NOT ALL INTERACTIONS NEED TO BE GLOBAL: QUERY-CENTERED MULTI-SCALE SPATIAL OBSERVATION FOR NEURAL OPERATORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural operators provide resolution-independent surrogates for partial differential equations, but accurately modeling spatial dependencies across different ranges remains challenging. While local structures and broad spatial configurations jointly determine the solution, existing operators typically rely on global representations or extensive spatial interactions to acquire long-range context. We propose the *Multi-Scale Spatial Observation Operator* (MSO), which reformulates long-range dependency modeling as a query-centered observation problem. MSO constructs a hierarchical spatial representation and propagates information independently along high-to-low and low-to-high resolution paths. For each query coordinate, it extracts spatially aligned features from all resolution levels and propagation directions, and adaptively aggregates these $2(L+1)$ observations through scale-wise attention. This design enables broad spatial observation while maintaining sparse query-side interaction, avoiding dense query-to-token mixing. Experiments on diverse PDE benchmarks covering incompressible and compressible fluid dynamics, shallow-water flows, and diffusion-reaction dynamics demonstrate strong predictive performance with a substantially smaller parameter footprint than representative neural operators. Further analyses show that different query locations favor different observation scales, supporting the effectiveness of query-adaptive multi-scale spatial observation.

1 INTRODUCTION

Neural operator learning aims to approximate mappings between infinite-dimensional function spaces and has become an important framework for modeling PDE-governed physical systems (Kovachki et al., 2023; Hao et al., 2023a; Raissi et al., 2019). A central challenge is capturing spatial dependencies over different ranges (Brandstetter et al., 2023; Pfaff et al., 2021). In many dynamical systems, the evolution at a query coordinate depends simultaneously on localized structures, such as sharp gradients and boundary layers, and broader patterns, such as transport and advection (Brunton et al., 2020; Kochkov et al., 2021). Thus, accurate query prediction requires both local detail and broad spatial context (Gupta et al., 2021). This raises a question: *does accessing global context necessarily require global interaction?*

Existing neural operators typically obtain long-range dependencies through global representations or broad spatial interaction. Spectral operators such as FNO (Li et al., 2021) efficiently capture global information in the frequency domain, but spectral truncation may suppress localized high-frequency structures (Rahaman et al., 2019; John Xu et al., 2020). Attention- and token-based operators, including Oformer (Li et al., 2023), GNOT (Hao et al., 2023b), Transolver and Transolver++ (Wu et al., 2024; Luo et al., 2025), and CViT (Wang et al., 2025), provide flexible spatial interactions but often expose each query to a large set of tokens, coupling long-range context with dense interaction. Hierarchical operators such as U-NO (Rahman et al., 2023) introduce multi-resolution representations, yet the observation range of an individual query remains implicit. These approaches suggest that the challenge is not whether a query needs broad context, but how that context should be organized and accessed.

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

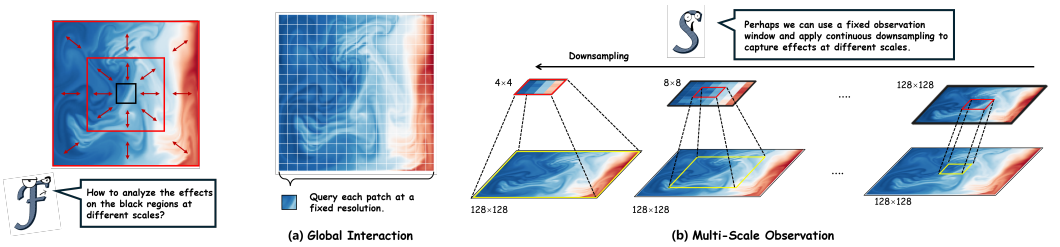

Figure 1: Multi-scale spatial observation. (a) *Global Interaction* captures long-range dependencies through broad spatial token interaction. (b) *Multi-Scale Observation* applies the same local window across progressively downsampled representations, expanding the effective physical observation range while preserving sparse, spatially aligned query interaction.

We observe that spatial resolution provides a natural prior for observation range. As shown in Figure 1, for a fixed local window of size $w \times w$, applying the same operator to a representation downsampled by 2^l corresponds to an effective physical footprint of approximately $(w2^l) \times (w2^l)$ in the original domain. Different resolutions therefore expose the same physical location to progressively broader neighborhoods, while preserving local interaction at each scale. This suggests an alternative to global query-to-token interaction: long-range context can be obtained through structured multi-scale observation, with spatial resolution controlling ***what range*** a query observes rather than ***how far*** it interacts.

Motivated by this perspective, we propose the *Multi-Scale Spatial Observation Operator* (MSO), a query-centered neural operator that explicitly organizes spatial context across resolutions and propagation directions. MSO builds hierarchical representations and propagates information through independent top-down and bottom-up paths enabling complementary information exchange across scales. For a query coordinate $\mathbf{x}$, it retrieves only the spatially aligned representation of $\mathbf{x}$ from each scale and direction, and adaptively aggregates the resulting $2(L+1)$ observations through scale-wise attention. Long-range context is thus acquired through cross-scale propagation rather than dense query-to-token interaction, enabling broad spatial observation with sparse query-side aggregation. The main contributions of this work are summarized as follows:

- We introduce a new perspective for spatial dependency modeling in neural operators: *global context does not necessarily require global interaction*. We formulate spatial resolution as an explicit prior for physical observation range, providing a structured alternative to dense spatial interaction.
- We propose the *Multi-Scale Spatial Observation Operator* (MSO), which combines hierarchical representations with independent top-down and bottom-up propagation. Each query accesses only spatially aligned observations across scales and directions, and aggregates them through scale-wise attention.
- We evaluate MSO on diverse PDE benchmarks covering incompressible and compressible fluid dynamics, shallow-water flows, and diffusion–reaction systems. MSO consistently achieves strong predictive performance against representative neural operators with a substantially smaller parameter budget, while ablations verify the benefits of multi-scale observation and complementary propagation paths.

2 Related Work

Neural Operators and Spectral Methods. Neural operators learn mappings between infinite-dimensional function spaces for resolution-independent PDE modeling (Lu et al., 2021; Li et al., 2021). DeepONet (Lu et al., 2021) uses branch and trunk networks to encode input functions and query coordinates, while FNO (Li et al., 2021) models integral kernels in the Fourier domain for efficient global interactions. Extensions include FFNO (Tran et al., 2023), Spherical FNO (Bonev et al., 2023), and Riesz Neural Operators (Liu et al., 2026), which improve efficiency or geometric

generality. These methods primarily capture spatial dependencies through global spectral representations.

Attention- and Transformer-Based Operators. Attention-based operators model spatial dependencies through learned token interactions (Li et al., 2023; Hao et al., 2023b; Wu et al., 2024; Wang et al., 2025). Oformer (Li et al., 2023) uses cross-attention for arbitrary query locations, GNOT (Hao et al., 2023b) handles heterogeneous geometric and physical inputs, and Transolver (Wu et al., 2024) and CViT (Wang et al., 2025) introduce physics-aware or patch-based token representations for spatial interaction. These approaches mainly obtain spatial context through interactions among distributed spatial tokens.

Hierarchical and Multi-Resolution Operators. Hierarchical operators exploit multi-resolution representations to capture spatial structures at different scales. U-NO (Rahman et al., 2023) adopts an encoder-decoder hierarchy, while WNO (Tripura & Chakraborty, 2023) and other multi-scale architectures further leverage localized representations across different resolutions to model spatial dependencies (Li et al., 2020; Liu et al., 2024). These methods organize information across scales without explicitly defining resolution as a query-dependent observation range.

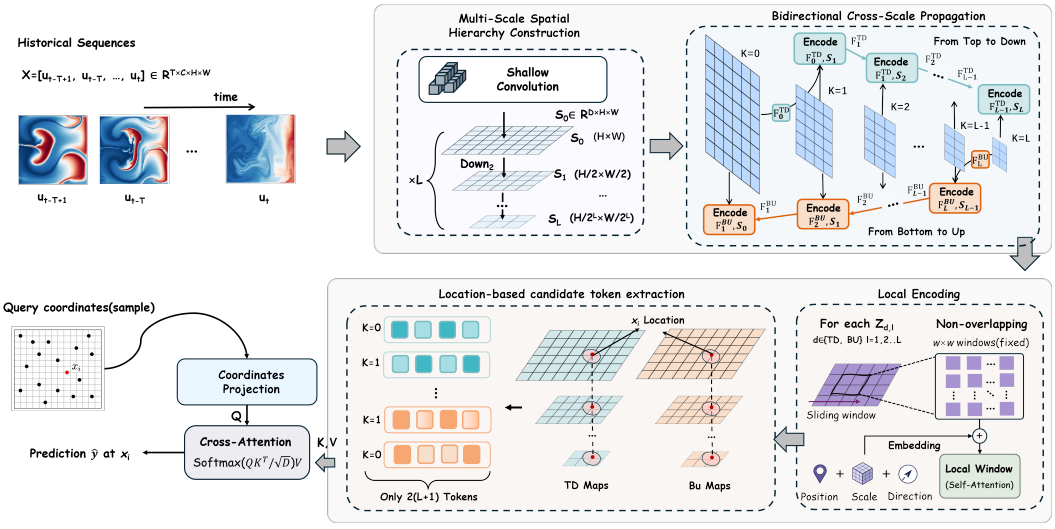

Figure 2: Architecture of Multi-Scale Spatial Operator. The framework features three stages: (1) *Multi-Scale Representation*: Temporal inputs are recursively downsampled into a feature pyramid. (2) *Bidirectional Propagation & Local Encoding*: Features are routed through independent bottom-up and top-down paths, then refined by local window self-attention using spatial, scale, and direction embeddings. (3) *Query-Centered Attention*: For any coordinate x_i, spatially aligned features are extracted across all scales. The query aggregates merely $2(L+1)$ co-located tokens via cross-attention to yield the point-wise prediction.

