Generative Refinement Learning for Temporal Interpolation in Scientific Machine Learning

Anonymous authors
Paper under double-blind review

Abstract

Scientific temporal interpolation aims to recover intermediate physical fields from sparsely sampled simulations or observations. Existing video frame interpolation methods often rely on transport-based or deterministic-regression biases, which can be misaligned with complex systems exhibiting nonlinear, multi-scale, and uncertain dynamics. We propose a novel **i**nterpolative **FL**ow **EX**pert (`iFLEX`), a generative framework that decomposes interpolation into numerical coarse interpolation and diffusion-based stochastic refinement learning. The coarse estimate captures endpoint-consistent low-frequency evolution, while the diffusion module learns nonlinear, high-frequency, and uncertain corrections. To condition the stochastic refinement process, `iFLEX` separately encodes high-dimensional physical fields through a shared spatial encoder and low-dimensional scalar variables through FiLM-based modulation. We evaluate `iFLEX` on simulated Navier-Stokes Kraichnan turbulence, real-world Shanghai radar precipitation, and sea surface temperature data, including various challenging out-of-distribution tests. Across these settings, `iFLEX` improves reconstruction accuracy and structural fidelity over deterministic, optical-flow-based, and diffusion-based video interpolation baselines, with especially strong gains for large-step interpolation and turbulent or intermittent dynamics.

1 Introduction

Scientific and engineering applications often rely on large-scale spatiotemporal datasets that describe the evolution of physical systems. These datasets are typically collected from numerical simulations, real-world observation stations, and data assimilation pipelines. Nevertheless, the cost of producing and storing high-frequency scientific data remains substantial, and scientists resort to storing them at some (often coarse) temporal resolution (e.g., ERA5 (Hersbach et al., 2020)). Moreover, scientific data products are often obtained by sampling with heterogeneous temporal sampling rates. For instance, meteorological data may be stored at different intervals ranging from minutes to hours, either due to data generation constraints or since retaining each timestep is expensive. This mismatch in sampling frequencies poses challenges for downstream tasks, such as inverse analysis and real-time decision support, which require high-frequency information. This motivates the need to recover high temporal-resolution physical fields from coarse data while preserving physical structures and reducing computational and storage costs.

A natural question is *how to recover these missing temporal states*. In principle, temporal resolution can be increased through either extrapolation or interpolation. Extrapolation treats the problem as a "forward prediction," estimating missing states by advancing the system from an initial condition. However, this strategy can accumulate errors over time, especially in fast-evolving or nonstationary systems. In contrast, interpolation leverages both the initial and final snapshots as "boundary information." The final state provides an additional constraint on the reconstructed trajectory, reducing the drift commonly observed in extrapolation. Therefore, interpolation offers a more stable and reliable strategy for reconstructing missing temporal segments. Despite this advantage, learning temporally continuous interpolants from discrete scientific data remains challenging. Since models are trained only on sampled snapshots, they may fail to preserve the long-term temporal continuity of the underlying dynamical system. Such continuity is important for

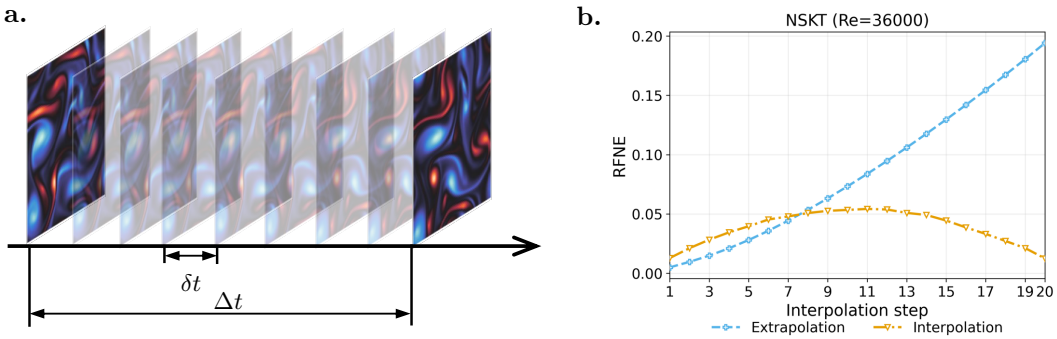


Figure 1: Motivation for boundary-conditioned temporal interpolation. (a) Scientific datasets are often available only at coarse temporal intervals Δt, while downstream tasks may require recovering intermediate states at finer resolution δt. (b) Extrapolation reconstructs missing states by advancing from the initial condition and can accumulate error over time, whereas interpolation uses both endpoint snapshots as boundary constraints. On a turbulence dataset (NSKT) with Reynolds number 36000, interpolation yields substantially lower relative errors (RFNE) across intermediate steps than extrapolation.

accurately evaluating interpolation quality in scientific datasets (Hodgkinson & Mahoney, 2020; Krishnapriyan et al., 2023). We illustrate this motivation in Figure 1.

At one level, this problem resembles video frame interpolation (VFI) (Dong et al., 2023) in computer vision (CV): given two endpoint states, estimate the missing intermediate states. However, scientific temporal interpolation differs fundamentally from classic video frame interpolation from CV. Many VFI methods are built on motion-compensation or optical-flow formulations, which assume that local image appearance is approximately preserved as it is displaced across frames (Horn & Schunck, 1981; Baker et al., 2011). This appearance-preserving transport prior is effective for natural videos with persistent objects of a given "size" or "scale" undergoing apparent motion, but it can be misaligned with scientific data governed by different physical processes such as advection, diffusion, forcing, dissipation, and nonlinear interactions, especially when multiple scales are involved. In such systems, physical structures may emerge, dissipate, interact, or transfer energy across scales, rather than merely undergo appearance-preserving spatial displacement. This distinction is critical since (CV and other) model architectures encode *implicit biases*. For example, flow-based methods in VFI bias interpolation toward spatial transport, while direct regression models bias predictions toward smooth deterministic mappings (Kye et al., 2025). Both assumptions can be misaligned with realistic physical systems, where fine-scale structures and uncertainty are often central to the dynamics. An illustrative example is presented in Figure 2. This motivates the central question of our work:

What inductive biases are appropriate for temporal interpolation of scientific spatiotemporal fields?

In this paper, we propose **i**nterpolative **FL**ow **EX**pert (`iFLEX`), a conditional diffusion framework designed specifically for scientific spatiotemporal interpolation. `iFLEX` aims to perform continuous temporal interpolation by combining numerical interpolation with generative refinement. Rather than use optical-flow priors (Jiang et al., 2018; Xu et al., 2019; Liu et al., 2020), which are common in VFI, we use numerical interpolation schemes to estimate the intermediate states between two available snapshots. This choice is motivated by the fact that optical-flow methods rely on assumptions such as color constancy and pixel-wise advection (Kye et al., 2025), but they do not hold for scientific variables like velocity, pressure, and temperature. In contrast, numerical interpolation can provide a stable approximation of the continuous evolution of physical systems. Furthermore, diffusion models (Ho et al., 2020; Song et al., 2021) are used to learn the numerical corrections after the interpolation due to their powerful capacity in learning high-frequency structures and stochastic components. By integrating coarse physical interpolation into probabilistic refinement, `iFLEX` reduces the learning difficulty and better captures the multi-scale structure present in scientific spatiotemporal data.

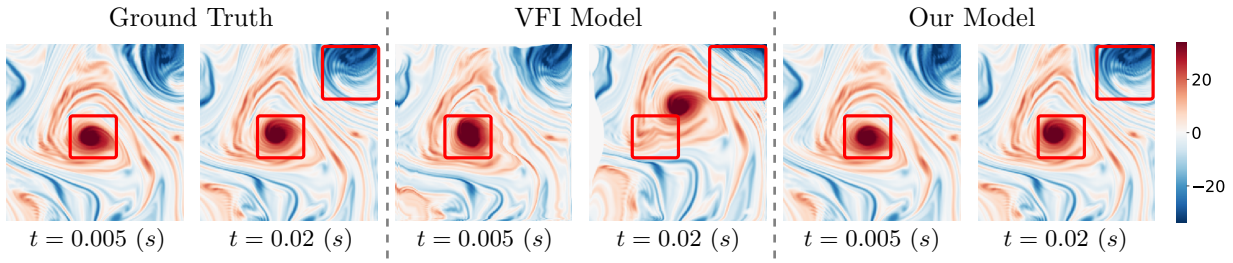


Figure 2: An illustrative comparison between an optical flow-based VFI model and our proposed probabilistic approach for continuous temporal interpolation on simulated turbulence data. Given the same endpoint states, conventional VFI models exhibit temporal inconsistency and distort coherent flow structures, reflecting the mismatch between video-oriented interpolation biases and physical dynamics. In contrast, our `iFLEX` produces intermediate fields that better preserve the evolution of vortical structures and fine-scale flow patterns, closely matching the ground truth across interpolation times.

We evaluate `iFLEX` on controlled simulations and real-world datasets, including Navier–Stokes turbulence, precipitation, and sea surface temperature. Our results show that `iFLEX` captures fine-scale structures and remains robust under challenging interpolation and distribution-shift settings. More broadly, our study highlights that scientific interpolation requires rethinking the implicit assumptions inherited from video models, and that aligning architectural bias with physical data characteristics is essential for reliable spatiotemporal modeling. This is particularly important in light of recent work (Krishnapriyan et al., 2021; 2023; Sakarvadia et al., 2025; Yu et al., 2025).

Overall, our contributions are summarized below:

- We propose a `iFLEX`, a *generative refinement learning* framework for scientific temporal interpolation. `iFLEX` first uses numerical interpolation to construct an endpoint-consistent coarse estimate, and then it applies a diffusion-based stochastic refinement module to learn nonlinear, high-frequency, and uncertain corrections.
- We introduce a *structured conditioning mechanism* tailored to scientific spatiotemporal fields. High-dimensional physical states are encoded through a shared spatial encoder and injected via gated feature modulation, while low-dimensional temporal and physical variables, including interpolation ratio, total interpolation interval, diffusion step, and available physical parameters, are incorporated through FiLM-based modulation (Perez et al., 2018).
- We study scientific temporal interpolation through the lens of *implicit architectural bias*. Our analysis contrasts video-oriented interpolation priors, such as appearance-preserving optical-flow transport and generic latent diffusion, with numerical interpolation, stochastic refinement mechanisms, and structured conditioning mechanisms that better align physical dynamics involving nonlinear interaction, dissipation, and multi-scale structure.
- We evaluate `iFLEX` across controlled simulations and real-world datasets, including turbulence flows, Shanghai radar precipitation, and sea surface temperature. Extensive experiments have shown consistent improvements of our method over deterministic, optical-flow-based, and diffusion-based video interpolation baselines in terms of reconstruction accuracy, structural fidelity, and physics preservation.

2 Related Work

Scientific Data Refinement. Scientific data refinement has become a central topic of research, attracting growing attention across scientific and engineering domains. Current efforts focus on several interrelated directions. First, machine learning (ML) methods have been applied to the full-field spatial reconstruction from sparse sensor measurements (Erichson et al., 2020; Fukami et al., 2021b), with extensions to spatiotemporal

forecasting that jointly recover and predict dynamic fields (Kutz et al., 2024; Song et al., 2024; Lyu et al., 2025). Second, researchers have developed neural PDE solvers that learn from low-resolution simulation data to generate either high-resolution flow predictions (Fukami et al., 2019; 2021a; Esmaeilzadeh et al., 2020; Ren et al., 2023) or numerically accurate coarse-grid solutions (Kochkov et al., 2021; Sun et al., 2023) by correcting discretization and resolution errors. Third, given the storage and computational cost of high-fidelity simulations, recent work has explored data compression and refinement through temporal interpolation (Gadirov et al., 2025a;b) and spatiotemporal downscaling (Jiang et al., 2023; Bassetti et al., 2024), enabling efficient reconstruction of fine-scale physical processes from compressed representations.

