taking their norm ratio. Trajectory nRMSE averages the stepwise errors e_t from Eq. (9) with equal temporal weight, while Final Step nRMSE measures the last prediction. These summaries capture different aspects of rollout accuracy and can produce different rankings. Tables 24 to 28 report diagnostics over the full 197-step APEBench or 100-step RealPDEBench rollout. Their terminal entries are identical to those in the corresponding horizon tables. Late Step nRMSE averages e_t over the final half of the forecast; the final-quarter metric in the

commutator intervention remains a separate diagnostic. Mean Corr. follows the main-text stepwise Pearson correlation. Energy Curve Error averages the absolute predicted–reference energy difference divided by the reference energy plus 10^{-6}. Energy is the spatial mean of the sum of squared evaluated field components.

Accuracy across horizons. CoMNO achieves the lowest prefix relative ℓ_2 at every reported horizon on Burgers, Cylinder, and FSI, together with the lowest terminal nRMSE. The Burgers and Cylinder results agree with Figure 5. On KS, F-FNO leads at 20 steps, whereas CoMNO leads at 50, 100, and 197 steps; FactFormer achieves the lowest terminal nRMSE. On Kolmogorov flow, CoMNO leads at 20, 50, and 100 steps and has the lowest terminal nRMSE, while FactFormer achieves the lowest 197-step prefix error. The horizon and terminal results therefore distinguish accuracy accumulated over the forecast from accuracy at its endpoint.

Late-rollout and physical fidelity. CoMNO achieves the lowest Trajectory nRMSE on all five systems. It also leads Late Step nRMSE on Burgers, Cylinder, and FSI, while FactFormer leads this metric on KS and Kolmogorov flow. CoMNO has the highest Mean Corr. on four systems, with F-FNO leading on KS. Energy Curve Error is lowest for CoMNO on Burgers, KS, Kolmogorov flow, and Cylinder, and for FactFormer on FSI. Together, these diagnostics show that CoMNO's trajectory accuracy is accompanied by strong late-rollout performance and faithful energy evolution across diverse dynamics.

Table 29: Physical diagnostics on Burgers. StepRes measures the normalized governing-operator magnitude; the remaining metric measures error relative to the reference evolution.

Model	StepRes ↓	Energy Curve Rel. Error ↓
DPOT	0.1943 ± 0.0055	0.1459 ± 0.0042
FactFormer	0.0410 ± 0.0012	0.0842 ± 0.0024
F-FNO	0.9517 ± 0.0072	2.0950 ± 0.0061
MemNO	0.0428 ± 0.0012	0.1637 ± 0.0047
PDE-Refiner	0.7718 ± 0.0020	0.3525 ± 0.0101
Poseidon	0.2437 ± 0.0070	0.1956 ± 0.0056
CoMNO	**0.0354 ± 0.0010**	**0.0751 ± 0.0021**

Table 30: Physical diagnostics on KS. StepRes measures the normalized governing-operator magnitude; the remaining metric measures error relative to the reference evolution.

Model	StepRes ↓	Gradient-Energy Curve Rel. Error ↓
DPOT	0.3730 ± 0.0011	1.1010 ± 0.0095
FactFormer	0.0501 ± 0.0014	0.7391 ± 0.0198
F-FNO	0.2592 ± 0.0070	0.2457 ± 0.0066
MemNO	0.0972 ± 0.0026	0.2882 ± 0.0077
PDE-Refiner	0.5851 ± 0.0058	0.3383 ± 0.0091
Poseidon	0.3856 ± 0.0004	2.4023 ± 0.0045
CoMNO	**0.0091 ± 0.0002**	**0.1135 ± 0.0030**

Table 31: Physical diagnostics on Cylinder. Vorticity error measures spatial rotational-structure error; enstrophy error measures aggregate rotational-intensity discrepancy.