3 METHOD

As shown in Figure 2, our architecture consists of three stages: (1) **Hierarchical Multi-Scale Representation**, which maps the temporal history to a base spatial representation and recursively constructs a multi-resolution feature hierarchy; (2) **Bidirectional Propagation and Local Encoding**, which propagates information across resolution levels along two independent directions and performs localized window attention within each level; and (3) **Query-Centered Scale-Wise Attention**, which samples spatially aligned features at an arbitrary query coordinate and aggregates them across scales and propagation directions.

3.1 Problem Formulation

Let $\Omega \subset \mathbb{R}^2$ be a bounded spatial domain. We consider learning a solution operator $\mathcal{G}_\theta$ for two-dimensional partial differential equations (PDEs). Given a history of T consecutive system states,

$$\mathbf{X} = [\mathbf{u}_{t-T+1}, \ldots, \mathbf{u}_t] \in \mathbb{R}^{T\times H\times W\times C}, \tag{1}$$

where $H \times W$ denotes the spatial discretization and C the physical channel dimension, the objective is to predict the solution at a future time step for an arbitrary query coordinate $\mathbf{x} \in \Omega$:

$$\hat{\mathbf{u}}_{t+\Delta t}(\mathbf{x}) = \mathcal{G}_\theta(\mathbf{X}, \mathbf{x}) \in \mathbb{R}^C. \tag{2}$$

3.2 Multi-Scale Spatial Hierarchy Construction

We first fold the temporal history into the channel dimension,

$$\mathbf{X}_{\text{in}} \in \mathbb{R}^{TC\times H\times W}, \tag{3}$$

and use a shallow convolutional stem to obtain the finest-scale representation:

$$\mathbf{S}^0 = \Phi_0(\mathbf{X}_{\text{in}}) \in \mathbb{R}^{D\times H\times W}. \tag{4}$$

We then recursively construct a hierarchy of $L + 1$ resolutions:

$$\mathbf{S}^{l+1} = \Phi_{l+1}\left(\text{Down}_2(\mathbf{S}^l)\right), \qquad l = 0, \ldots, L-1, \tag{5}$$

where Down_2 downsamples the spatial resolution by a factor of two and Φ_{l+1} preserves the channel dimension D. Thus,

$$\mathbf{S}^l \in \mathbb{R}^{D\times H/2^l\times W/2^l}, \tag{6}$$

with $\mathbf{S}^0$ the finest and $\mathbf{S}^L$ the coarsest representation.

Physical Scale Interpretation. At resolution level l, a fixed local window of size $w \times w$ corresponds to an effective footprint of approximately $(w2^l) \times (w2^l)$ in the original spatial domain. Hence, increasing l provides progressively broader spatial context without enlarging the local interaction window.

3.3 Bidirectional Cross-Scale Propagation and Local Encoding

To exchange complementary information across resolutions, we introduce two independent propagation paths. The **Top-Down (TD) path** propagates features from the finest to progressively coarser resolutions, whereas the **Bottom-Up (BU) path** propagates features from the coarsest back toward finer resolutions.

Top-Down (TD) High-to-Low Resolution Path. The TD stream propagates information from $\mathbf{S}^0$ toward progressively coarser levels:

$$\mathbf{F}_l^{\text{TD}} = \begin{cases} \Phi_0^{\text{TD}}(\mathbf{S}^0), & l = 0, \\ \Phi_l^{\text{TD}}\left(\text{Concat}\left[\mathbf{S}^l, \text{Down}_2\left(\mathbf{F}_{l-1}^{\text{TD}}\right)\right]\right), & l = 1, \ldots, L. \end{cases} \tag{7}$$

Bottom-Up (BU) Low-to-High Resolution Path. The BU stream propagates information from $\mathbf{S}^L$ toward progressively finer levels:

$$\mathbf{F}_l^{\text{BU}} = \begin{cases} \Phi_L^{\text{BU}}(\mathbf{S}^L), & l = L, \\ \Phi_l^{\text{BU}}\left(\text{Concat}\left[\mathbf{S}^l, \text{Up}_2\left(\mathbf{F}_{l+1}^{\text{BU}}\right)\right]\right), & l = L-1, \ldots, 0. \end{cases} \tag{8}$$

The two paths remain independent throughout propagation, producing complementary directional representations $\{\mathbf{F}_l^{\text{TD}}, \mathbf{F}_l^{\text{BU}}\}$ at each resolution level.

Coordinate and Scale Embedding. For each direction $d \in \{\text{TD}, \text{BU}\}$ and level l, we project the propagated features and augment them with normalized spatial coordinates, scale embeddings, and direction embeddings:

$$\mathbf{Z}_l^d = \Gamma_l^d(\mathbf{F}_l^d) + \Pi_{\text{coord}}(\mathbf{P}_l) + \mathbf{e}_l^{\text{scale}} + \mathbf{e}_d^{\text{dir}}, \tag{9}$$

where $\mathbf{P}_l$ denotes the normalized spatial grid, Γ_l^d is a linear projection, and Π_{coord} is a coordinate MLP.

Local Window Self-Attention. Each $\mathbf{Z}_l^d$ is partitioned into non-overlapping $w \times w$ windows, within which standard multi-head self-attention is applied:

$$\widetilde{\mathbf{Z}}_l^d = \mathrm{WinAttn}(\mathbf{Z}_l^d) + \mathbf{Z}_l^d. \tag{10}$$

A position-wise FFN with residual connections follows the attention block. Since the window size is fixed, intra-scale computation remains linear in the number of spatial locations.

3.4 Query-Centered Scale-Wise Attention

Rather than exposing a query coordinate $\mathbf{x}$ to spatial tokens across the entire domain, MSO aggregates context only across resolution levels and propagation directions at the query location.

Query Feature Formulation. For an arbitrary query coordinate $\mathbf{x} \in \Omega$, we interpolate the latest observed state $\mathbf{u}_t$ at $\mathbf{x}$ and combine it with its coordinate embedding:

$$\mathbf{q}_\mathbf{x} = \mathbf{W}_s \mathbf{u}_t(\mathbf{x}) + \Psi_{\mathrm{coord}}(\mathbf{x}) \in \mathbb{R}^D. \tag{11}$$

Spatially Aligned Scale Sampling. For each resolution level l and direction d, we obtain the representation at $\mathbf{x}$ by bilinear interpolation:

$$\mathbf{z}_l^d(\mathbf{x}) = \mathrm{Sample}\left(\widetilde{\mathbf{Z}}_l^d, \mathbf{x}\right) \in \mathbb{R}^D. \tag{12}$$

The resulting spatially aligned observation set is

$$\mathcal{O}(\mathbf{x}) = \left\{\mathbf{z}_l^{\mathrm{TD}}(\mathbf{x})\right\}_{l=0}^{L} \cup \left\{\mathbf{z}_l^{\mathrm{BU}}(\mathbf{x})\right\}_{l=0}^{L}, \qquad |\mathcal{O}(\mathbf{x})| = 2(L+1). \tag{13}$$

Scale-Wise Cross-Attention and Prediction. Stacking the observations gives

$$\mathbf{M}_\mathbf{x} \in \mathbb{R}^{2(L+1)\times D}. \tag{14}$$

Using $\mathbf{q}_\mathbf{x}$ as the query, we perform cross-attention (Vaswani et al., 2023) over this scale-direction sequence:

$$\mathbf{Q}_\mathbf{x} = \mathbf{q}_\mathbf{x}\mathbf{W}_Q, \qquad \mathbf{K}_\mathbf{x} = \mathbf{M}_\mathbf{x}\mathbf{W}_K, \qquad \mathbf{V}_\mathbf{x} = \mathbf{M}_\mathbf{x}\mathbf{W}_V, \tag{15}$$

and obtain

$$\mathbf{h}_\mathbf{x} = \mathrm{Softmax}\left(\frac{\mathbf{Q}_\mathbf{x}\mathbf{K}_\mathbf{x}^\top}{\sqrt{D}}\right)\mathbf{V}_\mathbf{x}. \tag{16}$$

The prediction is then produced by a residual FFN and an output projection:

$$\mathbf{r}_\mathbf{x} = \mathrm{FFN}\left(\mathbf{q}_\mathbf{x} + \mathbf{h}_\mathbf{x}\right), \tag{17}$$

$$\hat{\mathbf{u}}_{t+\Delta t}(\mathbf{x}) = \Psi_{\mathrm{out}}\left(\mathbf{r}_\mathbf{x}\right). \tag{18}$$

Because each query attends to only $2(L+1)$ observations, the query decoding cost is $\mathcal{O}(Q(L+1)D)$ for Q evaluation points, independent of the total number of spatial tokens. This decouples spatial feature propagation from query aggregation while retaining multi-scale spatial context.

4 Experiments

We evaluate our method through benchmark comparisons, ablation studies, analyses of spatial observation, noise robustness, and computational efficiency. Unless otherwise specified, we use relative L_2 error as the evaluation metric. Detailed dataset specifications and training parameters are provided in Appendix A and C.

4.1 Experimental Setup

Benchmarks. For a comprehensive comparison across diverse physical dynamics, we evaluate our framework under two standard benchmark protocols: (1) **CViT Benchmark Setting:** Following CViT (Wang et al., 2025), we evaluate on incompressible Navier–Stokes (NS), Diffusion–Reaction (DR) and Shallow Water Equations (SWE) from PDEArena (Gupta & Brandstetter, 2022), as well as

Table 1: Performance comparison on the incompressible Navier–Stokes (NS), Compressible Navier–Stokes (CNS), and Diffusion–Reaction (DR) benchmarks.