Generative Learning Dynamics. Generative models have recently advanced spatiotemporal interpolation (Bassetti et al., 2024; Naiman et al., 2024; Ren et al., 2026), prediction (Rühling Cachay et al., 2023; Li et al., 2024; Erichson et al., 2025), and super-resolution (Du et al., 2024; Lopez-Gomez et al., 2025) across various scientific domains. These models learn the probability distribution of complex dynamic fields, allowing both unconditional generation from random noise and conditional synthesis from coarse-grained inputs or related physical variables. Denoising Diffusion Probabilistic Models (DDPMs) (Ho et al., 2020) have gained attention for their stable training and ability to generate high-fidelity samples. For instance, Dyffusion (Rühling Cachay et al., 2023) and its spherical extension (Rühling Cachay et al., 2024) demonstrate long-range forecasting of fluid dynamics and global climate data. Moreover, diffusion models can be combined with a conditional neural field for diverse spatiotemporal data restoration tasks (Du et al., 2024), such as zero-shot flow reconstruction from sparse sensor measurements and super-resolution generation. Beyond simulation and reanalysis datasets, diffusion methods also show strong performance in radar precipitation forecasting (Yu et al., 2024) by capturing both statistical variability and physical realism. Despite these advances, the application of generative models to spatiotemporal interpolation problems in scientific machine learning (SciML) remains largely unexplored.

Video Frame Interpolation. VFI is a well-established problem in CV that aims to synthesize intermediate frames between two given frames. Conventional video frame interpolation approaches primarily rely on optical flow to model motion between frames, such as linear (Jiang et al., 2018) and quadratic (Xu et al., 2019; Liu et al., 2020) motion estimation. These VFI methods are deterministic and effective under moderate motion, but they often struggle with large displacements and rapid appearance changes (Kye et al., 2025). To mitigate these issues, generative models have been introduced to capture uncertainty and produce diverse yet physically consistent interpolations. Among them, diffusion models have recently emerged as a powerful framework that reformulates VFI as a conditional generative denoising process. The typical diffusion-based methods operate directly in pixel space, iteratively refining the interpolated frame conditioned on the start and end frames (Jain et al., 2024), while others perform diffusion in a latent space to improve computational efficiency and implicitly capture motion dynamics (Danier et al., 2024). Further advancements incorporate explicit motion guidance using flow-based methods for enhanced temporal coherence (Huang et al., 2024), or apply diffusion to residual refinements that correct pre-interpolated frames (Chavez et al., 2025). These generative diffusion frameworks offer improved stability and physical consistency compared with traditional flow-based interpolators.

3 Temporal Interpolation as a Boundary Value Problem

Problem Setup. We study conditional temporal interpolation for scientific spatiotemporal fields. Let $\mathbf{u}(t, \mathbf{x})$ denote the physical field of interest at time t and spatial location $\mathbf{x}$, evolving over a spatial grid $\Omega \in \mathbb{R}^{H \times W}$. The underlying dynamics are governed by an unknown or partially known nonlinear operator: $\frac{\partial \mathbf{u}}{\partial t} = \mathcal{F}(\mathbf{u}; \boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ represents system parameters such as Reynolds numbers. Given two observed endpoint snapshots $\mathbf{u}(t_0) \in \mathbb{R}^{H \times W}$ and $\mathbf{u}(t_1) \in \mathbb{R}^{H \times W}$ with $t_0 < t_1$, our goal is to recover an intermediate state $\widehat{\mathbf{u}}_\tau \in \mathbb{R}^{H \times W}$, where $t_\tau = (1 - \tau)t_0 + \tau t_1$ and $\tau \in (0, 1)$. Unlike forecasting, which is typically formulated as an initial-value problem, temporal interpolation uses both endpoints and can be viewed as a boundary-value problem:

$$\frac{\partial \mathbf{u}}{\partial t} = \mathcal{F}(\mathbf{u}), \quad \mathbf{u}(t_0) = \mathbf{u}_0, \quad \mathbf{u}(t_1) = \mathbf{u}_1. \tag{1}$$

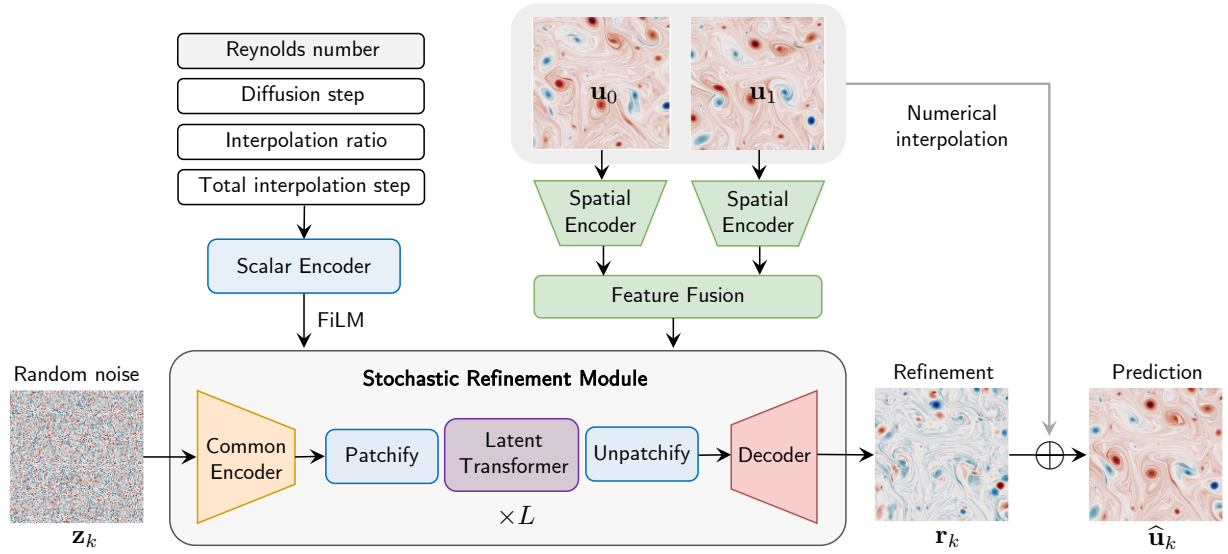


Figure 3: Overview of `iFLEX`, which decomposes temporal interpolation into a boundary-consistent low-frequency estimate and diffusion-based high-frequency refinement. Numerical interpolation provides a physically consistent coarse estimate of the temporal variation, while the stochastic refinement module learns residual corrections that recover fine-scale structures. A shared spatial encoder maps the endpoint states ($\mathbf{u}_0$ and $\mathbf{u}_1$) into a latent feature space, and the resulting spatial features are injected into the low-resolution layers of the refinement module. Meanwhile, a scalar encoder embeds accessible physical variables such as Reynolds number, together with time-specific parameters, including the interpolation ratio, total interpolation interval, and diffusion timestep. These scalar features are incorporated into each layer of the encoder-transformer-decoder denoising network through FiLM-based modulation (Perez et al., 2018).

The two boundary snapshots constrain the admissible solution trajectories to those that connect $\mathbf{u}_0$ and $\mathbf{u}_1$. However, for complex physical processes, the intermediate dynamics remain *underdetermined*, especially in the presence of *large* temporal gaps, *noisy* observations, and *multi-scale* behavior. To this end, we formulate the learning objective as estimating the conditional distribution of intermediate fields,

$$p(\mathbf{u}_\tau | \mathbf{u}_0, \mathbf{u}_1, \boldsymbol{\lambda}, \mathbf{p}), \tag{2}$$

instead of a deterministic mapping. Here, $\mathbf{p}$ denotes all other related parameters, including interpolation ratio τ, total interpolation interval, and diffusion timesteps.

4 Method

In this work, we propose to decompose the interpolation problem into a coarse boundary-consistent estimate and a finer probabilistic refinement. The numerical interpolation component encodes a stable, boundary-aware prior, while the diffusion refinement learns the nontrivial high-frequency residual dynamics needed to recover realistic intermediate states. Our setup is designed to infer physically plausible intermediate fields under limited temporal observations. The overview of our framework is presented in Figure 3. Specifically, our end-to-end framework comprises three main components: (i) deterministic boundary-consistent coarse interpolation, (ii) stochastic refinement in the physical space, and (iii) structured continuous-time conditioning.

4.1 Boundary-Aware Coarse Interpolation

Given two endpoint snapshots $\mathbf{u}_0$ and $\mathbf{u}_1$, our first step is to construct a coarse estimate of the intermediate state at $t_\tau = (1-\tau)t_0 + \tau t_1$, where $\tau \in (0, 1)$. This estimate serves as a boundary-consistent reference trajectory that satisfies the endpoint constraints and captures the lowest-order temporal variation between the two observations. Formally, we define a deterministic interpolation operator $\mathcal{I}(\cdot)$: $\widetilde{\mathbf{u}}_\tau = \mathcal{I}(\mathbf{u}_0, \mathbf{u}_1; \tau)$, where

$\widetilde{\mathbf{u}}_\tau$ denotes the coarse prediction at interpolation ratio τ. A simple and stable choice is linear interpolation in the physical space:

$$\widetilde{\mathbf{u}}_\tau = (1-\tau)\mathbf{u}_0 + \tau\mathbf{u}_1, \tag{3}$$

which satisfies the boundary conditions: $\mathcal{I}(\mathbf{u}_0, \mathbf{u}_1; 0) = \mathbf{u}_0$ and $\mathcal{I}(\mathbf{u}_0, \mathbf{u}_1; 1) = \mathbf{u}_1$. Although this linear estimate is not sufficient to recover nonlinear or multi-scale dynamics, it provides a physically meaningful low-order approximation that is robust, computationally inexpensive, and free from video-specific assumptions such as optical-flow warping.

This design has two advantages. First, it explicitly preserves the boundary structure of the interpolation problem, ensuring that the model is anchored by both endpoints $\mathbf{u}_0$ and $\mathbf{u}_1$. Second, it reduces the burden on the stochastic refinement model by separating low-frequency temporal variation from high-frequency nonlinear corrections. As a result, the subsequent diffusion model can focus on learning the uncertain and physically complex components of the dynamics, rather than rediscovering the coarse interpolation between endpoints.

4.2 Stochastic Refinement Learning

The deterministic interpolation module provides a coarse global temporal trend $\widetilde{\mathbf{u}}_\tau$, but it also introduces a "numerical" error gap. To address it, we present a stochastic refinement module that learns the residual evolution between the interpolated field $\widetilde{\mathbf{u}}_\tau$ and the true intermediate state $\mathbf{u}_\tau$. We formulate this process as learning local fine-scale corrections $\mathbf{r}_\tau$, given by

$$\mathbf{r}_\tau = \mathbf{u}_\tau - \widetilde{\mathbf{u}}_\tau. \tag{4}$$

The final interpolated state is then reconstructed as $\widehat{\mathbf{u}}_\tau = \widetilde{\mathbf{u}}_\tau + \widehat{\mathbf{r}}_\tau$, where $\widehat{\mathbf{r}}_\tau$ is sampled from a conditional diffusion model. This residual formulation reduces the burden on the diffusion model: rather than generating the full physical field, the model focuses on learning the nonlinear correction beyond the boundary-consistent interpolation.