Model	Vorticity Rel. ℓ_2 ↓	Enstrophy Rel. Error ↓
DPOT	0.6939 ± 0.0198	0.5250 ± 0.0150
FactFormer	0.6920 ± 0.0197	0.3945 ± 0.0113
F-FNO	0.6125 ± 0.0175	0.2594 ± 0.0074
MemNO	0.6109 ± 0.0174	0.3747 ± 0.0107
PDE-Refiner	0.6085 ± 0.0174	0.3080 ± 0.0088
Poseidon	0.6727 ± 0.0192	0.3977 ± 0.0114
CoMNO	**0.5979 ± 0.0170**	**0.2508 ± 0.0075**

Table 32: Physical diagnostics on FSI. Vorticity error measures spatial rotational-structure error; enstrophy error measures aggregate rotational-intensity discrepancy.

Model	Vorticity Rel. ℓ_2 ↓	Enstrophy Rel. Error ↓
DPOT	0.6835 ± 0.0083	0.2028 ± 0.0054
FactFormer	0.5996 ± 0.0061	0.2648 ± 0.0071
F-FNO	0.5622 ± 0.0051	0.2492 ± 0.0067
MemNO	0.5126 ± 0.0038	0.2317 ± 0.0062
PDE-Refiner	0.5550 ± 0.0149	0.2320 ± 0.0062
Poseidon	0.7246 ± 0.0083	0.2137 ± 0.0054
CoMNO	**0.5109 ± 0.0037**	**0.2010 ± 0.0054**

Table 33: Physical diagnostics on Kolmogorov flow. StepRes measures the normalized governing-operator magnitude; KE and enstrophy errors measure deviations from the reference evolution.

Model	StepRes ↓	KE Error ↓	Enstrophy Error ↓
DPOT	0.395 ± 0.009	0.346 ± 0.009	0.396 ± 0.011
FactFormer	0.155 ± 0.004	0.304 ± 0.008	0.387 ± 0.016
F-FNO	0.655 ± 0.018	5.618 ± 0.151	0.420 ± 0.011
MemNO	0.151 ± 0.004	0.234 ± 0.006	$\mathbf{0.306 \pm 0.008}$
PDE-Refiner	0.380 ± 0.010	0.645 ± 0.017	0.509 ± 0.014
Poseidon	0.304 ± 0.008	0.561 ± 0.015	0.566 ± 0.015
CoMNO	$\mathbf{0.136 \pm 0.004}$	$\mathbf{0.175 \pm 0.005}$	0.368 ± 0.016

C.5 Physical Consistency

Diagnostics. StepRes and VortRel follow Eq. (11). StepRes measures the relative discrepancy between each next prediction and the APEBench reference time-stepper applied to the preceding prediction. KS and Kolmogorov flow use their corresponding reference time-steppers. VortRel measures relative error in the spatial vorticity field. Energy, gradient-energy, and kinetic-energy errors divide the temporal RMS of the predicted-reference observable difference by the reference curve's RMS, floored at ϵ_{phys}. Enstrophy error averages stepwise relative absolute differences. Gradient energy is the spatial mean of squared field gradients; enstrophy is the spatial mean of squared vorticity.

Result. Tables 29–33 report the physical diagnostics across all systems. CoMNO combines the lowest StepRes with the closest energy evolution on Burgers and gradient-energy evolution on KS. On Kolmogorov flow it leads StepRes and kinetic-energy accuracy; MemNO has the lowest enstrophy error. On Cylinder and FSI, CoMNO leads both vorticity and enstrophy errors. Together, these improvements connect lower field-prediction error to more accurate spatial vorticity and aggregate rotational intensity.

Table 34: Ablation on Burgers. Each variant removes one of the three modules evaluated in the main text.