Method	**Model**	**# Params**	L^2 **error** (%)		
			NS	**CNS**	**DR**
Spectral	FNO (Li et al., 2021)	0.5 M	9.12	9.60	12.00
	FFNO (Tran et al., 2023)	1.3 M	8.39	5.20	5.71
	RNO (Liu et al., 2026)	35 M	2.62	0.79	0.86
Attention	GNOT (Hao et al., 2023b)	1.8 M	17.20	4.20	3.11
	Oformer (Li et al., 2023)	1.9 M	13.50	6.25	1.92
	Transolver++ (Luo et al., 2025)	300 M	11.71	3.82	3.66
	CViT-S (Wang et al., 2025)	13 M	3.75	2.71	1.13
	CViT-B (Wang et al., 2025)	30 M	3.18	1.99	1.11
	CViT-L (Wang et al., 2025)	92 M	2.35	1.29	0.68
Foundation	MPP-S (McCabe et al., 2024)	30 M	–	–	1.12
	MPP-L (Pre-trained) (McCabe et al., 2024)	400 M	–	–	0.98
	DPOT-Ti (Hao et al., 2024)	7 M	12.50	3.97	3.21
	DPOT-S (Hao et al., 2024)	30 M	9.91	3.37	3.79
	DPOT-L (Pre-trained) (Hao et al., 2024)	500 M	7.98	2.16	2.32
	DPOT-L (Fine-tuned) (Hao et al., 2024)	500 M	2.78	1.31	0.73
	DPOT-H (Pre-trained) (Hao et al., 2024)	1.03 B	3.79	1.80	1.91
Ours	**MSO**	16 M	**1.18**	**0.31**	**0.19**

Compressible Navier–Stokes (CNS) from PDEBench (Takamoto et al., 2022). We strictly follow the data preprocessing, training schedules, and optimization configurations of CViT without introducing task-specific tuning. (2) **FNO Benchmark Setting:** To evaluate our model under an independent operator setup, we consider the 64×64 Navier–Stokes dataset from FNO (Li et al., 2021), strictly following the data preparation and training pipeline of Transolver++ (Luo et al., 2025).

Baselines. We compare our method with representative neural operator architectures, including spectral methods (FNO, FFNO, RNO), attention/token-based operators (GNOT, Oformer, Transolver, CViT), and foundation surrogates (DPOT, MPP). To ensure a fair comparison, all baseline models are trained using identical open-source scripts under unified parameter configurations. We report the best performance between our reproduced results and the metrics presented in their original literature.

Evaluation. We use relative L_2 error as the primary prediction metric. For the ablation experiments. All compared models are evaluated using the protocol associated with each benchmark.

4.2 Main Results

As reported in Table 1 (PDEArena and PDEBench), Table 2 (64×64 NS), and Table 3 (SWE), our method achieves the lowest relative L_2 error across all evaluated benchmarks.

On the 128×128 suite, MSO obtains relative L_2 errors of 0.0118, 0.0031, and 0.0019 on NS, CNS, and DR, respectively, while achieving an error of 0.0109 on SWE. On the 64×64 NS benchmark, MSO attains a relative L_2 error of 0.0147. The consistent performance across incompressible and compressible fluid dynamics, shallow-water flows, and diffusion–reaction systems demonstrates the effectiveness of the proposed multi-scale spatial observation mechanism across diverse PDE dynamics. Notably, MSO achieves these results with substantially fewer parameters than foundation-scale models such as DPOT, indicating that the observed performance is not attributable to increased parameter capacity.

Table 2: Performance comparison on the incompressible 64×64 NS benchmark.

Model	**# Params**	L^2 **error** (↓)
FNO	0.9 M	15.56%
CViT	13 M	7.94%
Transolver	-	9.00%
Transolver++	-	7.19%
RNO	35 M	4.02%
Ours	5 M	**1.47%**

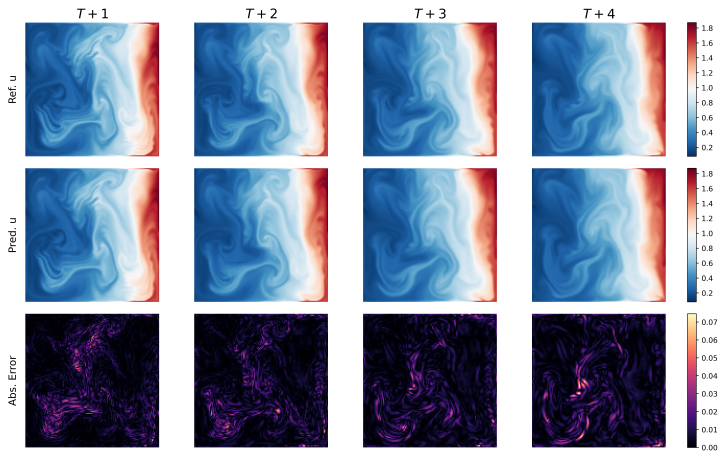

Figure 3: Incompressible Navier-Stokes (128×128) Autoregressive Forecasting.

4.3 Ablation Study

We investigate the contribution of the key components of MSO on the 64×64 NS benchmark. All variants use the same architecture, training protocol, and batch size of 8, with only the specified component modified.

Table 3: Performance comparison on the Shallow Water Equations (SWE) benchmark.

Model	**# Params**	L^2 **error** (↓)
FNO	268 M	3.97%
CViT	90 M	1.56%
RNO	35 M	2.48%
Ours	16 M	**1.09%**

We consider five variants. (1) *w/o Top-Down* removes the TD propagation path and retains only the BU observations. (2) *w/o Bottom-Up* removes the BU propagation path and retains only the TD observations. (3) *Uniform Fusion* replaces query-centered scale-wise attention with equal-weight averaging over all spatially aligned observations. (4) *Static Fusion* assigns learnable but query-independent weights to the aligned observations. (5) *Global Token Fusion* replaces query-centered aggregation with dense spatial token interaction, allowing each query to access all tokens across the multi-scale representations.

Table 4: Ablation study on the 64×64 NS dataset. Training memory corresponds to the peak consumption per epoch with a batch size of 8.

Model	**Tokens per Query**	**Rel.** L^2 **error** (↓)	**Peak Memory (MB)**
w/o Top-down	5	1.92%	806
w/o Bottom-up	5	1.88%	806
Uniform Fusion	10	2.45%	1,245
Static Fusion	10	2.29%	1,293
Global Token Fusion	10,912	3.30%	72,766
Ours	10	**1.47%**	1,308

As shown in Table 4, the full MSO achieves the lowest relative L_2 error of **1.47**%. Removing either propagation path leads to degraded performance, with *w/o Top-Down* and *w/o Bottom-Up* reaching **1.92**% and **1.88**%, respectively, indicating that the two directions provide complementary information across resolutions. Replacing adaptive aggregation with *Uniform Fusion* further increases the error to **2.45**%, showing that simply combining multi-scale observations is insufficient. *Static Fusion* improves upon uniform averaging, achieving **2.29**%, but remains inferior to the full query-adaptive design. This comparison indicates that the relative importance of different spatial observations depends on the query rather than being fixed across the domain.

In contrast, *Global Token Fusion* yields a relative L_2 error of **3.30**% while requiring approximately **55.6**× the GPU memory of MSO. In the 64×64 setting, dense fusion exposes each query to $11{,}000$ spatial tokens across the multi-scale representations, compared with only $2(L+1)$ spatially aligned observations in MSO. The resulting performance degradation and substantially higher memory cost demonstrate that broad spatial context can be obtained without dense query-to-token interaction. Together, these results support the complementary roles of bidirectional propagation and query-adaptive multi-scale observation in MSO.

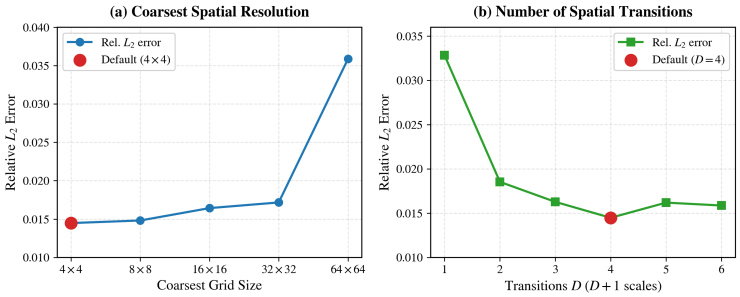

Figure 4: Analysis of spatial observation hierarchy on the 64×64 NS benchmark. **(a)** Prediction error across different coarsest observation levels (from 4×4 to 64×64). **(b)** Effect of hierarchy depth (D transitions across fixed boundaries 64×64 to 4×4).

4.4 Analysis of Spatial Observation

Coarsest Resolution and Receptive Field. We first evaluate the effective spatial observation range by varying the coarsest grid size from 64×64 down to 4×4 while preserving the finest resolution (Figure 4a). Relying solely on the single-scale 64×64 representation results in severe performance degradation, with the relative L_2 error increasing from 0.01448 to 0.03586. Progressively incorporating coarser observations monotonically improves prediction fidelity. This confirms that local patch representations alone cannot resolve long-range convective dynamics, whereas broader spatial views provide essential macroscopic context that anchors query predictions.

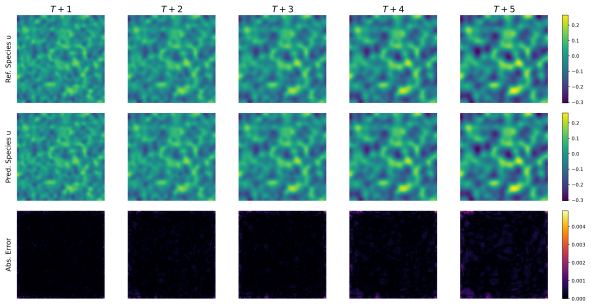

Figure 5: Diffusion–Reaction (128×128) Autoregressive Forecasting.

Granularity of Scale Transitions. To isolate the effect of multi-scale hierarchy depth without altering the receptive field boundaries, we fix the terminal resolutions to 64×64 and 4×4 and adjust the number of intermediate transition levels D ($D+1$ total scales; Figure 4b). Direct scale bridging ($D = 1$) yields an inferior error of 0.03282, reflecting insufficient information exchange across disparate spatial frequencies. The error reaches a clear optimum of 0.01448 at $D = 4$. Further increasing scale density ($D \geq 5$) yields no additional gain and slightly degrades accuracy (e.g., 0.01586 at $D = 6$), indicating that overly dense subdivision introduces representational redundancy without supplementary predictive utility.