We use a denoising diffusion model (Ho et al., 2020; Song et al., 2021) to learn the conditional distribution

$$p_\theta(\mathbf{r}_\tau | \mathbf{u}_0, \mathbf{u}_1, \boldsymbol{\lambda}, \mathbf{p}), \tag{5}$$

where $\boldsymbol{\lambda}$ and $\mathbf{p}$ represent the available physical variables and time-specific parameters, respectively, and where θ denotes the network parameters. In the training stage, the clean residual is gradually corrupted by Gaussian noise, which is given by

$$\mathbf{z}_n = \alpha_n \mathbf{r}_\tau + \sigma_n \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad n = 1, 2, \cdots, N, \tag{6}$$

where n is the diffusion timestep and α_n, σ_n are determined by the noise schedule, and N is the final diffusion timestep, i.e., the maximum-noise state. The denoising network is trained to predict the injected noise from the noisy residual and conditioning information:

$$\widehat{\boldsymbol{\epsilon}} = \epsilon_\theta(\mathbf{z}_n, n | \mathbf{u}_0, \mathbf{u}_1, \boldsymbol{\lambda}, \mathbf{p}). \tag{7}$$

The training objective is formulated as

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{r}_\tau, n, \boldsymbol{\epsilon}}[\|\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{z}_n, n | \mathbf{u}_0, \mathbf{u}_1, \boldsymbol{\lambda}, \mathbf{p})\|_2^2]. \tag{8}$$

At inference time, we first compute the coarse estimate $\widetilde{\mathbf{u}}_\tau$ using the deterministic interpolation operator; and we then initialize the residual variable from Gaussian noise: $\mathbf{z}_N \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and iteratively denoise it using the learned reverse process, conditioned on the endpoint states and available physical variables:

$$\mathbf{z}_{n-1} \sim p_\theta(\mathbf{z}_{n-1} | \mathbf{z}_n, \mathbf{u}_0, \mathbf{u}_1, \boldsymbol{\lambda}, \mathbf{p}). \tag{9}$$

After the final denoising step, we obtain the residual sample $\widehat{\mathbf{r}}_0 = \mathbf{z}_0$, and reconstruct the intermediate field as $\widehat{\mathbf{u}}_\tau = \widetilde{\mathbf{u}}_\tau + \widehat{\mathbf{r}}_\tau$. Overall, the conditional diffusion module serves as a numerically-grounded refinement mechanism, learning the high-frequency and uncertain components to recover the true intermediate trajectory.

This approach is quite general; and, to evaluate it, we adopt the encoder-transformer-decoder architecture of the `FLEX` framework (Erichson et al., 2025) as the backbone of our stochastic refinement module, as `FLEX` is naturally suited to residual learning.

4.3 Structured Continuous-time Conditioning

Conditioning is central to scientific temporal interpolation since the target state is strongly constrained by the endpoint fields, interpolation ratio, and the physical regime. This differs from text-to-image generation (Rombach et al., 2022), where the condition is typically low-dimensional semantic information (e.g., text) and the task is highly uncertain, requiring strong generative capacity. In contrast, interpolative generation is comparatively lower uncertainty, but it requires strong reconstruction ability. Namely, the scientific interpolation model should preserve boundary consistency, recover fine-scale structures, and respect continuous temporal evolution (Krishnapriyan et al., 2023). Moreover, our conditioning inputs are heterogeneous, combining high-dimensional spatial fields $\mathbf{u}_0, \mathbf{u}_1$ with low-dimensional scalar variables such as the interpolation ratio τ, total interval Δt, available physical parameters $\boldsymbol{\lambda}$, and diffusion step n. Therefore, directly adopting conventional conditioning schemes such as dense cross-attention at all layers is unnecessary and computationally expensive. We instead design a *structured conditioning mechanism* that treats different condition types according to their roles: endpoint fields are encoded as spatial features and injected through a learned gate (Dauphin et al., 2017; Hu et al., 2018) to adaptively control boundary information, while scalar variables are incorporated through FiLM modulation (Perez et al., 2018) to globally adapt the denoising dynamics across temporal locations, physical regimes, and noise levels.

As shown in Figure 3, for spatial conditioning, we encode the two endpoint fields using a shared encoder E_ψ. At resolution level l, this gives

$$\mathbf{f}_0^{(\ell)} = E_\psi^{(\ell)}(\mathbf{u}_0), \qquad \mathbf{f}_1^{(\ell)} = E_\psi^{(\ell)}(\mathbf{u}_1), \tag{10}$$

where $\mathbf{f}_0^{(\ell)}, \mathbf{f}_1^{(\ell)} \in \mathbb{R}^{H_\ell \times W_\ell \times C_\ell}$. The shared encoder ensures that the initial and final states are represented in the same latent feature space. To capture the boundary structure and temporal change between the two endpoints, we construct a fused spatial condition map (i.e., Feature Fusion part in Figure 3)

$$\mathbf{c}_{\text{sp}}^{(\ell)} = \psi_\ell \left(\left[\mathbf{f}_0^{(\ell)}, \mathbf{f}_1^{(\ell)}, \frac{1}{2}\left(\mathbf{f}_0^{(\ell)} + \mathbf{f}_1^{(\ell)} \right), \mathbf{f}_1^{(\ell)} - \mathbf{f}_0^{(\ell)} \right] \right), \tag{11}$$

where $[\cdot]$ denotes channel-wise concatenation and ψ_ℓ is a lightweight projection module. The average feature provides a stable representation of shared structures, while the difference feature encodes the endpoint-to-endpoint change. These spatial condition maps are injected into the low-resolution layers of the denoising network, where they guide the global structure of the residual without incurring the cost of dense high-resolution conditioning.

For continuous-time and physics conditioning, we collect the scalar variables into

$$\mathbf{s} = [\boldsymbol{\lambda}, \tau, \Delta t, n]. \tag{12}$$

These variables provide global context rather than spatial patterns, so we encode them as channel-wise scale and shift parameters that modulate the denoising features via the FiLM scheme. This allows the same denoising backbone to adapt to different time intervals, physical regimes, and noise levels without concatenating scalar quantities as image channels. First, we embed them using a Fourier feature map followed by a multilayer perceptron:

$$\mathbf{e}_{\text{sc}} = \text{MLP}\big(\Phi(\mathbf{s})\big), \tag{13}$$

where $\Phi(\cdot)$ denotes the Fourier positional embedding. At each denoising block ℓ, this scalar embedding generates feature-wise scale and shift parameters,

$$\boldsymbol{\gamma}^{(\ell)}, \boldsymbol{\beta}^{(\ell)} = W_\ell(\mathbf{e}_{\text{sc}}), \tag{14}$$

which modulate the hidden feature map $\mathbf{h}^{(\ell)}$ through FiLM:

$$\widetilde{\mathbf{h}}^{(\ell)} = \left(1 + \boldsymbol{\gamma}^{(\ell)}\right) \odot \mathbf{h}^{(\ell)} + \boldsymbol{\beta}^{(\ell)}. \tag{15}$$

Here, $\boldsymbol{\gamma}^{(\ell)}$ and $\boldsymbol{\beta}^{(\ell)}$ are broadcast across spatial dimensions, enabling global temporal and physical variables to control the channel responses of the denoising network.

The spatial and scalar conditioning signals are then combined inside the stochastic refinement module. For layers where spatial conditioning is applied, we update the modulated feature by

$$\mathbf{h}_{\text{cond}}^{(\ell)} = \widetilde{\mathbf{h}}^{(\ell)} + \mathbf{g}^{(\ell)} \odot W_c^{(\ell)} \left(\mathbf{c}_{\text{sp}}^{(\ell)} \right), \tag{16}$$

where $W_c^{(\ell)}$ aligns the condition map with the hidden feature dimension and

$$\mathbf{g}^{(\ell)} = \sigma \left(W_g^{(\ell)} \left(\mathbf{c}_{\text{sp}}^{(\ell)} \right) \right) \tag{17}$$

is a learned gate. This gated injection allows the model to adaptively determine where and how strongly endpoint information should influence the denoising process.

Overall, the conditioning mechanism can be summarized as

$$\boldsymbol{\epsilon}_\theta = \mathcal{D}_\theta \left(\mathbf{z}_n, n; \mathcal{C}_\psi \left(\mathbf{u}_0, \mathbf{u}_1, \tau, \Delta t, \boldsymbol{\lambda} \right) \right), \tag{18}$$

where $\mathcal{D}_\theta$ denotes the denoising backbone and $\mathcal{C}_\psi$ denotes the structured conditioning network. This design preserves the separation between denoising and conditioning: spatial fields guide the residual through learned low-resolution feature maps, while continuous-time and physical variables globally modulate the denoising dynamics. As a result, the model incorporates scientific context without collapsing into a direct deterministic mapping from endpoints to the target state.

4.4 Training and Inference

Multi-resolution Training. While continuous-time embeddings provide flexibility for temporal interpolation, they can introduce temporal blurring when the supervision signal is weak (Kye et al., 2025), and there can be non-trivial subtleties, depending on the architecture (Yu et al., 2025). To mitigate this issue, we incorporate numerical interpolation as a prior, and we adopt a multi-resolution training strategy. Specifically, during training, we vary the total number of interpolation steps, T, which allows the model to be supervised at diverse intermediate time stamps. This strategy helps expose `iFLEX` to interpolation targets at different time scales and strengthens its ability to resolve time-dependent structures.

Arbitrary Inference. The proposed framework supports arbitrary-time inference by explicitly conditioning on the target time stamp t_τ. This allows the model to generate system snapshots at any intermediate time $t_k \in (t_0, t_1)$ without relying on fixed temporal resolution. Moreover, dense temporal sequences can be reconstructed from sparsely sampled data with flexible resolution and improved efficiency within a single forward pass.

5 Experiments

In this section, we evaluate `iFLEX` across a range of spatiotemporal systems to assess both accuracy and generalization under increasing dynamical complexity. We combine controlled simulations with real-world datasets to provide a comprehensive and realistic assessment.

We first consider Navier–Stokes Kraichnan Turbulence (NSKT) data, where we construct an easy-to-hard protocol by varying the Reynolds number Re. Models are trained on mid-range regimes (e.g., Re={2000,4000,8000,16000,32000}) and evaluated on unseen values (e.g., $Re = 36000$), forming a principled out-of-distribution (OOD) generalization test. This setup isolates how model inductive biases affect performance as the dynamics transition from smooth to highly turbulent regimes. We further evaluate on real-world precipitation and sea surface temperature datasets, which introduce additional challenges such as noise and non-stationarity. Together, these experiments probe model robustness across regimes of increasing difficulty and distribution shift.

5.1 Datasets

NSKT. Fluid flows play a central role across science and engineering, from environmental and material sciences to geophysics, astrophysics, and chemical engineering. Turbulence is a chaotic behavior inherent

to many flows, posing major challenges for modeling and prediction. The governing equations for fluid turbulence can be described using the incompressible NS equations:

$$\nabla \cdot \mathbf{u} = 0, \quad \nabla \times \mathbf{u} = \mathbf{w}, \quad \frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} = -\frac{1}{\rho} \nabla \mathbf{p} + \nu \nabla^2 \mathbf{u}, \tag{19}$$

where $\mathbf{u}$ and $\mathbf{p}$ denote the velocity and pressure fields, while ρ and ν represent the fluid density and kinematic viscosity, respectively. Here, we consider the Kraichnan model, which offers a simplified framework for analyzing turbulent and chaotic dynamics in fluid flows. The NSKT dataset in this paper is built as an extension of SuperBench (Ren et al., 2025). Training data are generated at Reynolds numbers $Re = \{2000, 4000, 8000, 16000, 32000\}$, while testing data are produced from a distinct initial condition with $Re = \{600, 1000, 2000, 4000, 8000, 12000, 16000, 24000, 32000, 36000\}$. The simulations have an original spatial resolution of 2048^2, which is cropped to 256^2 for model training and evaluations. The interpolation step is defined as 20. We use the vorticity field as a representative example for analyzing the model performance.