Model	Rel. ℓ_2 ↓	Traj. nRMSE ↓	Final nRMSE ↓	HF Error ↓	Mean Corr. ↑
w/o Operator Commutator	0.0738 ± 0.0021	0.0783 ± 0.0023	0.1848 ± 0.0052	1.0506 ± 0.0302	0.9966 ± 0.0008
w/o Closure Memory	0.0716 ± 0.0020	0.0716 ± 0.0021	0.1946 ± 0.0056	0.5029 ± 0.0144	0.9958 ± 0.0009
w/o Closure Experts	0.0650 ± 0.0019	0.0650 ± 0.0018	0.1775 ± 0.0051	0.5624 ± 0.0062	0.9952 ± 0.0008
CoMNO	$\mathbf{0.0568 \pm 0.0016}$	$\mathbf{0.0568 \pm 0.0017}$	$\mathbf{0.1412 \pm 0.0040}$	$\mathbf{0.4288 \pm 0.0012}$	$\mathbf{0.9978 \pm 0.0002}$

Table 35: Ablation on Cylinder. Each variant removes one of the three modules evaluated in the main text.

Model	Rel. ℓ_2 ↓	Traj. nRMSE ↓	Final nRMSE ↓	HF Error ↓	Mean Corr. ↑
w/o Operator Commutator	0.1328 ± 0.0038	0.1252 ± 0.0036	0.1911 ± 0.0054	0.8212 ± 0.0234	0.9858 ± 0.0002
w/o Closure Memory	0.1405 ± 0.0040	0.1325 ± 0.0038	0.2034 ± 0.0058	0.8263 ± 0.0238	0.9841 ± 0.0009
w/o Closure Experts	0.1269 ± 0.0036	0.1210 ± 0.0035	0.1810 ± 0.0052	0.8113 ± 0.0031	0.9872 ± 0.0005
CoMNO	$\mathbf{0.1151 \pm 0.0034}$	$\mathbf{0.1161 \pm 0.0033}$	$\mathbf{0.1618 \pm 0.0048}$	$\mathbf{0.8006 \pm 0.0028}$	$\mathbf{0.9874 \pm 0.0004}$

C.6 Ablation Study

Module removals. We retrain the three variants in the main text under the same data splits, normalization, training windows, optimization, checkpoint-selection rule, and random seeds. Removing *Operator Commutator* disables cross-scale evidence and its state and memory discrepancy penalties. Removing *Closure Memory* disables the recurrent closure state and its coupling to local features. Removing *Closure Experts* disables the adaptive closure correction and its magnitude, diversity, and routing-entropy penalties. Tables 34 and 35 compare these variants with the full CoMNO model using the five tabulated main-text prediction metrics. The main-text ablation figure additionally visualizes RMSE, computed after inverse normalization, as a complementary field-scale diagnostic.

Contributions to prediction. The full model leads all five metrics on both systems. On Burgers, removing Operator Commutator causes the largest increase in relative ℓ_2, trajectory nRMSE, and HF Error. Removing Closure Memory causes the largest terminal-error increase, while removing Closure Experts produces the largest loss of spatial correlation. The three components therefore contribute differently to trajectory, endpoint, and spatial accuracy. On Cylinder, removing Closure Memory causes the largest degradation across all five metrics, making recurrent context particularly important for this measured-flow task. Removing either Operator Commutator or Closure Experts also worsens every metric, supporting their combined use with Closure Memory.

Table 36: Timing on Burgers.

Model	Training s/run ↓	Inference ms/window ↓
DPOT	3025.5905	1650.1274
FactFormer	59327.3033	1322.1693
F-FNO	**1856.1637**	884.5422
MemNO	54294.9713	7606.0083
PDE-Refiner	3397.8745	1636.9655
Poseidon	3123.3230	**735.9401**
CoMNO	10166.1472	770.1263

Table 37: Timing on KS.

Model	Training s/run ↓	Inference ms/window ↓
DPOT	10492.0486	1835.1998
FactFormer	44211.3145	1331.7986
F-FNO	**3695.9106**	809.9820
MemNO	64371.3925	4562.4230
PDE-Refiner	24183.0945	1740.7213
Poseidon	10604.6931	**587.7483**
CoMNO	15090.1923	918.9086

Table 38: Timing on Kolmogorov flow.