Query-Adaptive Scale Selection We analyze the scale-wise attention of MSO on the 64×64 NS benchmark. For each query $\mathbf{x}$, MSO assigns weights $\alpha_l^d(\mathbf{x})$ to the $2(L+1)$ spatially aligned observations across resolution levels and propagation directions. As shown in Figure 6, different query

locations exhibit distinct observation preferences: smooth regions tend to emphasize coarser representations, while regions with sharper spatial variations rely more on finer scales. These results indicate that the effective observation range is query-dependent, supporting adaptive scale-wise aggregation in MSO.

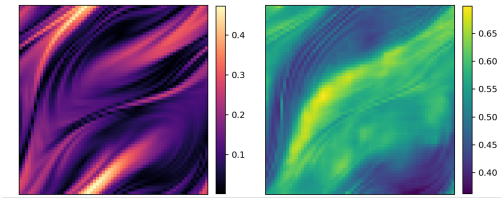

Figure 6: **Query-adaptive scale selection on the NS benchmark.** (a): Prediction physical field. (b): Scale preference scores (brighter indicates coarser scales). Smooth regions attend to coarser scales (brighter), while areas with sharp spatial variations focus on finer scales (darker).

Result Visualization. Finally, we examine the qualitative fidelity of our operator by visualizing predicted solution fields and corresponding point-wise absolute error distributions across diverse fluid benchmarks (Figures 3, Figures 5 and Figures 7).

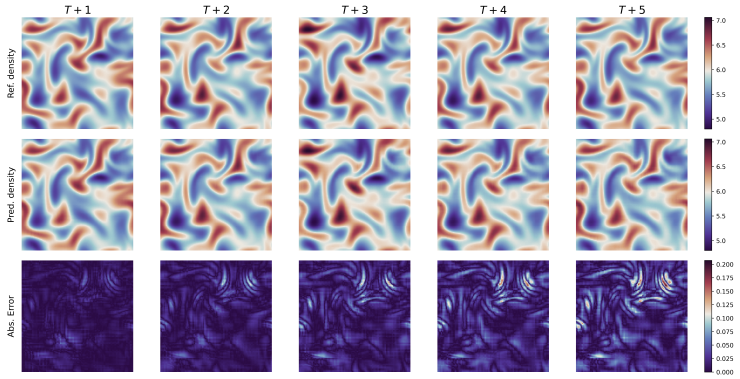

Figure 7: Compressible Navier-Stokes (128×128) Autoregressive Forecasting.

5 Discussion

In this work, we proposed a query-centered multi-scale neural operator that organizes spatial context across resolution levels. By combining independent high-to-low and low-to-high propagation paths with spatially aligned query sampling, MSO provides broad spatial observations without dense query-to-token interaction. Experiments across PDEArena, PDEBench, and FNO benchmarks demonstrate strong predictive performance across diverse PDE dynamics with a substantially smaller parameter footprint than representative baselines.

Limitations. The current framework focuses on spatial observation and does not explicitly address temporal error accumulation. Its performance may therefore degrade during long-horizon autoregressive rollouts.

Future Work.A natural direction is to extend the proposed observation mechanism to the temporal dimension for more stable long-horizon prediction. The query-side interaction cost is bounded by $\mathcal{O}(Q(L+1)D)$, compared with $\mathcal{O}(QND)$ for dense spatial attention, motivating further exploration of scalable PDE solvers on complex geometries. This property also motivates applying the framework to higher-resolution, three-dimensional, and multi-physics settings.

AI USE STATEMENT

In this work, we used generative AI tools to assist with coding and manuscript writing. Generative AI was not used to produce conclusions unsupported by objective experimental evidence. All AI-assisted content, including but not limited to code and manuscript text, was manually reviewed and verified by the authors. We take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

To ensure the full reproducibility of our empirical results, we provide detailed model architectures, multi-scale token extraction procedures, training hyperparameters, and evaluation protocols in the main text and Appendix A. We explicitly specify the physical governing equations, grid resolutions, and data splits for all evaluated benchmark datasets (including Navier-Stokes, Compressible Navier-Stokes, Shallow Water Equations, and Diffusion-Reaction) in Appendix C. Furthermore, all random seeds, initialization schemes, and hardware-specific deterministic flags are strictly fixed as documented. To guarantee a fair and standardized comparison, all baseline models are implemented and evaluated using identical open-source scripts and unified configurations. The complete source code and dataset downloading instructions for the proposed framework will be made publicly available upon acceptance.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. URL https://arxiv.org/abs/1607.06450.

Boris Bonev, Thorsten Kurth, Christian Hundt, Jaideep Pathak, Maximilian Baust, Karthik Kashinath, and Anima Anandkumar. Spherical fourier neural operators: Learning stable dynamics on the sphere, 2023. URL https://arxiv.org/abs/2306.03838.

Johannes Brandstetter, Daniel Worrall, and Max Welling. Message passing neural pde solvers, 2023. URL https://arxiv.org/abs/2202.03376.

Steven L. Brunton, Bernd R. Noack, and Petros Koumoutsakos. Machine learning for fluid mechanics. *Annual Review of Fluid Mechanics*, 52(1):477–508, January 2020. ISSN 1545-4479. doi: 10.1146/annurev-fluid-010719-060214. URL http://dx.doi.org/10.1146/annurev-fluid-010719-060214.

Gaurav Gupta, Xiongye Xiao, and Paul Bogdan. Multiwavelet-based operator learning for differential equations. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

Jayesh K. Gupta and Johannes Brandstetter. Towards multi-spatiotemporal-scale generalized pde modeling, 2022. URL https://arxiv.org/abs/2209.15616.

Zhongkai Hao, Songming Liu, Yichi Zhang, Chengyang Ying, Yao Feng, Hang Su, and Jun Zhu. Physics-informed machine learning: A survey on problems, methods and applications, 2023a. URL https://arxiv.org/abs/2211.08064.

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*, ICML'23. JMLR.org, 2023b.

Zhongkai Hao, Chang Su, Songming Liu, Julius Berner, Chengyang Ying, Hang Su, Anima Anandkumar, Jian Song, and Jun Zhu. Dpot: auto-regressive denoising operator transformer for large-scale pde pre-training. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2023. URL https://arxiv.org/abs/1606.08415.

Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J Henaff, Matthew Botvinick, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=fILj7WpI-g.

Zhi-Qin John Xu, Yaoyu Zhang, Tao Luo, Yanyang Xiao, and Zheng Ma. Frequency principle: Fourier analysis sheds light on deep neural networks. *Communications in Computational Physics*, 28(5):1746–1767, November 2020. ISSN 1991-7120. doi: 10.4208/cicp.oa-2020-0085. URL http://dx.doi.org/10.4208/cicp.OA-2020-0085.

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), May 2021. ISSN 1091-6490. doi: 10.1073/pnas.2101784118. URL http://dx.doi.org/10.1073/pnas.2101784118.

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.

Zijie Li, Kazem Meidani, and Amir Barati Farimani. Transformer for partial differential equations' operator learning, 2023. URL https://arxiv.org/abs/2205.13671.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Multipole graph neural operator for parametric partial differential equations, 2020. URL https://arxiv.org/abs/2006.09535.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations, 2021. URL https://arxiv.org/abs/2010.08895.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection, 2017. URL https://arxiv.org/abs/1612.03144.

Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the CoordConv solution. In *Advances in Neural Information Processing Systems*, 2018a.

Shouyi Liu, Xiaokang Yang, and Yuntian Chen. Riesz neural operator for solving partial differential equations. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Vjw7q1quNt.

Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation, 2018b. URL https://arxiv.org/abs/1803.01534.

Xinliang Liu, Bo Xu, Shuhao Cao, and Lei Zhang. Mitigating spectral bias for the multiscale operator learning, 2024. URL https://arxiv.org/abs/2210.10890.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. URL https://arxiv.org/abs/2103.14030.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, March 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00302-5. URL http://dx.doi.org/10.1038/s42256-021-00302-5.

Huakun Luo, Haixu Wu, Hang Zhou, Lanxiang Xing, Yichen Di, Jianmin Wang, and Mingsheng Long. Transolver++: An accurate neural solver for pdes on million-scale geometries, 2025. URL https://arxiv.org/abs/2502.02414.

Michael McCabe, Bruno Régaldo-Saint Blancard, Liam Parker, Ruben Ohana, Miles Cranmer, Alberto Bietti, Michael Eickenberg, Siavash Golkar, Geraud Krawezik, Francois Lanusse, Mariel Pettee, Tiberiu Tesileanu, Kyunghyun Cho, and Shirley Ho. Multiple physics pretraining for spatiotemporal surrogate models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. *PyTorch: an imperative style, high-performance deep learning library*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter W. Battaglia. Learning mesh-based simulation with graph networks, 2021. URL https://arxiv.org/abs/2010.03409.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5301–5310. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/rahaman19a.html.

Md Ashiqur Rahman, Zachary E. Ross, and Kamyar Azizzadenesheli. U-no: U-shaped neural operators, 2023. URL https://arxiv.org/abs/2204.11127.

M. Raissi, P. Perdikaris, and G.E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2018.10.045. URL https://www.sciencedirect.com/science/article/pii/S0021999118307125.

Makoto Takamoto, Timothy Praditia, Raphael Leiteritz, Dan MacKinlay, Francesco Alesiani, Dirk Pflüger, and Mathias Niepert. Pdebench: an extensive benchmark for scientific machine learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Mingxing Tan, Ruoming Pang, and Quoc V. Le. Efficientdet: Scalable and efficient object detection, 2020. URL https://arxiv.org/abs/1911.09070.

Alasdair Tran, Alexander Mathews, Lexing Xie, and Cheng Soon Ong. Factorized fourier neural operators, 2023. URL https://arxiv.org/abs/2111.13802.

Tapas Tripura and Souvik Chakraborty. Wavelet neural operator for solving parametric partial differential equations in computational mechanics problems. *Computer Methods in Applied Mechanics and Engineering*, 404:115783, 2023. ISSN 0045-7825. doi: https://doi.org/10.1016/j.cma.2022.115783. URL https://www.sciencedirect.comscience/article/pii/S0045782522007393.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Sifan Wang, Jacob H Seidman, Shyam Sankaran, Hanwen Wang, George J. Pappas, and Paris Perdikaris. Cvit: Continuous vision transformer for operator learning, 2025. URL https://arxiv.org/abs/2405.13998.