Sea Temperature Data. We also consider the ocean temperature data used in HelmFluid (Xing et al., 2023), which is originally adapted from E.U. Copernicus Marine Service Information (the ECMWF reanalysis dataset) (E.U. Copernicus Marine Service Information (CMEMS). Marine Data Store (MDS), 2023). We follow the region selection procedure in (Xing et al., 2023), but we use a spatial resolution of 128×128. The time interval is one day for this dataset. We set the interpolation step to 10.

Shanghai Radar Data. The Shanghai Radar dataset (Chen et al., 2020) was collected from volume scans at approximately 6-minute intervals between 2015 and 2018 in Shanghai, China. We use the dataset following the preprocessing procedure in (Yu et al., 2024). The data cover a spatial domain of 501 km $\times$ 501 km, discretized on a 128×128 grid. The interpolation step is set to 8 due to the significant difficulty of this real-world dataset.

5.2 Implementations

Evaluation Metrics. To rigorously test the performance of our proposed `iFLEX` model, we employ a multi-faceted assessment strategy (Ren et al., 2025), including both data-centric and physics-based evaluations. The data-centric metrics measure the pixel-wise difference between the spatiotemporally reconstructed results and the ground truth. Specifically, we consider relative Frobenius norm errors (RFNE), Pearson correlation (PC), and the Structural Similarity Index Measure (SSIM) (Wang et al., 2004). Note that SSIM is evaluated in the normalized space due to the varying channel ranges across different types of scientific data, which ensures consistency and comparability in the assessment. Moreover, since our objective focuses on scientific data, it is also crucial to assess the preservation of domain-specific physical properties, such as energy spectra. The details of these metrics can be found in Appendix A.2.

Baseline Models. We compare our approach against three representative baselines: a U-Net model (Ronneberger et al., 2015) as a standard SciML architecture, Super SloMo (Jiang et al., 2018) as a deterministic VFI method, and EDEN (Zhang et al., 2025) as a state-of-the-art diffusion-based interpolation model.

5.3 General Results

In this part, we evaluate the overall interpolation performance of `iFLEX` compared to baseline models across all three benchmark settings.

Simulated NSKT Data. For this controlled simulation dataset, we evaluate the physics generalization by training all models on $Re = \{2000, 4000, 8000, 16000, 32000\}$ and testing them on unseen Reynolds numbers $Re = \{12000, 24000, 36000\}$. Table 1 shows that `iFLEX` consistently outperforms all baselines under Reynolds-number generalization. Across all test regimes, `iFLEX` achieves the lowest RFNE and the highest PC and SSIM, with especially large gains in structural fidelity. For example, at a highly turbulent regime ($Re = 36000$), `iFLEX` reduces RFNE from 24.82% for Super SloMo, 26.46% for EDEN, and 30.15% for U-Net to 3.86%, while improving SSIM from 0.83, 0.65, and 0.45 to 0.96. Similar trends hold at $Re = 12000$ and

Table 1: OOD performance evaluations for NSKT data across different Reynolds numbers. All the models are trained with datasets using $Re = \{2000, 4000, 8000, 16000, 32000\}$. The total interpolation step is 20 for all models. The best results are marked in bold. "IT" denotes the inference time in milliseconds, which is computed based on the time required to interpolate one frame.

			$Re = 12000$			$Re = 24000$			$Re = 36000$		
Method	#Params [M]	IT [ms]	RFNE↓	PC↑	SSIM↑	RFNE↓	PC↑	SSIM↑	RFNE↓	PC↑	SSIM↑
U-Net	260.33	29.41	34.07%	0.93	0.45	41.81%	0.90	0.36	30.15%	0.94	0.55
EDEN	194.00	9.53	29.80%	0.96	0.65	35.40%	0.94	0.53	26.46%	0.96	0.70
Super SloMo	39.59	3.54	14.93%	0.99	0.83	20.83%	0.98	0.76	24.82%	0.97	0.71
`iFLEX`	110.4	35.39	**3.79%**	**1.00**	**0.96**	**6.81%**	**1.00**	**0.92**	**3.86%**	**1.00**	**0.96**

$Re = 24000$, indicating that the proposed model remains robust when evaluated at unseen physical regimes. These results suggest that `iFLEX` benefits from an architectural bias better aligned with scientific temporal interpolation, particularly in long-range continuous interpolation settings. The deterministic encoder–decoder models, such as U-Net, struggle to interpolate long-range dynamics and extrapolate across Reynolds numbers because they learn a direct mapping from the endpoint states to the target snapshot, which can be sensitive to changes in the underlying flow regime. Super SloMo, a deterministic VFI method, performs better than U-Net, but its optical-flow-based transport bias constrains its long-horizon interpolation performance for fluid data where turbulent structures deform, dissipate, and interact across scales. Although EDEN is a more expressive diffusion-based model, its performance is limited here by compressed latent reconstruction and VAE decoder block artifacts, which directly affect evaluation metrics on high-frequency vorticity fields. Overall, `iFLEX` presents lower errors than all baselines, while preserving coherent vortex structures across both interpolation time and Reynolds-number shifts. Additional comparative snapshots can be found in Figure 8 of Appendix B.

Shanghai Radar Data. Table 2 reports performance on two real-world datasets that pose distinct interpolation challenges. The Shanghai radar dataset is particularly difficult because precipitation data are highly intermittent, noisy, and locally non-stationary: rain cells can intensify, dissipate, split, or merge within short time intervals. On this dataset, `iFLEX` achieves the best performance across all metrics. It also obtains the highest PC and SSIM, indicating better reconstruction accuracy and stronger preservation of spatial precipitation structures. However, the RFNE values remain high for all methods due to the sparse and heavy-tailed feature of precipitation fields, where small spatial misalignments or errors in high-intensity regions can produce large relative deviations. The improvement over Super SloMo suggests that optical-flow-based transport is insufficient for radar precipitation interpolation, where changes are not purely motion-driven. The gain over EDEN, a strong diffusion-based video interpolation model, suggests that directly adopting a video-oriented latent generative formulation may introduce a misaligned implicit design bias for scientific fields. Our method that combines boundary-aware numerical interpolation with stochastic refinement learning provides a more suitable inductive bias for these dynamics. Additional visualization of comparative snapshots is shown in Figure 9 of Appendix B.

SST Data. The SST dataset presents a different regime: sea surface temperature evolves more smoothly and has stronger large-scale spatial coherence, but accurate interpolation still requires preserving subtle gradients and low-amplitude temporal variations. Furthermore, we evaluate the cross-region domain shift problem in this dataset. All models are trained on data from the Atlantic South, Indian Ocean, and North Pacific regions, and tested on the held-out South Pacific region. The test set differs geographically from the training distribution, requiring models to generalize across regional ocean dynamics. Despite this domain shift, `iFLEX` achieves the best overall performance, with the lowest RFNE (1.81%) and the highest PC (0.998) and SSIM (0.87). The smaller performance gaps compared with the Shanghai radar are consistent with the smoother dynamics of SST, where interpolation is less dominated by abrupt local changes. Nevertheless, `iFLEX` consistently improves over deterministic regression, optical-flow-based interpolation, and the diffusion-based EDEN baseline, showing that our framework remains effective under geophysical domain shift. We present the comparative snapshots in Figure 10 of Appendix B.

Table 2: Performance summary on the Shanghai radar and SST datasets. The total number of interpolation steps is set to 8 for the Shanghai radar and 10 for the SST data across all models. The best results are marked in bold. "IT" denotes inference time in milliseconds, measured as the time required to interpolate a single frame.

		Shanghai Radar Data				Sea Surface Temperature			
Method	#Params [M]	IT [ms]	RFNE↓	PC↑	SSIM↑	IT [ms]	RFNE↓	PC↑	SSIM↑
U-Net	260.33	13.25	48.51%	0.85	0.82	13.27	2.29%	0.996	0.80
EDEN	194.00	2.32	56.54%	0.77	0.77	2.45	2.48%	0.996	0.77
Super SloMo	39.59	1.45	36.77%	0.91	0.87	1.45	2.28%	0.996	0.79
`iFLEX`	110.4	10.53	**32.69%**	**0.93**	**0.88**	9.35	**1.81%**	**0.998**	**0.87**

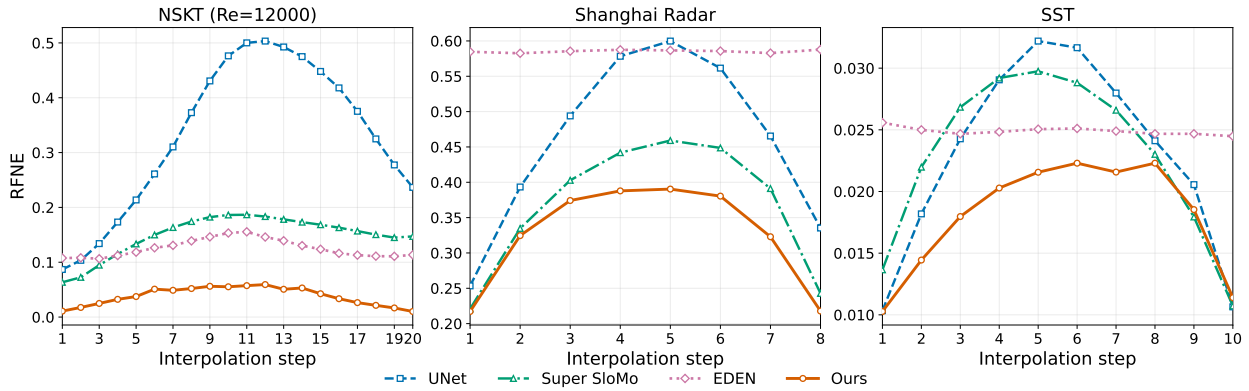


Figure 4: Multi-step interpolation performance across NSKT ($Re = 12000$), Shanghai radar, and SST datasets. We use the RFNE metric at each interpolation step for all models.

5.4 Analysis and Discussions

In this part, we use the NSKT dataset at $Re = 12000$ as a representative case for the analysis. All model predictions are generated under the same $Re = 12000$ evaluation setting. We examine model behavior from four perspectives: multi-step interpolation error distribution, physical structure preservation, accuracy-efficiency analysis, and generalization to unseen interpolation horizons.