Model	Training s/run ↓	Inference ms/window ↓
DPOT	**3619.6117**	1857.3645
FactFormer	14742.0839	1344.8393
F-FNO	6178.5432	759.3715
MemNO	67652.6916	1839.8797
PDE-Refiner	8849.2068	1698.4824
Poseidon	5008.0558	**533.5882**
CoMNO	13160.5317	623.1423

Table 39: Timing on Cylinder.

Model	Training s/run ↓	Inference ms/window ↓
DPOT	20071.4915	256.9994
FactFormer	129743.4437	633.8724
F-FNO	**16556.6802**	397.9526
MemNO	282159.0751	750.1990
PDE-Refiner	32084.0626	571.6319
Poseidon	18560.8863	**167.8150**
CoMNO	98906.4574	937.2157

Table 40: Timing on FSI.

Model	Training s/run ↓	Inference ms/window ↓
DPOT	19286.0141	**141.7249**
FactFormer	157273.1993	893.1758
F-FNO	**15575.2524**	142.2549
MemNO	272467.3243	1357.5088
PDE-Refiner	26305.7649	1342.5798
Poseidon	18151.8495	481.2714
CoMNO	124610.9605	1672.3246

D HYPERPARAMETERS AND COMPUTATIONAL COST

Model configuration. The local encoder and base operator use width 64, with two local blocks and four base blocks. Closure Memory has width 32 and operates on a grid downsampled by a factor of two. The base operator retains 12 Fourier modes and applies a spectral block every two blocks. Operator Commutator uses scales $\{2, 4\}$ and is recomputed every two prediction steps; intermediate steps use zero current evidence. Gradients propagate through this construction. Closure Experts comprise five expert families, with the boundary expert active on non-periodic domains and circular padding on periodic domains.

Optimization. CoMNO uses AdamW with a constant learning rate of 2×10^{-4}, weight decay 10^{-4}, batch size 4, and gradient clipping at 1.0. Four observed frames initialize a 20-step training rollout, with teacher forcing disabled. Training runs for at most 500 epochs, with validation after each epoch and early stopping after 10 epochs without improvement. All methods use the same five random seeds, data splits, forecast windows, and metric implementation. Normalization statistics are fitted to each modality's training split.

Loss weights. The rollout loss has unit weight and includes a trajectory-relative term weighted by 0.05. Temporal weights have unit mean and increase linearly, with the final-step weight twice the first-step weight. The Fourier-amplitude loss has weight 5×10^{-3}. Raw state and memory commutator penalties have weights 0.02 and 0.01. Memory and closure-magnitude regularization each use 10^{-4}; expert-diversity and routing-entropy terms use 10^{-3} and 10^{-5}. The temporal component of memory regularization has coefficient 0.1. These weights define the default $\beta = 1$ setting in Appendix B.

Validation selection. Checkpoints minimize a validation score consisting of full-trajectory relative ℓ_2 plus half the late-window nRMSE and half the final-step nRMSE. The late window is the final third of the forecast. Architecture and loss weights are selected on Burgers validation data and transferred across systems, with physical channels, grid spacing, and boundary handling set for each task. Baselines retain their official architectures and pretrained initializations where applicable, while following the common training and evaluation protocol stated in the main text, including the shared epoch budget, validation schedule, and checkpoint criterion.

Computational cost. Tables 36–40 report training time per run and inference time per complete forecast window. Training includes optimization, validation, and checkpoint saving. Inference includes memory initialization, recursive prediction, commutator or refinement computations, denormalization, and device synchronization, with batch size one and gradients disabled. Runs use one NVIDIA RTX PRO 6000 GPU with 96 GB memory, PyTorch 2.8.0, and CUDA 12.8.

CoMNO combines competitive APEBench inference time with lower prediction errors. Its inference cost is higher on Cylinder and FSI. Each commutator update adds a provisional transition on the observation grid and each configured coarse grid, so the timing results account for the computation used to obtain the accuracy gains.