Haixu Wu, Huakun Luo, Haowen Wang, Jianmin Wang, and Mingsheng Long. Transolver: A fast transformer solver for pdes on general geometries, 2024. URL https://arxiv.org/abs/2402.02366.

Yuxin Wu and Kaiming He. Group normalization, 2018. URL https://arxiv.org/abs/1803.08494.

A Model Architecture and Implementation Details

This section details the algorithmic execution and architectural configurations of the proposed Multi-Scale Observation (MSO) framework.

A.1 Algorithm Flowchart and Pseudocode

Algorithm 1 outlines the forward pass for computing query predictions. The spatial context is organized into $L + 1$ resolution levels. The framework maintains two independent propagation paths: a top-down pathway and a bottom-up pathway, yielding two sets of hierarchical feature maps.

For an arbitrary continuous query coordinate $\mathbf{x}_q$, the algorithm identifies the nearest spatially aligned discrete index at each resolution level l. Exactly one token is extracted from the top-down feature map, and one from the bottom-up feature map at that specific spatial location. This extraction process yields a strictly bounded token set containing $2(L + 1)$ tokens for each query. A cross-attention mechanism is subsequently applied exclusively to this set to decode the final prediction.

Algorithm 1 Forward of MSO

Require: Top-down feature pyramids $\mathcal{P}_{TD} = \{H_{TD}^{(0)}, \ldots, H_{TD}^{(L)}\}$, Bottom-up feature pyramids $\mathcal{P}_{BU} = \{H_{BU}^{(0)}, \ldots, H_{BU}^{(L)}\}$, Query coordinates $\mathbf{x}_q$.
Ensure: Predicted state $\mathbf{y}_q$ at $\mathbf{x}_q$.
1: Initialize the spatially aligned token set $S_q \leftarrow \emptyset$
2: **for** $l = 0$ to L **do**
3: Compute the corresponding discrete spatial index $\text{idx}_l(\mathbf{x}_q)$ at resolution level l
4: Extract top-down token: $t_{TD}^{(l)} \leftarrow H_{TD}^{(l)}[\text{idx}_l(\mathbf{x}_q)]$
5: Extract bottom-up token: $t_{BU}^{(l)} \leftarrow H_{BU}^{(l)}[\text{idx}_l(\mathbf{x}_q)]$
6: Append tokens to the query set: $S_q \leftarrow S_q \cup \{t_{TD}^{(l)}, t_{BU}^{(l)}\}$
7: **end for**
8: Process query embedding: $\mathbf{q} \leftarrow \text{Linear}(\mathbf{x}_q)$
9: Compute cross-attention using $\mathbf{q}$ and S_q: $\mathbf{z}_q \leftarrow \text{CrossAttention}(\text{Query} = \mathbf{q}, \text{Key/Value} = S_q)$
10: Decode the aggregated spatial feature: $\mathbf{y}_q \leftarrow \text{Decoder}(\mathbf{z}_q)$
11: **return** $\mathbf{y}_q$

A.2 Detailed Network Architecture

The input historical states are initially reshaped from $[B, T, H, W, C]$ to $[B, T \times C, H, W]$. The model utilizes a shared convolutional pyramid, followed by independent top-down and bottom-up branches, related to multi-scale feature propagation in feature-pyramid networks (Lin et al., 2017; Liu et al., 2018b). Unlike fused bidirectional pyramids (Tan et al., 2020), the two branches remain separate. For an architecture with $L + 1$ spatial scales, the two branches produce a total of $2(L + 1)$ candidates per queried coordinate. Each candidate consists of a physical-state delta and a token of dimension `delta_dim`. Windowed self-attention (Liu et al., 2021) is applied to refine tokens within each scale, followed by a query-to-candidate multi-head cross-attention mechanism Vaswani et al. (2023); Jaegle et al. (2022) that fuses all $2(L + 1)$ candidates at each requested coordinate.

The residual convolutional blocks (He et al., 2015) comprise two 3×3 convolutions, GroupNorm (Wu & He, 2018) (using the largest channel divisor not exceeding 8), and GELU activations (Hendrycks & Gimpel, 2023). Spatial downsampling is performed via $2\times$ average pooling, and upsampling uses bilinear interpolation. Coordinates are normalized to $[0, 1]^2$; explicitly providing spatial coordinates follows coordinate-aware feature design (Liu et al., 2018a). The input channels (`in_channels`) strictly equal the output dimension (`out_dim`), as the network predicts the physical increment added to the latest observed state.

Initialization and Normalization. LayerNorm (Ba et al., 2016) operations use $\epsilon = 10^{-5}$, and multi-head attention is configured with zero dropout. Scale, direction, and candidate embeddings are initialized from a zero-mean Gaussian distribution with a standard deviation of 0.02. To bias the

initial operator toward identity (persistence), the physical delta heads and the final residual head are initialized using a zero-mean Gaussian with a standard deviation of 0.001 and zero bias. All other convolutional and linear layers follow the PyTorch 2.9.0 default initialization (Paszke et al., 2019). The random seed is fixed to 42 for Python, NumPy, and PyTorch.

Table 5 details the specific architectural configurations across different benchmark datasets.

Table 5: Detailed architectural configurations for evaluated datasets.

Configuration	NS64	NS128	CNS	SWE	DR
Grid Resolution	64×64	128×128	128×128	96×192	128×128
History $\times$ Channels	10×1	10×3	10×4	2×2	2×2
Scale Feature Channels	32,64,128,128,128	32,64,128,256,256,256	32,64,128,128,128,128	32,64,128,256,256,256	32,64,128,128,128,128
Base Dimension	32	32	32	32	32
Maximum Dimension	128	256	256	256	256
Token Dimension	64	64	64	64	64
Attention Heads	4	8	8	8	8
Attention Depth	3	1	1	1	1
Window Size	4	4	4	4	4
MLP Ratio	1	2	2	2	2
Padding Mode	circular	circular	circular	zeros	circular
Total Parameters	4,697,645	15,902,861	15,902,861	15,891,582	15,902,861

A.3 Optimization and Hyperparameters

All models are optimized using the AdamW optimizer with $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$, and without AMSGrad. The learning rate is scheduled using the OneCycleLR policy, updated at every optimizer step. The scheduler uses cosine annealing with `pct_start` $= 0.3$, `div_factor` $= 25$, `final_div_factor` $= 10000$, and disabled three-phase scheduling.

Training is executed with CUDA mixed precision, utilizing float16 autocast and `GradScaler`, while the delta Mean Squared Error (MSE) is explicitly accumulated in float32 for numerical stability. Deterministic operations are enforced by setting `cudnn.deterministic = True` and `cudnn.benchmark = False`. Data loaders are configured with zero workers (`num_workers = 0`), dropping no incomplete batches (`drop_last = False`), and shuffling enabled exclusively during training. Table 6 summarizes the dataset-specific hyperparameters.

It should be noted that while the total number of training epochs is uniformly set to 300, the disparate number of temporal windows across datasets results in significantly different optimizer update counts. Specifically, the total updates over 300 epochs are 37,500 (NS64), 1,560,000 (NS128), 1,856,400 (CNS), 3,780,000 (SWE), and 152,100 (DR).

Table 6: Hyperparameter settings for model optimization.

Hyperparameter	NS64	NS128	CNS	SWE	DR
Epochs	300	300	300	300	300
Batch Size (Train/Test)	$8/8$	$4/4$	$16/16$	$4/4$	$16/16$
Max Learning Rate	1×10^{-3}	1×10^{-3}	1×10^{-3}	1×10^{-3}	5×10^{-4}
Weight Decay	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}	1×10^{-5}
Gradient Clipping	none	10.0	none	none	1.0
Train Queries per Sample	4,096 (all)	1,024	16,384 (all)	18,432 (all)	1,024
Evaluation Queries	full grid	full grid	full grid	full grid	full grid
Steps per Epoch	125	5,200	6,188	12,600	507

B Extended Complexity Analysis

This section provides a theoretical derivation of the computational complexity for the query-interaction mechanism, comparing standard global attention models with the proposed Multi-Scale Observation (MSO) framework.

Let $N = H \times W$ denote the total number of spatial tokens at the native grid resolution, and Q denote the number of evaluation queries. Let D represent the hidden embedding dimension of the tokens.

Global Spatial Mixing Complexity. In standard token-based neural operators (e.g., global Vision Transformers or unconstrained cross-attention decoders), the final query prediction requires computing the attention weights between all Q queries and all N spatial tokens. The computational complexity of this operation is $\mathcal{O}(Q \cdot N \cdot D)$. When evaluating the full spatial grid, $Q \approx N$, resulting in a quadratic complexity of $\mathcal{O}(N^2 \cdot D)$. As the physical domain scales to higher resolutions or three-dimensional systems ($N = H \times W \times D_z$), the term N grows cubically, making the $\mathcal{O}(Q \cdot N)$ complexity a primary bottleneck for memory and computation.

MSO Query Complexity. The MSO framework completely decouples the query interaction cost from the native grid size N. As detailed in Algorithm 1, for any arbitrary query coordinate, the model extracts exactly one spatially aligned token from the top-down pathway and one from the bottom-up pathway across $L + 1$ resolution levels. Consequently, the size of the key-value context set for each query is strictly bounded to $K = 2(L + 1)$ tokens.

The cross-attention mechanism in MSO is restricted to interacting the Q queries exclusively with their respective K multi-scale tokens. The computational complexity of this decoding step is therefore bounded by:

$$\mathcal{O}(Q \cdot K \cdot D) = \mathcal{O}(Q \cdot L \cdot D) \tag{19}$$

Since the number of spatial scales L is a small constant structurally independent of N (e.g., $L = 5$ or 6, whereas $N = 16,384$ for a 128×128 grid), the condition $L \ll N$ strictly holds. By reducing the asymptotic complexity from $\mathcal{O}(Q \cdot N)$ to $\mathcal{O}(Q \cdot L)$, the MSO framework avoids dense spatial mixing. This $\mathcal{O}(Q \cdot L)$ bottleneck-free property structurally resolves the memory explosion issue, providing a scalable mathematical foundation for extending the operator to high-resolution and 3D physical modeling.