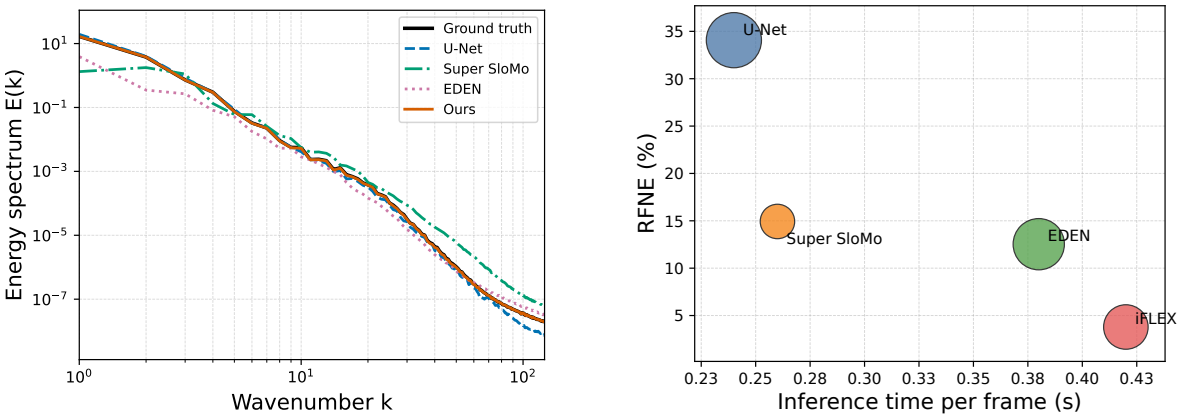
Multi-step Error Analysis. Figure 4 reports RFNE across interpolation steps for NSKT, Shanghai radar, and SST. A consistent trend is that most models exhibit a clear mid-interval error peak: errors increase as the target moves away from the first snapshot, reach their maximum near the middle of the interval, and decrease as the target approaches the last snapshot. This reflects the increasing ambiguity of interpolation when the target is farthest from both boundary states. EDEN exhibits a relatively flat error curve and shows more uniform behavior across interpolation steps; however, this temporal uniformity does not necessarily lead to lower reconstruction error. U-Net learns a deterministic endpoint-to-target mapping, which may not reliably capture the effect of temporal position across the full interpolation interval.

Across all datasets, `iFLEX` achieves the best overall performance, especially for large-step interpolation on NSKT data, where nonlinear turbulent dynamics make mid-interval prediction challenging. On SST, where the dynamics are smoother and overall errors are lower, `iFLEX` remains best or near-best across most steps. These results show that our proposed method improves multi-step interpolation performance, especially in the difficult large-step regime where endpoint information alone is least informative.

Physics Preservation. The energy spectrum comparison in Figure 5(a) highlights important differences in the physical fidelity of each model. U-Net captures the low-wavenumber structure reasonably well but loses energy at higher wavenumbers, consistent with the smoothing behavior of deterministic regression. Super SloMo overestimates high-wavenumber energy, suggesting that optical-flow-based warping can introduce

spurious small-scale artifacts in turbulent fields. EDEN exhibits the opposite behavior: it preserves the high-wavenumber regime relatively well, but it deviates from the ground truth at low wavenumbers. This indicates that diffusion-based video interpolation can recover fine-scale spectral content, but it may still fail to preserve the dominant large-scale energy-containing modes without an inductive bias tailored to scientific dynamics. In contrast, `iFLEX` more closely matches the ground-truth spectrum across both low- and high-wavenumber regimes, suggesting that boundary-aware coarse interpolation, stochastic refinement, and structured conditioning module provide a better implicit bias for preserving the multi-scale structure of turbulent flows.

Efficiency Analysis. We further analyze the accuracy-efficiency trade-off using parameter count and per-frame inference time in Figure 5(b). Although `iFLEX` is diffusion-based, its refinement formulation and small number of DDIM sampling steps keep inference relatively efficient. Compared with the deterministic U-Net, `iFLEX` uses fewer parameters while improving reconstruction accuracy on challenging scientific datasets. Compared with EDEN, `iFLEX` avoids a heavier latent generative pipeline and achieves stronger accuracy but with more inference cost. Super SloMo remains competitive in inference time, but its optical-flow bias limits accuracy on non-transport-dominated scientific fields. Overall, `iFLEX` provides a favorable trade-off between reconstruction quality, model size, and inference cost.



(a) Energy Spectrum

(b) Accuracy-Efficiency

Figure 5: Additional analysis of interpolation behavior. (a) Energy spectrum comparison evaluates whether models preserve multi-scale physical structures. (b) Efficiency analysis compares reconstruction quality against inference cost and model size.

Generalization to Different Interpolation Steps. Figure 6 evaluates generalization to unseen interpolation horizons on NSKT at $Re = 36000$. All models are trained with 20 interpolation steps and tested with 12, 16, and 24 steps. Across all three testing horizons, `iFLEX` achieves lower RFNE than Super SloMo and EDEN over the full interpolation interval. This demonstrates that the proposed continuous-time conditioning and multi-resolution training allow the model to adapt to interpolation steps that differ from the training setting. Moreover, the baselines exhibit different failure modes. Super SloMo accumulates large errors as the target step moves away from the first endpoint, especially for 12 and 16 testing steps, where the testing horizons are shorter than the training horizon. This shows limited robustness to changes in temporal discretizations. EDEN produces smoother error trends but remains consistently above `iFLEX`, indicating that latent diffusion can not necessarily provide accurate interpolation under changed temporal horizons. These results demonstrate that `iFLEX` generalizes reliably across interpolation horizons, especially in the large-step regime.

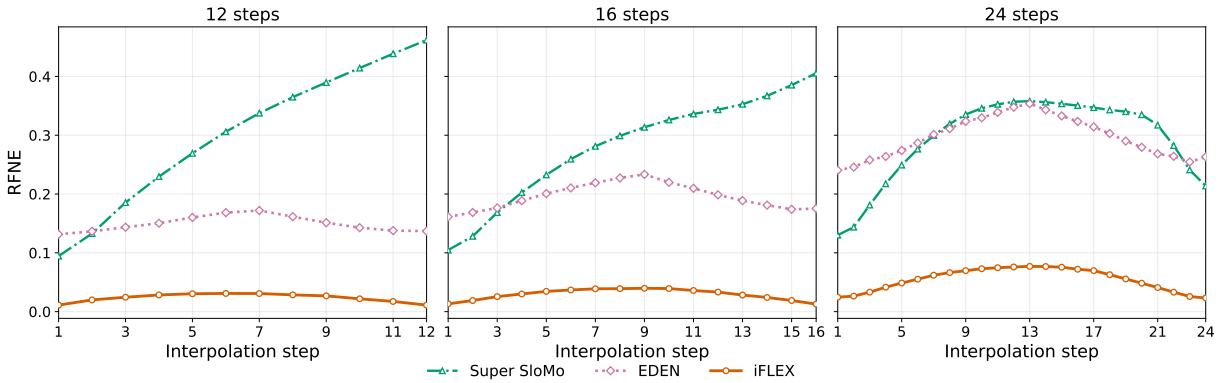
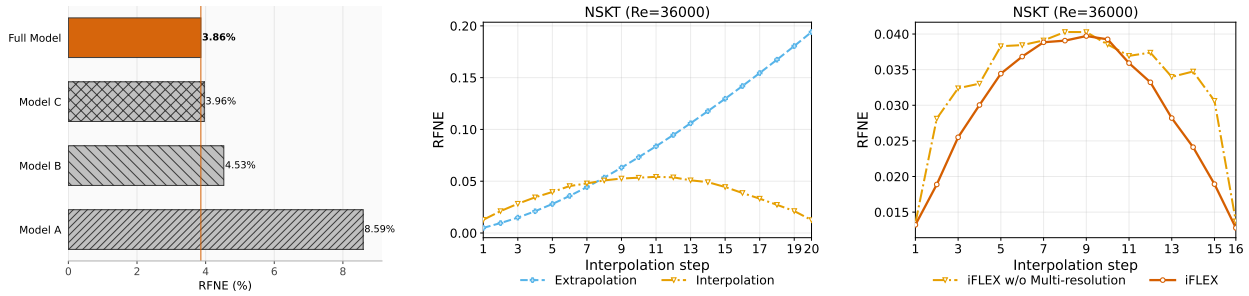


Figure 6: Generalization to unseen interpolation horizons on NSKT at $Re = 36000$. All models are trained with 20 interpolation steps and evaluated with 12, 16, and 24 steps. We report RFNE at each target step for Super SloMo, EDEN, and `iFLEX`.



(a) Performance Summary (b) Full Model vs. Model A (c) Full Model vs. Model C

Figure 7: Ablation study on the NSKT dataset. The results are evaluated at $Re = 36000$. (a) Overall RFNE comparison between the full model and three ablated variants. Model A replaces the proposed interpolation-and-refinement decomposition with direct residual learning from the first endpoint; Model B removes the total interpolation interval Δt from scalar conditioning; and Model C removes multi-resolution training. (b) Effect of boundary-consistent interpolation and refinement (Interpolation), compared to extrapolative learning in the residual space (Extrapolation). (c) Effect of multi-resolution training, compared with fixed-step training under a different testing horizon.

5.5 Ablation Study

In this part, we perform ablations to evaluate three core design choices in `iFLEX`: conditioning on total time interval; multi-resolution training; and the interpolation-and-refinement scheme.

- First, we define Model A as an extrapolation-style variant of `iFLEX`. Specifically, we modify the diffusion learning objective by replacing our interpolation-based refinement target: $\mathbf{r}_\tau = \mathbf{u}_\tau - \mathcal{I}(\mathbf{u}_0, \mathbf{u}_1; \tau)$ with a direct residual from the first endpoint: $\mathbf{r}_\tau^0 = \mathbf{u}_\tau - \mathbf{u}_0$. We further remove the final state $\mathbf{u}_1$ from the conditioning inputs, so the model predicts intermediate states using only the initial snapshot. This ablation removes the numerical interpolation prior and requires the diffusion model to learn both the coarse temporal evolution and fine-scale corrections from a single boundary. Thus, it provides a direct comparison between temporal interpolation and extrapolation-style prediction.

- Second, we define Model B, where we remove the total interpolation interval Δt from the scalar conditioning variables. This tests whether the model needs access to the physical time span between endpoint states, since the same interpolation ratio τ may correspond to different dynamical changes under different Δt.

- Third, we define Model C, where we remove multi-resolution training and train only at the target interpolation resolution. This tests whether exposure to multiple temporal resolutions improves continuous-time generalization and large-step interpolation.

We use the NSKT dataset as a representative case for the ablation study. All variants are trained under the same setup and evaluated at $Re = 36000$. Figure 7(a) summarizes the RFNE results for all models, where each removed component degrades performance, with the full model achieving the lowest RFNE of 3.86%.

Replacing the proposed interpolation-and-refinement design with direct extrapolative learning from the first endpoint leads to significantly high error, showing the benefit of anchoring the stochastic refinement around a boundary-consistent numerical estimate. Moreover, we further analyze the step-wise performance in Figure 7(b), which shows that direct extrapolative learning from the first endpoint suffers from error accumulation: RFNE grows with the interpolation step and reaches its maximum at the final target. In contrast, our interpolation-based refinement uses both endpoints to construct a boundary-aware coarse estimate, leading to the largest error near the middle of the interval, where the target is least constrained by either boundary. This demonstrates that the numerical interpolation prior reduces long-horizon error accumulation and provides a better residual target for diffusion-based refinement.

Model B, which removes the total interpolation interval from the conditioning variables, also increases error, demonstrating the effectiveness of providing the physical time span between endpoint snapshots. This is important for scientific interpolation because the same interpolation ratio τ can correspond to different amounts of physical evolution under different temporal gaps. Furthermore, Model C, which removes multi-resolution training, causes a smaller but consistent degradation. In this ablation, models are trained with 20 interpolation steps and evaluated with 16 steps. As shown in Figure 7(c), removing multi-resolution training yields larger errors over most target steps. This shows that multi-resolution training improves generalization to unseen interpolation horizons. Overall, the ablation study confirms that the performance of `iFLEX` comes from the combination of physical time conditioning, boundary-consistent interpolation-based refinement, and multi-resolution training.

6 Conclusions

We present `iFLEX`, a generative framework for continuous temporal interpolation of scientific data. Unlike many video interpolation methods that rely on optical-flow-based warping, `iFLEX` uses numerical interpolation to construct a boundary-aware coarse estimate and then applies diffusion-based stochastic refinement to recover nonlinear, fine-scale, and uncertain corrections. This design preserves endpoint consistency while avoiding video-specific transport assumptions that may be misaligned with physical fields. We further introduce a structured conditioning mechanism that separately handles high-dimensional physical fields and low-dimensional temporal and physical variables through spatial feature injection and FiLM-based modulation. Experiments on simulated turbulence and real-world datasets demonstrate that `iFLEX` improves reconstruction quality and robustness under challenging interpolation and distribution-shift settings. Overall, our results suggest that scientific temporal interpolation benefits from models whose inductive biases are designed, at least weakly, around physical spatiotemporal structure, rather than being inherited directly from natural-video interpolation.

Broader Impact Statement

This work aims to improve temporal interpolation for scientific spatiotemporal data, with potential benefits for reducing storage and simulation costs and supporting downstream tasks such as weather analysis, climate diagnostics, fluid simulation post-processing, and real-time decision support. However, generated intermediate fields should not be treated as substitutes for direct observations or high-fidelity simulations without validation, especially in safety-critical settings. The model may inherit biases from training data, sensor noise, simulation artifacts, or distribution shifts, which could lead to misleading reconstructions if uncertainty is ignored. Practical use should therefore include uncertainty assessment, out-of-distribution evaluation, physical consistency checks, and domain-expert validation. More broadly, our work emphasizes

the importance of designing interpolation models whose inductive biases are aligned with scientific data, rather than directly transferring assumptions from natural-video modeling.

References

Simon Baker, Daniel Scharstein, James P Lewis, Stefan Roth, Michael J Black, and Richard Szeliski. A database and evaluation methodology for optical flow. *International journal of computer vision*, 92(1): 1–31, 2011.

Seth Bassetti, Brian Hutchinson, Claudia Tebaldi, and Ben Kravitz. Diffesm: Conditional emulation of temperature and precipitation in earth system models with 3d diffusion models. *Journal of Advances in Modeling Earth Systems*, 16(10):e2023MS004194, 2024.

Victor Fonte Chavez, Claudia Esteves, and Jean-Bernard Hayet. Time-adaptive video frame interpolation based on residual diffusion. *arXiv preprint arXiv:2504.05402*, 2025.

Lei Chen, Yuan Cao, Leiming Ma, and Junping Zhang. A deep learning-based methodology for precipitation nowcasting with radar. *Earth and Space Science*, 7(2):e2019EA000812, 2020.

Duolikun Danier, Fan Zhang, and David Bull. Ldmvfi: Video frame interpolation with latent diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 1472–1480, 2024.

Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *International conference on machine learning*, pp. 933–941. PMLR, 2017.

Jiong Dong, Kaoru Ota, and Mianxiong Dong. Video frame interpolation: A comprehensive survey. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(2s):1–31, 2023.

Pan Du, Meet Hemant Parikh, Xiantao Fan, Xin-Yang Liu, and Jian-Xun Wang. Conditional neural field latent diffusion model for generating spatiotemporal turbulence. *Nature Communications*, 15(1):10416, 2024.

N Benjamin Erichson, Lionel Mathelin, Zhewei Yao, Steven L Brunton, Michael W Mahoney, and J Nathan Kutz. Shallow neural networks for fluid flow reconstruction with limited sensors. *Proceedings of the Royal Society A*, 476(2238):20200097, 2020.

N Benjamin Erichson, Vinicius Mikuni, Dongwei Lyu, Yang Gao, Omri Azencot, Soon Hoe Lim, and Michael W Mahoney. Flex: A backbone for diffusion-based modeling of spatio-temporal physical systems. *arXiv preprint arXiv:2505.17351*, 2025.

Soheil Esmaeilzadeh, Kamyar Azizzadenesheli, Karthik Kashinath, Mustafa Mustafa, Hamdi A Tchelepi, Philip Marcus, Mr Prabhat, Anima Anandkumar, et al. Meshfreeflownet: A physics-constrained deep continuous space-time super-resolution framework. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–15. IEEE, 2020.

E.U. Copernicus Marine Service Information (CMEMS). Marine Data Store (MDS). Global Ocean Physics Reanalysis. `https://doi.org/10.48670/moi-00021`, 2023. (Accessed on 23 September 2023).

Kai Fukami, Koji Fukagata, and Kunihiko Taira. Super-resolution reconstruction of turbulent flows with machine learning. *Journal of Fluid Mechanics*, 870:106–120, 2019.

Kai Fukami, Koji Fukagata, and Kunihiko Taira. Machine-learning-based spatio-temporal super resolution reconstruction of turbulent flows. *Journal of Fluid Mechanics*, 909:A9, 2021a.

Kai Fukami, Romit Maulik, Nesar Ramachandra, Koji Fukagata, and Kunihiko Taira. Global field reconstruction from sparse sensors with voronoi tessellation-assisted deep learning. *Nature Machine Intelligence*, 3(11):945–951, 2021b.

Hamid Gadirov, Jos BTM Roerdink, and Steffen Frey. Flint: Learning-based flow estimation and temporal interpolation for scientific ensemble visualization. *IEEE Transactions on Visualization and Computer Graphics*, 2025a.

Hamid Gadirov, Qi Wu, David Bauer, Kwan-Liu Ma, Jos BTM Roerdink, and Steffen Frey. Hyperflint: Hypernetwork-based flow estimation and temporal interpolation for scientific ensemble visualization. In *Computer Graphics Forum*, pp. e70134. Wiley Online Library, 2025b.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The ERA5 Global Reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

L. Hodgkinson and M. W. Mahoney. Multiplicative noise and heavy tails in stochastic optimization. Technical Report Preprint: arXiv:2006.06293, 2020.

Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

Zhilin Huang, Yijie Yu, Ling Yang, Chujun Qin, Bing Zheng, Xiawu Zheng, Zikun Zhou, Yaowei Wang, and Wenming Yang. Motion-aware latent diffusion models for video frame interpolation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1043–1052, 2024.

Siddhant Jain, Daniel Watson, Eric Tabellion, Ben Poole, Janne Kontkanen, et al. Video interpolation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7341–7351, 2024.

Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9000–9008, 2018.

Peishi Jiang, Zhao Yang, Jiali Wang, Chenfu Huang, Pengfei Xue, TC Chakraborty, Xingyuan Chen, and Yun Qian. Efficient super-resolution of near-surface climate modeling using the fourier neural operator. *Journal of Advances in Modeling Earth Systems*, 15(7):e2023MS003800, 2023.

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.

A. S. Krishnapriyan, A. Gholami, S. Zhe, R. M. Kirby, and M. W. Mahoney. Characterizing possible failure modes in physics-informed neural networks. Technical Report Preprint: arXiv:2109.01050, 2021.

Aditi S Krishnapriyan, Alejandro F Queiruga, N Benjamin Erichson, and Michael W Mahoney. Learning continuous models for continuous physics. *Communications Physics*, 6(1):319, 2023.

J Nathan Kutz, Maryam Reza, Farbod Faraji, and Aaron Knoll. Shallow recurrent decoder for reduced order modeling of plasma dynamics. *arXiv preprint arXiv:2405.11955*, 2024.

Dahyeon Kye, Changhyun Roh, Sukhun Ko, Chanho Eom, and Jihyong Oh. Acevfi: A comprehensive survey of advances in video frame interpolation. *arXiv preprint arXiv:2506.01061*, 2025.

Lizao Li, Robert Carver, Ignacio Lopez-Gomez, Fei Sha, and John Anderson. Generative emulation of weather forecast ensembles with diffusion models. *Science Advances*, 10(13):eadk4489, 2024.

Yihao Liu, Liangbin Xie, Li Siyao, Wenxiu Sun, Yu Qiao, and Chao Dong. Enhanced quadratic video interpolation. In *European conference on computer vision*, pp. 41–56. Springer, 2020.

Ignacio Lopez-Gomez, Zhong Yi Wan, Leonardo Zepeda-Núñez, Tapio Schneider, John Anderson, and Fei Sha. Dynamical-generative downscaling of climate model ensembles. *Proceedings of the National Academy of Sciences*, 122(17):e2420288122, 2025.

Dongwei Lyu, Rie Nakata, Pu Ren, Michael W Mahoney, Arben Pitarka, Nori Nakata, and N Benjamin Erichson. Rapid wavefield forecasting for earthquake early warning via deep sequence to sequence learning. *Nature Communications*, 2025.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pp. 23–40. Springer, 2024.

Ilan Naiman, N. Benjamin Erichson, Pu Ren, Michael W. Mahoney, and Omri Azencot. Generative modeling of regular and irregular time series data via koopman vaes. In *ICLR*. OpenReview.net, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Pu Ren, Chengping Rao, Yang Liu, Zihan Ma, Qi Wang, Jian-Xun Wang, and Hao Sun. Physr: Physics-informed deep super-resolution for spatiotemporal data. *Journal of Computational Physics*, 492:112438, 2023.

Pu Ren, N Benjamin Erichson, Junyi Guo, Shashank Subramanian, Omer San, Zarija Lukic, and Michael W Mahoney. Superbench: A super-resolution benchmark dataset for scientific machine learning. *Journal of Data-centric Machine Learning Research*, 2025.

Pu Ren, Rie Nakata, Maxime Lacour, Ilan Naiman, Nori Nakata, Jialin Song, Zhengfa Bi, Osman Asif Malik, Dmitriy Morozov, Omri Azencot, et al. Learning earthquake ground motions via conditional generative modeling. *Nature Communications*, 2026.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Salva Rühling Cachay, Bo Zhao, Hailey Joren, and Rose Yu. Dyffusion: A dynamics-informed diffusion model for spatiotemporal forecasting. *Advances in neural information processing systems*, 36:45259–45287, 2023.

Salva Rühling Cachay, Brian Henn, Oliver Watt-Meyer, Christopher S Bretherton, and Rose Yu. Probablistic emulation of a global climate model with spherical dyffusion. *Advances in Neural Information Processing Systems*, 37:127610–127644, 2024.

M. Sakarvadia, K. Hegazy, A. Totounferoush, K. Chard, Y. Yang, I. Foster, and M. W. Mahoney. The false promise of zero-shot super-resolution in machine-learned operators. Technical Report Preprint: arXiv:2510.06646, 2025.

Jialin Song, Zezheng Song, Pu Ren, N Benjamin Erichson, Michael W Mahoney, and Xiaoye S Li. Forecasting high-dimensional spatio-temporal systems from sparse measurements. *Machine Learning: Science and Technology*, 5(4):045067, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*. OpenReview.net, 2021.

Zhiqing Sun, Yiming Yang, and Shinjae Yoo. A neural pde solver with temporal stencil modeling. In *International Conference on Machine Learning*, pp. 33135–33155. PMLR, 2023.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Lanxiang Xing, Haixu Wu, Yuezhou Ma, Jianmin Wang, and Mingsheng Long. Helmfluid: Learning helmholtz dynamics for interpretable fluid prediction. *arXiv preprint arXiv:2310.10565*, 2023.

Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. *Advances in Neural Information Processing Systems*, 32, 2019.

A. Yu, D. C. Maddix, B. Han, X. Zhang, A. F. Ansari, O. Shchur, C. Faloutsos, A. G. Wilson, M. W. Mahoney, and Y. Wang. Understanding the implicit biases of design choices for time series foundation models. Technical Report Preprint: arXiv:2510.19236, 2025.