C Experimental Setup and Datasets

This section details the governing equations of the physical systems evaluated in our benchmark suite, the implementation protocols for the baseline models, and the mathematical definition of the evaluation metrics.

C.1 Governing Equations and Datasets

Our framework is evaluated on five benchmark configurations covering incompressible and compressible fluid dynamics, diffusion–reaction systems, and shallow-water dynamics. For the NS128, CNS, DR, and SWE benchmarks, we strictly follow the dataset construction, spatial and temporal discretization, task formulation, and train/test partitions of the publicly released CViT implementation (Wang et al., 2025). For the NS64 benchmark, we strictly follow the standard benchmark configuration and data preparation protocol of Transolver (Luo et al., 2025).

Incompressible Navier–Stokes (NS). The 2D incompressible Navier–Stokes equations model the dynamics of viscous fluids. We evaluate on two standard configurations. The first is the canonical NS64 benchmark, which models the vorticity field $w(x, t)$ at a resolution of 64×64. The model takes 10 historical time steps with one vorticity channel as input to predict the subsequent state. The dataset construction and benchmark protocol strictly follow Transolver (Luo et al., 2025). The second configuration is the PDEArena NS128 dataset at a 128×128 resolution. The dataset contains 7,800 trajectories, with 6,500 trajectories for training and 1,300 for testing, and each trajectory contains 14 temporal frames. Following the CViT benchmark protocol, the model takes the preceding 10 time steps to predict the next time step, and performance is evaluated over a four-step autoregressive rollout (Wang et al., 2025).

Compressible Navier–Stokes (CNS). The 2D compressible Navier–Stokes equations involve the conservation of mass, momentum, and energy, and exhibit complex phenomena such as shock waves. The benchmark is taken from PDEBench and follows the configuration used by CViT. The dataset contains 9,000 training trajectories and 1,000 test trajectories. The input consists of preceding temporal observations and the task is formulated as next-state prediction followed by autoregressive rollout evaluation. The remaining dataset-specific parameters and preprocessing strictly follow the publicly released CViT implementation (Wang et al., 2025; Takamoto et al., 2022).

Diffusion–Reaction (DR). The 2D diffusion–reaction equations govern the spatiotemporal evolution of coupled concentration fields. The benchmark is based on the PDEBench dataset. The original data are generated on a 512×512 spatial grid with 501 temporal points and are downsampled to 128×128 with 101 temporal points for the benchmark used by CViT. The resulting dataset contains 900 training trajectories and 100 test trajectories. The model uses the same next-state prediction formulation and data preprocessing as specified by the publicly released CViT implementation (Wang et al., 2025; Takamoto et al., 2022).

Shallow Water Equations (SWE). The 2D Shallow Water Equations describe the propagation of gravity waves in a fluid layer. The SWE benchmark is generated using the PDEArena framework and is evaluated on a rectangular grid with a spatial resolution of 96×192. The underlying spatial grid is 192×96, with spatial resolutions of $h = 1.875^\circ$ and $\Delta y = 3.75^\circ$, and a temporal resolution of $\Delta t = 48\,\mathrm{h}$. Each trajectory contains 11 temporal frames. The dataset consists of 6,600 trajectories, including 5,600 training trajectories and 1,000 test trajectories. Following CViT, the input utilizes the preceding two time steps to predict the next time step, and evaluation is performed over a five-step autoregressive rollout (Wang et al., 2025; Gupta & Brandstetter, 2022).

C.2 Baseline Implementation Protocols

To ensure a rigorously fair and standardized comparison, we employ a "best-of-both" evaluation strategy. Baseline models, including Fourier Neural Operator (FNO), Continuous Vision Transformer (CViT), and Transolver, are trained and evaluated using identical open-source evaluation scripts and data loaders under unified parameter configurations (e.g., matching the number of epochs and using identical query coordinates).

For baseline metrics reported in Tables 1, 2, and 3, we select the lowest relative L_2 error between our internally reproduced results (using their official source code) and the metrics officially reported in their respective original literature. This protocol isolates the architectural efficiency of our Multi-Scale Observation (MSO) method from extraneous hyperparameter tuning.

C.3 Evaluation Metrics

The predictive performance of all models is measured using the relative L_2 error. Let $\mathbf{y}_q$ denote the ground truth physical state at spatial coordinate $\mathbf{x}_q$, and $\hat{\mathbf{y}}_q$ denote the corresponding model prediction. For a spatial grid containing N total query points, the relative L_2 error for a single temporal prediction is defined as:

$$\text{Relative } L_2 = \frac{\sqrt{\sum_{q=1}^{N} \|\mathbf{y}_q - \hat{\mathbf{y}}_q\|_2^2}}{\sqrt{\sum_{q=1}^{N} \|\mathbf{y}_q\|_2^2}} \tag{20}$$

The final metrics reported in our tables represent the average relative L_2 error evaluated across all samples in the designated test sets.

D Additional Qualitative Visualizations

This section provides supplementary visual analyses to illustrate the predictive behavior and internal attention mechanisms of the Multi-Scale Observation (MSO) framework.

D.1 Spatiotemporal Rollouts and Error Distributions

To evaluate the model's capacity to capture long-term physical dynamics, we visualize the autoregressive spatiotemporal rollouts, As shown in Figure 8 - Figure 14.

D.2 Visualization of Multi-Scale Attention Allocation

As detailed in Section A.1, the query mechanism in MSO is strictly bounded to interacting with $2(L+1)$ spatially aligned tokens. Figure 15 visualizes the cross-attention weights assigned to these hierarchical tokens for a selected query coordinate.

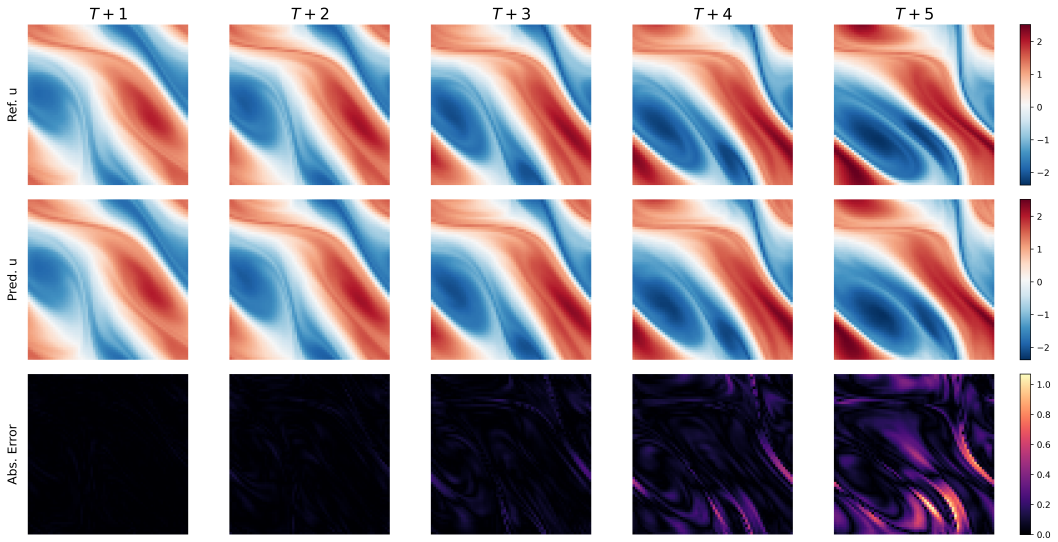

Figure 8: Incompressible Navier-Stokes (64×64) Autoregressive Forecasting.

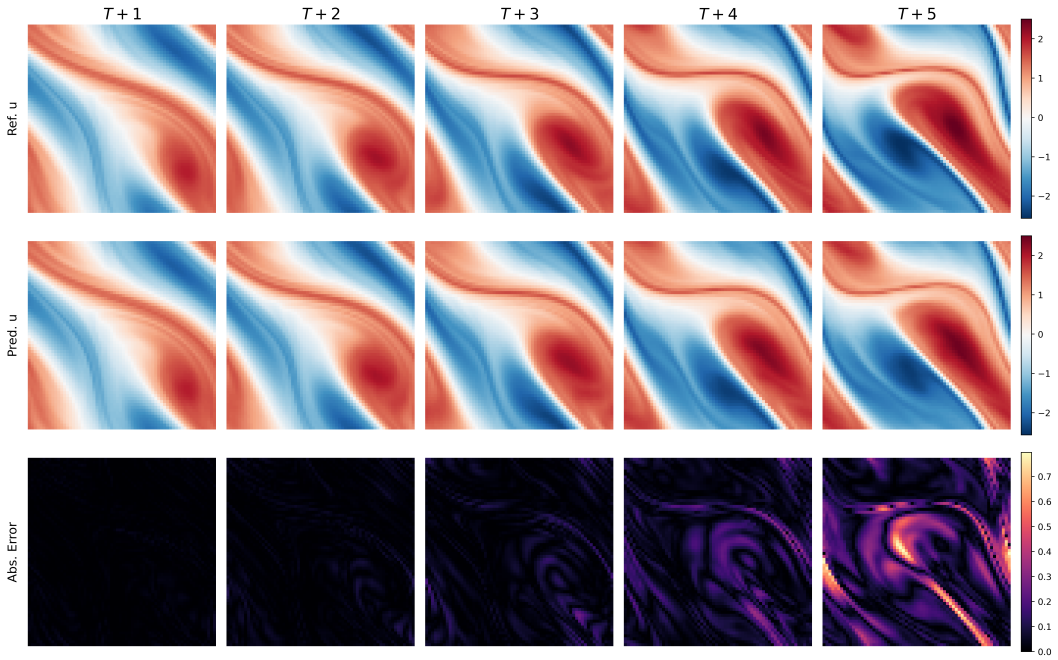

Figure 9: Incompressible Navier-Stokes (64×64) Autoregressive Forecasting.

This visualization demonstrates how the framework dynamically balances information extraction across different spatial scales.

D.3 Visualization of Out-of-Distribution Autoregression

We attempted to have the trained next-step prediction model generate continuous output; the results are shown in Figures 16, 17, and 18.

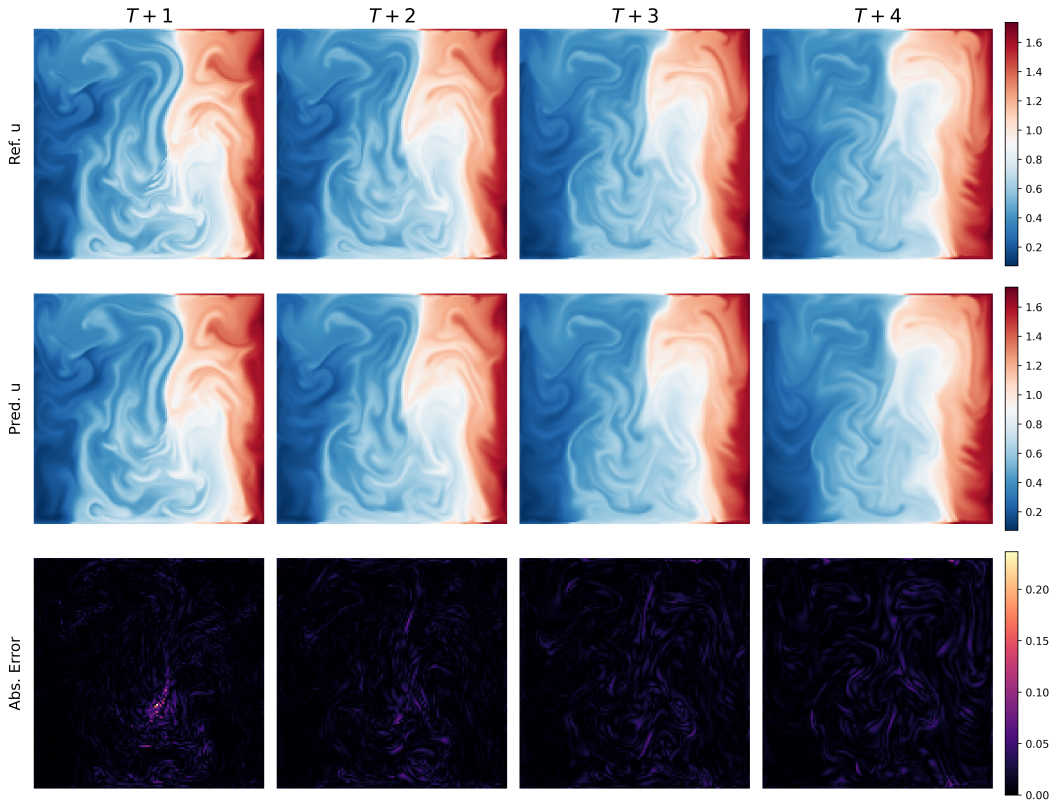

Figure 10: Incompressible Navier-Stokes (128×128) Autoregressive Forecasting.

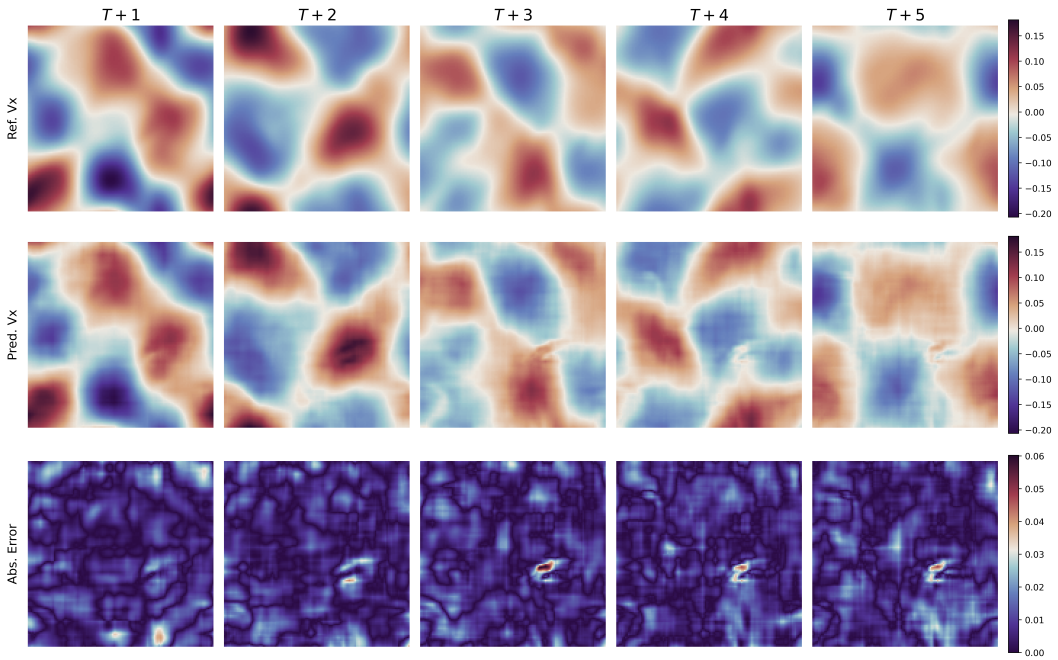

Figure 11: Compressible Navier-Stokes (128×128) Autoregressive Forecasting.

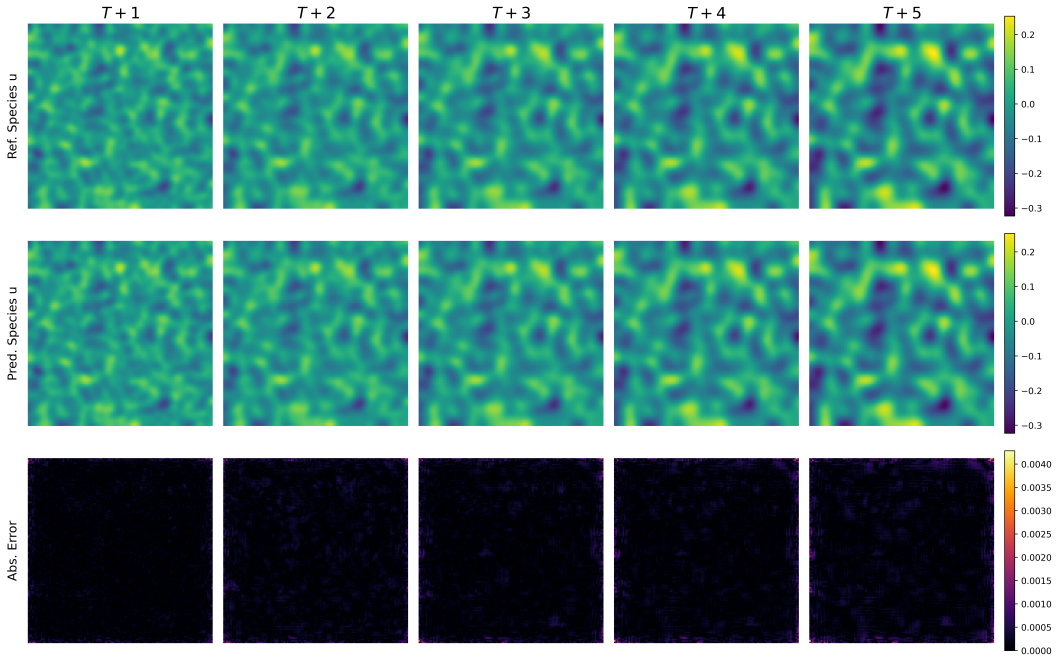

Figure 12: Diffusion–Reaction Autoregressive Forecasting.

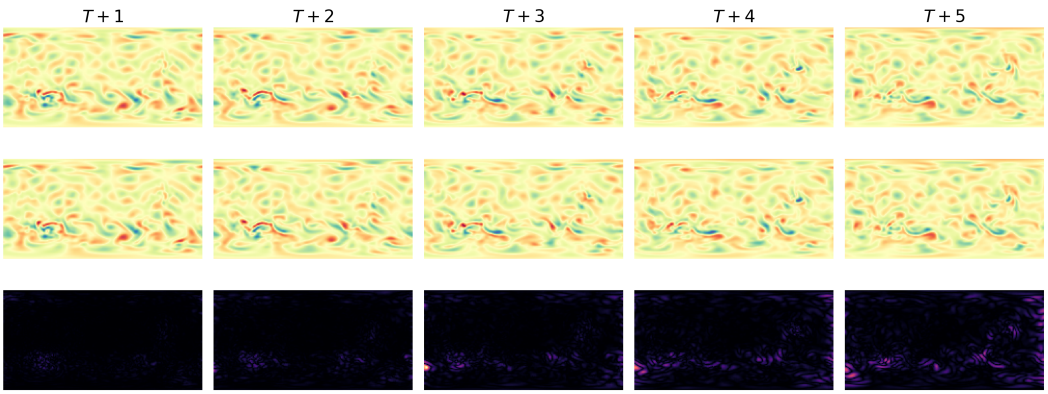

Figure 13: Shallow Water Equation Autoregressive Forecasting.

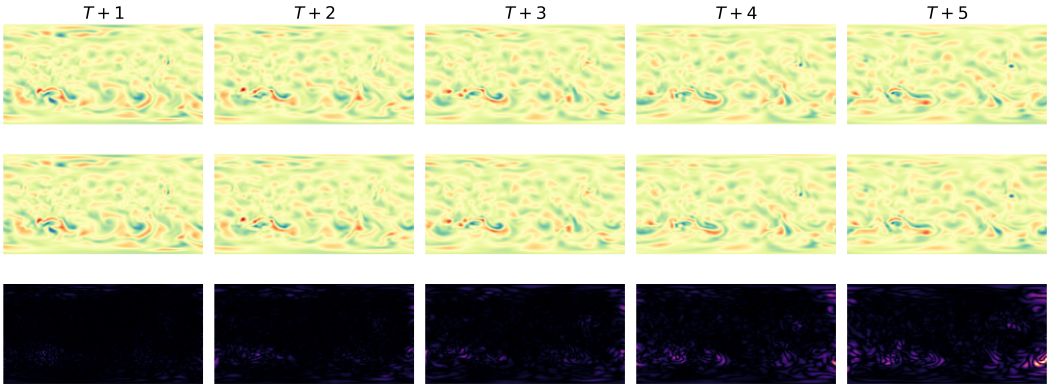

Figure 14: Shallow Water Equation Autoregressive Forecasting.