Demin Yu, Xutao Li, Yunming Ye, Baoquan Zhang, Chuyao Luo, Kuai Dai, Rui Wang, and Xunlai Chen. Diffcast: A unified framework via residual diffusion for precipitation nowcasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27758–27767, 2024.

Zihao Zhang, Haoran Chen, Haoyu Zhao, Guansong Lu, Yanwei Fu, Hang Xu, and Zuxuan Wu. Eden: Enhanced diffusion for high-quality large-motion video frame interpolation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2105–2115, 2025.

A Implementation Details

A.1 Datasets

We evaluate models on a diverse set of scientific spatiotemporal datasets spanning both controlled simulations and real-world observations. Table 3 summarizes the datasets used in our experiments. For datasets with known Reynolds numbers, we incorporate them into the model as physical conditioning variables.

Table 3: Summary of the used scientific datasets. "Interp. Length" denotes the interpolation length for each dataset.

Dataset	Interp. Length	Resolution	(Training, Validation, Test)	Observed State	Reynolds Numbers
NSKT	20	256×256	$(16000, 4000, 2000)$	Vorticity	$600 \sim 36000$
Shanghai Radar	8	128×128	$(1545, 850, 850)$	Precipitation	Unknown
Sea Surface Temperature	10	128×128	$(172765, 18278, 66222)$	Temperature	Unknown

A.2 Evaluation Metrics

We evaluate interpolation quality using three complementary metrics: relative Frobenius norm error (RFNE), Pearson correlation (PC), and structural similarity index (SSIM). RFNE measures the relative reconstruction error, PC evaluates the linear correlation between the predicted and ground-truth physical fields, and SSIM captures the structural similarity of spatial patterns. Let $\widehat{\mathbf{u}}_\tau$ and $\mathbf{u}_\tau$ denote the predicted and ground-truth intermediate fields. We compute RFNE and PC in the original physical space:

$$\mathrm{RFNE} = \frac{\|\widehat{\mathbf{u}}_\tau - \mathbf{u}_\tau\|_F}{\|\mathbf{u}_\tau\|_F}, \tag{20}$$

and

$$\mathrm{PC} = \frac{\left\langle \widehat{\mathbf{u}}_\tau - \overline{\widehat{\mathbf{u}}}_\tau,\, \mathbf{u}_\tau - \overline{\mathbf{u}}_\tau \right\rangle}{\left\| \widehat{\mathbf{u}}_\tau - \overline{\widehat{\mathbf{u}}}_\tau \right\|_F \|\mathbf{u}_\tau - \overline{\mathbf{u}}_\tau\|_F}, \tag{21}$$

where $\overline{\widehat{\mathbf{u}}}_\tau$ and $\overline{\mathbf{u}}_\tau$ denote spatial averages. For SSIM, we first normalize both predicted and ground-truth fields to a common range and then compute

$$\mathrm{SSIM} = \frac{(2\mu_{\widehat{\mathbf{u}}}\mu_{\mathbf{u}} + c_1)(2\sigma_{\widehat{\mathbf{u}}\mathbf{u}} + c_2)}{(\mu^2_{\widehat{\mathbf{u}}} + \mu^2_{\mathbf{u}} + c_1)(\sigma^2_{\widehat{\mathbf{u}}} + \sigma^2_{\mathbf{u}} + c_2)}, \tag{22}$$

where $\mu_{\widehat{\mathbf{u}}}$ and $\mu_{\mathbf{u}}$ are pixel sample means, $\sigma^2_{\widehat{\mathbf{u}}}$ and $\sigma^2_{\mathbf{u}}$ are sample variances, $\sigma_{\widehat{\mathbf{u}}\mathbf{u}}$ is the sample covariance, and c_1, c_2 are stabilization constants. Lower RFNE indicates better reconstruction accuracy, while higher PC and SSIM indicate better correlation and structural fidelity.

We also consider the energy spectrum for evaluating the macroscopic physical properties of fluid data. Specifically, the energy spectrum indicates the distribution of turbulence kinetic energy (TKE) across different length scales, ensuring an accurate representation of both large and small structures.

A.3 Model and Experiment Configurations

All experiments are conducted on four NVIDIA A100 GPUs with 40 GB memory in a standard workstation environment. To ensure a fair comparison, all models are trained and evaluated under the same dataset splits, interpolation horizons, and evaluation protocols. We describe the model architectures, conditioning inputs, optimization objectives, and inference settings for `iFLEX` and all baselines below.

U-Net. The U-Net baseline is a deterministic supervised interpolation model. It takes the two endpoint frames as a two-channel input and directly predicts a single intermediate frame, which is optimized against the

ground truth using an ℓ_1 reconstruction loss. We follow the U-Net architecture used in Erichson et al. (2025), with base width 128, two residual blocks per resolution level, channel multipliers $(1, 2, 4, 8)$ for 128×128 and 256×256 inputs, dropout rate 0.15, scale-shift normalization, residual up/downsampling, and multi-head attention at feature resolutions 32, 16, and 8. For all datasets, spatial conditioning is implemented by concatenating the two endpoint frames along the channel dimension. Scalar conditioning variables, including the interpolation ratio, total interpolation interval, and available Reynolds numbers, are encoded using Fourier embeddings and injected into the U-Net through the residual-block embedding pathway.

Table 4: Training parameters for U-Net across all three datasets.

Dataset	Epochs	Batch Size	Learning Rate	Optimizer
NSKT	150	48	1×10^{-4}	Adam
Shanghai	400	48	1×10^{-4}	Adam
SST	50	64	1×10^{-4}	Adam

Super SloMo. We follow the default Super SloMo setup of Jiang et al. (2018), which uses a two-network optical-flow-based interpolation architecture, and adapt it from RGB video to single-channel scientific fields. Specifically, the first U-Net takes the two endpoint frames concatenated along the channel dimension and predicts bidirectional endpoint flows. The second U-Net takes the endpoint frames, endpoint flows, estimated intermediate flows, and warped frames as input, and predicts flow residuals together with a visibility map. Super SloMo is conditioned on the two endpoint frames and the target interpolation time. The target time is not incorporated through an embedding layer, but instead enters analytically through the flow interpolation and warping coefficients. We do not provide available Reynolds numbers to Super SloMo. The model is trained with supervised reconstruction loss against one intermediate target frame, and we omit the VGG perceptual loss since it is designed for natural images and is not appropriate for our scientific data.

Table 5: Training parameters for Super SloMo across all three datasets.

Dataset	Epochs	Batch Size	Learning Rate	Optimizer
NSKT	200	256	6×10^{-5}	Adam
Shanghai	400	128	1×10^{-4}	Adam
SST	200	128	1×10^{-4}	Adam

EDEN. We follow the two-stage latent generative interpolation setup of Zhang et al. (2025). The first stage trains a conditional variational autoencoder (VAE) to learn a compact latent representation of the intermediate frame conditioned on the two endpoint frames. The second stage trains a diffusion transformer (DiT) (Peebles & Xie, 2023; Ma et al., 2024) in the VAE latent space to generate the latent code of an arbitrary intermediate frame. At inference time, the DiT samples a latent representation conditioned on the endpoints and the requested interpolation ratio, and the VAE decoder maps the generated latent back to the physical field. The EDEN VAE uses a patch-based transformer architecture with single-channel inputs and outputs, patch size 16, hidden dimension 768, 12 attention heads, MLP ratio 4, and latent dimension 16. The encoder and decoder each use four transformer blocks with temporal attention over the start, end, and target frames. The endpoint frames are embedded as conditioning tokens, while the target frame is encoded into low-resolution latent tokens parameterized by a diagonal Gaussian posterior. For scientific data, we make two implementation changes: we disable reconstruction clamping, which is designed for bounded image pixels but is unsuitable for normalized physical fields such as SST, and we remove image-style loss scaling that assumes inputs lie in $[0, 1]$. We select checkpoints using the validation ℓ_1 error instead of PSNR. The EDEN DiT operates entirely in the learned VAE latent space. It uses latent dimension 16, transformer width 768, 12 transformer blocks, 12 attention heads, and MLP ratio 4. The model receives noisy latent tokens and predicts the velocity field under a linear transport objective. Conditioning is provided by endpoint frame tokens, the interpolation ratio, and a normalized endpoint-difference statistic computed from the cosine similarity between the two endpoint frames. The denoising timestep, interpolation ratio, and endpoint-difference embedding are combined and injected through adaptive layer normalization in the DiT blocks.

Table 6: Training parameters for EDEN across all three datasets.

Dataset	Epochs (VAE)	Batch Size (VAE)	Learning Rate (VAE)	Epochs (DiT)	Batch Size (DiT)	Learning Rate (DiT)	Optimizer
NSKT	300	48	1.25×10^{-5}	300	48	1×10^{-4}	AdamW
Shanghai	300	256	5×10^{-5}	400	256	5×10^{-5}	AdamW
SST	100	256	5×10^{-5}	50	256	5×10^{-4}	AdamW

Our Model. `iFLEX` is a conditional diffusion-based interpolation model. Instead of modeling the full physical field directly, we train the diffusion model in the refinement space defined relative to the linear interpolation estimate. We use a cosine log-SNR diffusion schedule, velocity prediction, and an ℓ_1 denoising objective. At inference time, we sample the refinement using a DDIM-style sampler with three denoising steps and add the generated residual to the linear interpolation estimate. The `iFLEX` architecture consists of a condition-aware denoising encoder-decoder, a shared spatial encoder, and a scalar encoder. Across all experiments, we use channel widths $[64, 128, 256, 512]$, two residual blocks per resolution level in both the encoder and decoder, eight attention heads, and an MLP ratio of 2. The same architecture is used for NSKT, Shanghai radar, and SST experiments, with only dataset-specific input resolution and temporal horizon adjusted. Scalar conditioning variables, including the interpolation ratio, total interpolation interval, available physical parameters, and diffusion timestep, are Fourier-embedded and injected into the denoising network through the timestep-conditioning pathway. Spatial conditioning is provided by the two endpoint frames. A shared-weight endpoint encoder processes $\mathbf{u}_0$ and $\mathbf{u}_1$, and the resulting features are fused using the endpoint features, their average, and their difference. The fused spatial condition maps are injected into the denoising encoder and decoder at 32×32 and 16×16 resolutions through gated feature modulation. The training configurations are also applied to the ablation study in Section 5.5.

Table 7: Training parameters for `iFLEX` across all three datasets.

Dataset	Epochs	Batch Size	Learning Rate	Optimizer
NSKT	150	48	1×10^{-4}	Adam
Shanghai	300	48	1×10^{-4}	Adam
SST	50	128	1×10^{-4}	Adam

B Additional Results

We provide qualitative comparisons of model predictions against the ground truth across all datasets to assess spatial fidelity beyond evaluation metrics. These visualizations highlight how different interpolation strategies affect the reconstruction of coherent structures, fine-scale details, and physically meaningful patterns. In particular, they reveal failure modes that are not fully captured by scalar metrics, such as over-smoothing, spatial distortion, and loss of localized dynamics. We present representative snapshot comparisons for highly turbulent NSKT dynamics ($Re = 36000$) in Figure 8, real-world precipitation data in Figure 9, and sea surface temperature data in Figure 10.