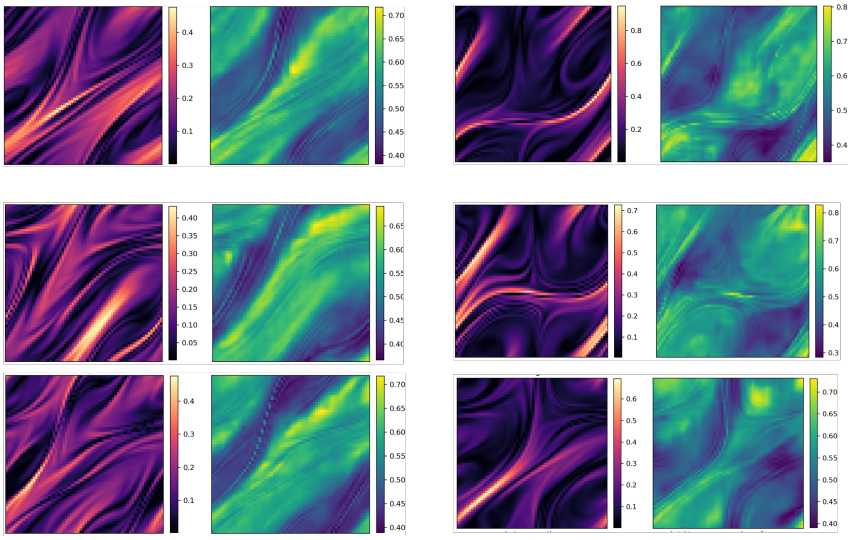

Figure 15: **Multi-scale attention allocation.** Visualization of the attention weights distributed across the $2(L+1)$ spatially aligned tokens for a specific query point. The layout illustrates the explicit decoupling and integration of fine-to-coarse and coarse-to-fine spatial context.

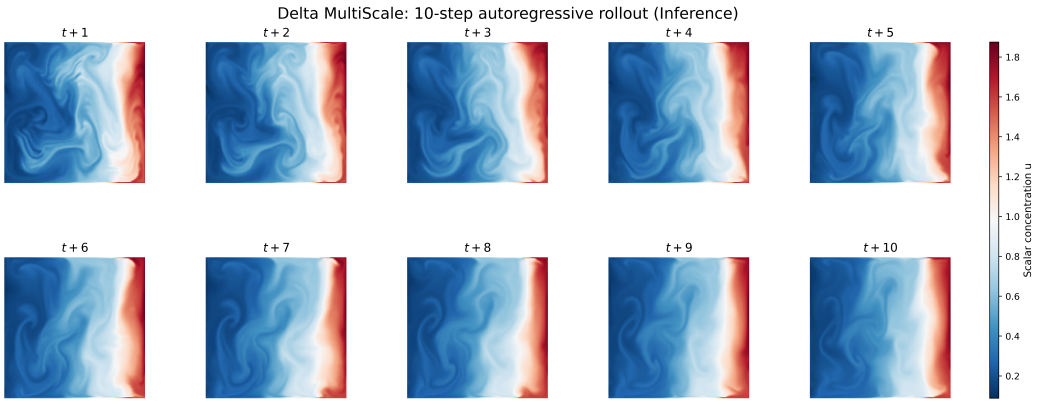

Figure 16: Autoregressive 10-step.

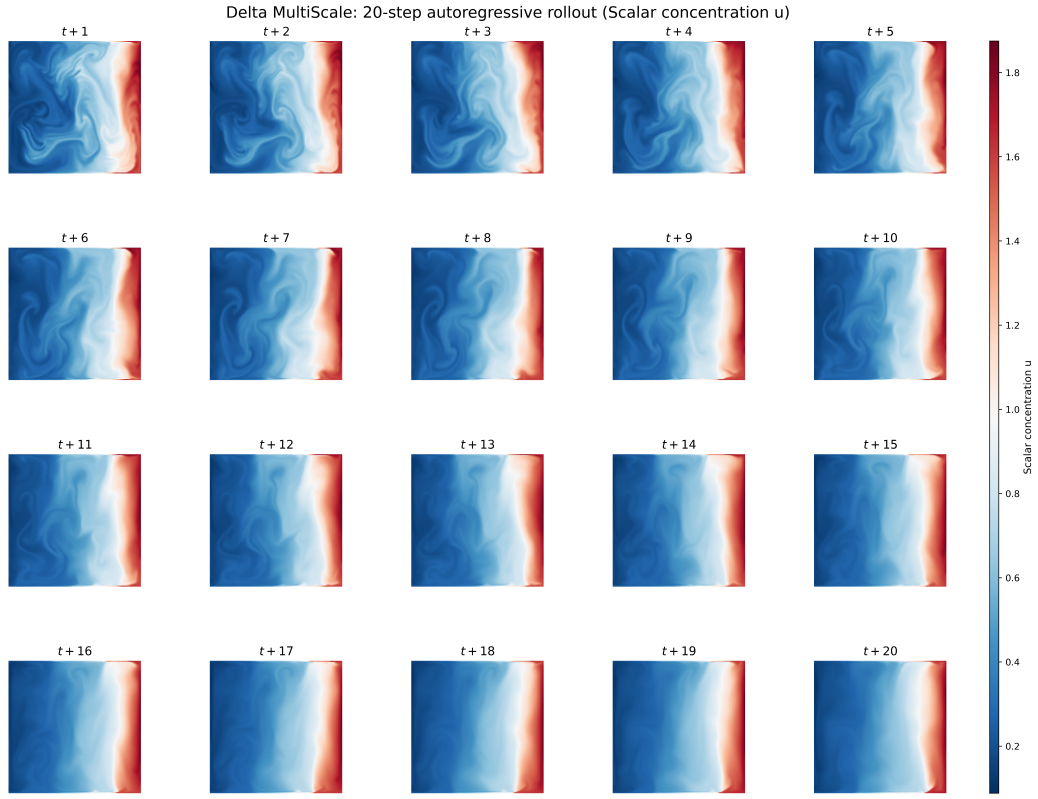

Figure 17: Autoregressive 20-step.

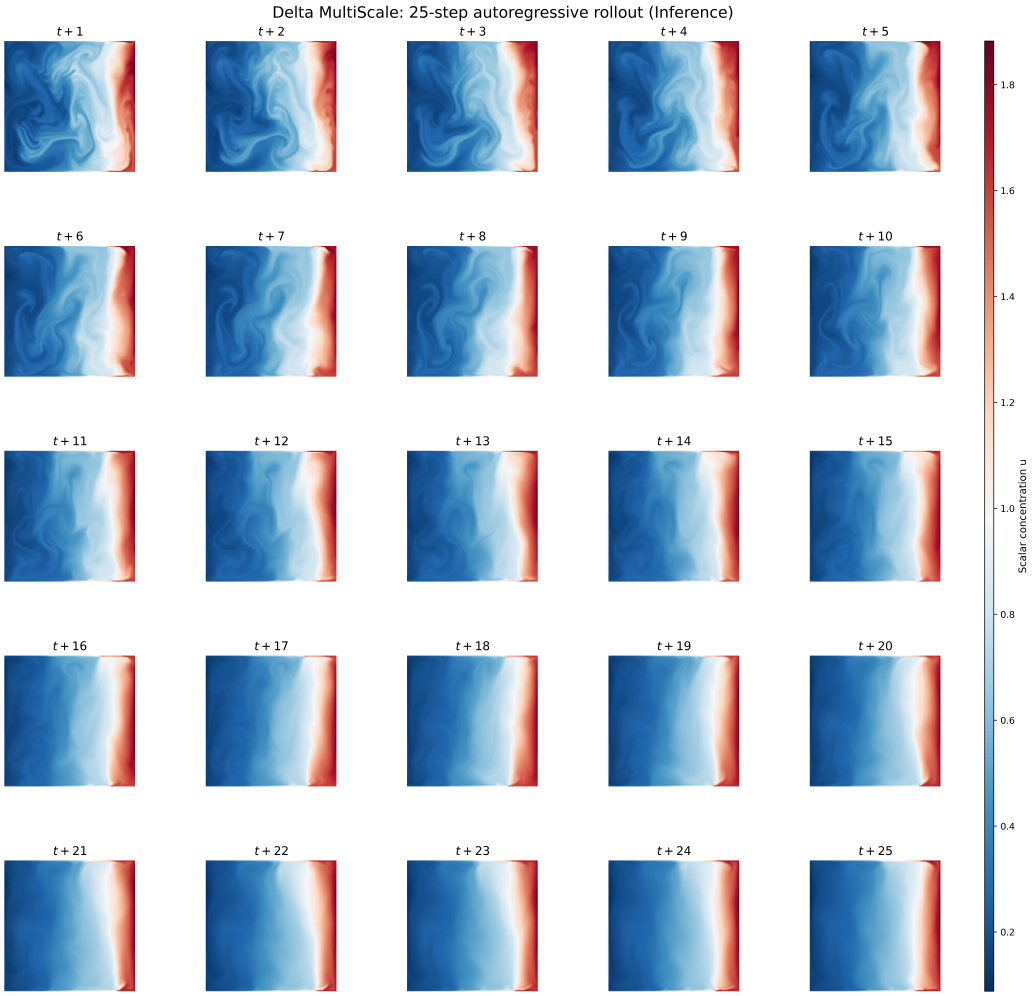

Figure 18: Autoregressive 25-step.