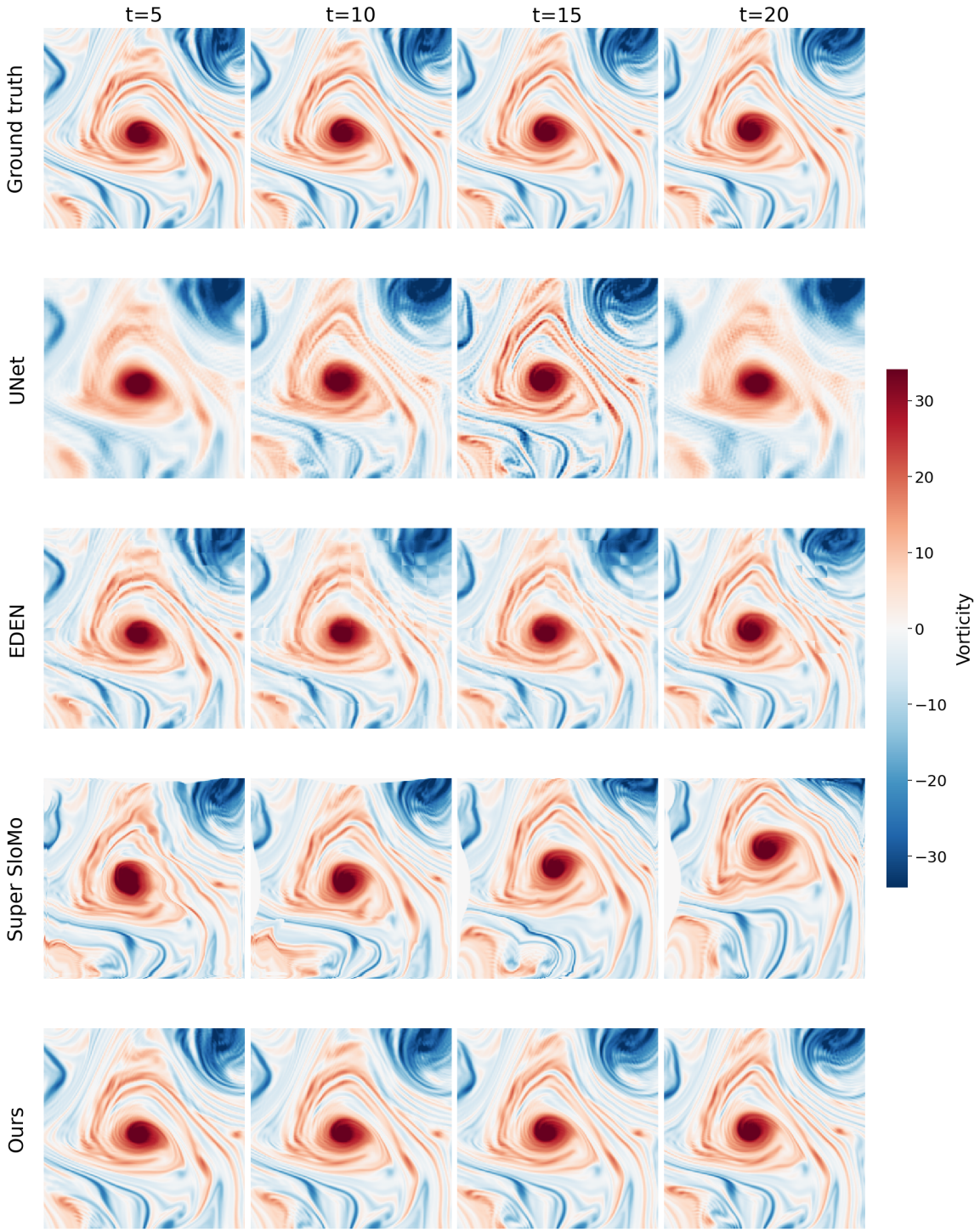


Figure 8: Visualization of ground truth and all model predictions for NSKT ($Re = 36000$) data. The labeled time denotes the interpolation step index.

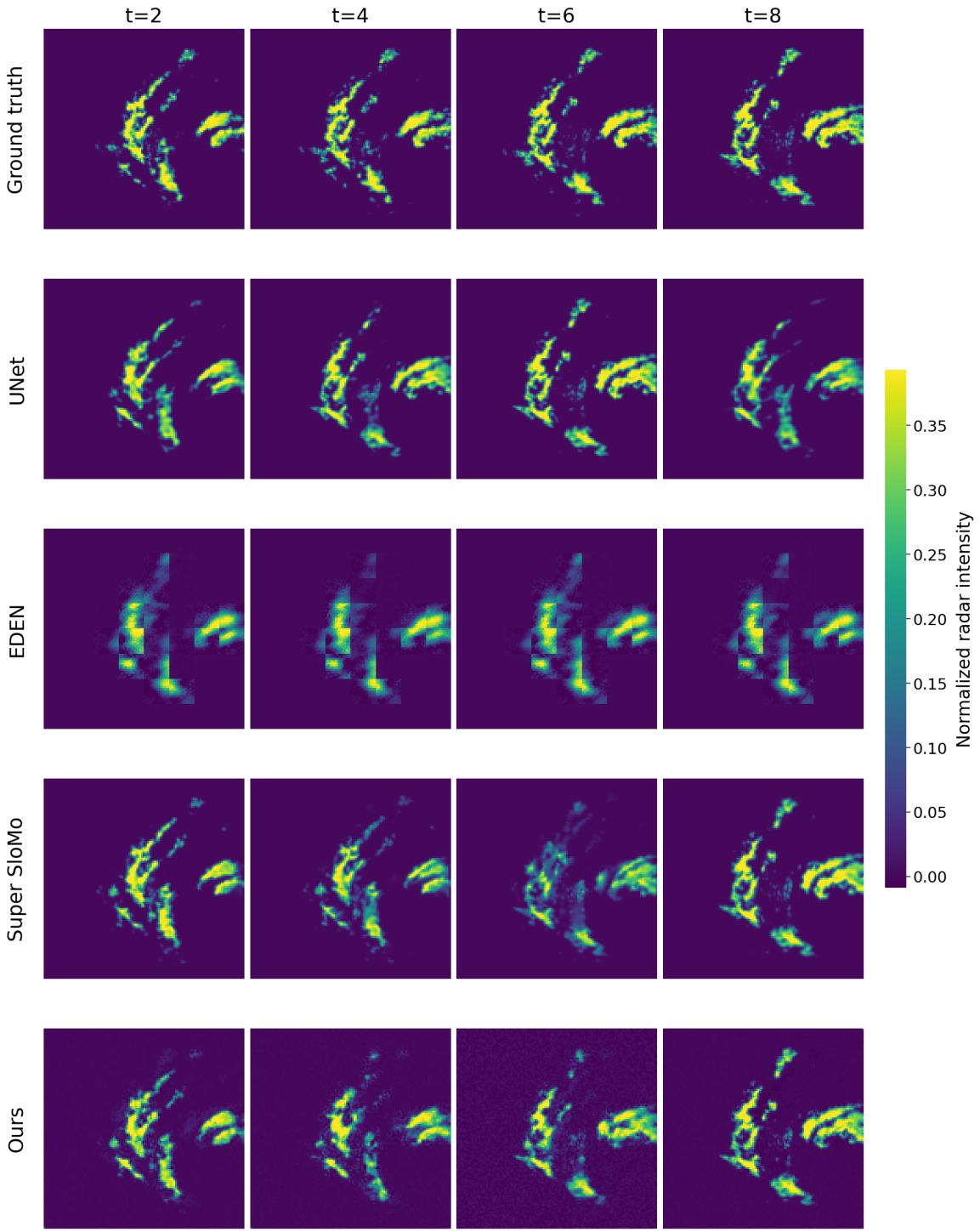


Figure 9: Visualization of ground truth and all model predictions for Shanghai radar data. The labeled time denotes the interpolation step index.

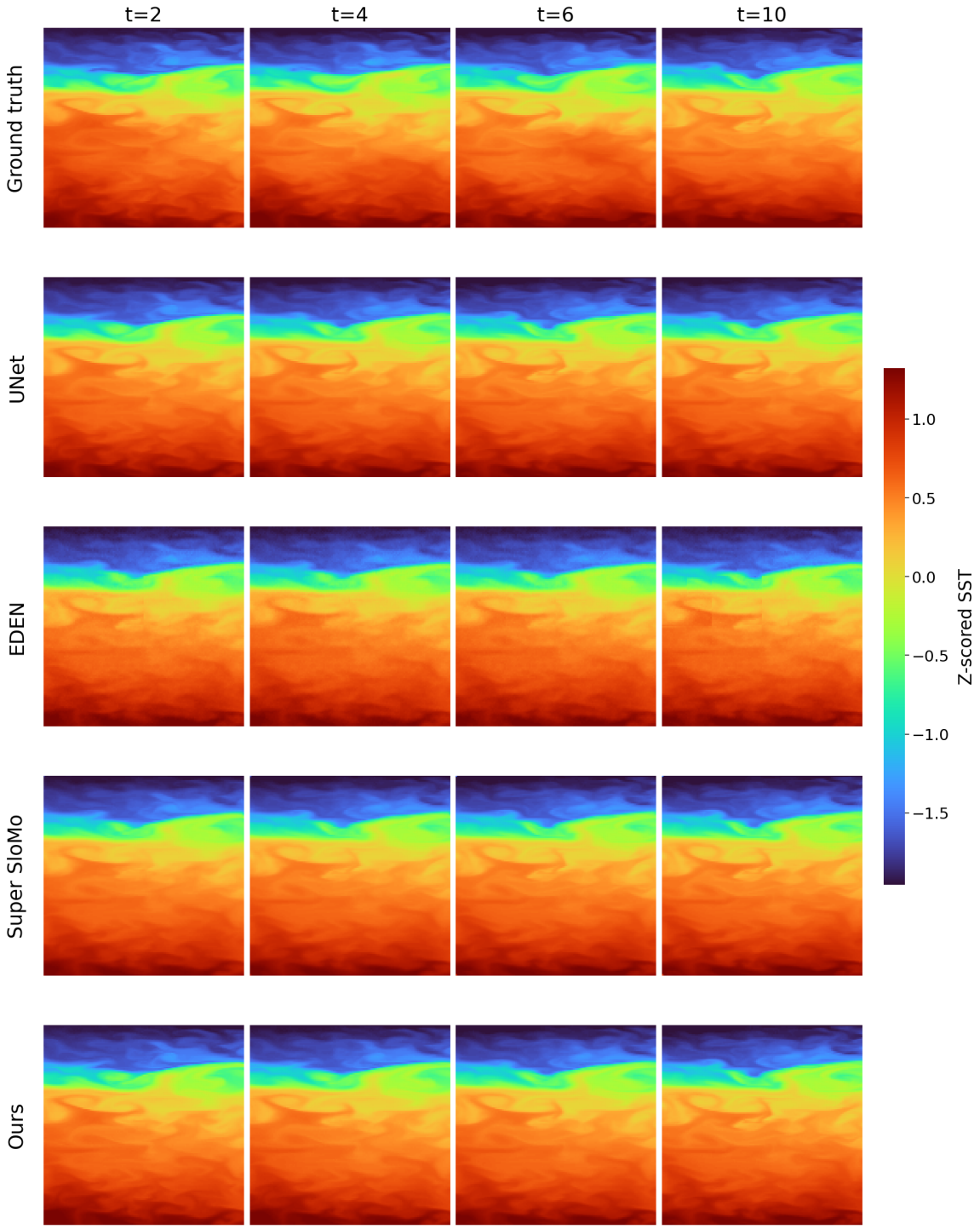


Figure 10: Visualization of ground truth and all model predictions for SST data. The labeled time denotes the interpolation step index.