PhyRE-Net: A Physics-Anchored Network with Active Spatial Gating for Industrial Infrared Thermal Forecasting

Shuotong Yang, **Haitao Zhang**
State Key Laboratory of Networking and Switching Technology
Beijing Key Lab of Intelligent Telecommunication Software and Multimedia
Beijing University of Posts and Telecommunications
yangshuotong@bupt.edu.cn, zht@bupt.edu.cn

Abstract

Spatiotemporal forecasting of infrared thermal fields is a critical technology for the predictive maintenance of industrial electrical control equipment. However, many existing methods are designed under index-based temporal representations and global regression objectives, which are not well aligned with schedule-driven, non-stationary industrial regimes and the long-tailed spatial distribution of safety-critical hotspots. To address these issues, this paper proposes PhyRE-Net, a physical-time-anchored network architecture featuring active spatial gating. Specifically, we first design the multi-period resolution rotary positional embeddings mechanism within the backbone, which injects absolute temporal features into attention geometry to align predicted phases with non-stationary production schedules. We then devise a reinforcement learning-driven spatial gating mechanism to dynamically modulate feature responses, actively shifting the computational focus from the dominant background to sparse but critical thermal anomalies. Extensive experiments on a large-scale real-world dataset covering seven types of heterogeneous electrical equipment show that PhyRE-Net achieves consistent improvements over strong baselines across all evaluated metrics. The source code is available at https://github.com/YST-10/PhyRE-Net.

1 Introduction

Electrical control equipment is central to industrial automation, where any failure can disrupt entire production lines. In this context, infrared thermography (IRT) provides non-contact, full-field temperature maps and reveals localized heating patterns that point-based sensors miss [Kim *et al.*, 2021]. Therefore, the spatiotemporal thermal sequences captured by IRT can be used to model thermal evolution and forecast future fields, enabling earlier warning and maintenance.

To achieve high-accuracy spatiotemporal thermal forecasting, existing approaches can be broadly categorized into three paradigms. The first comprises spatiotemporal recurrent neural networks (RNNs) [Lipton *et al.*, 2015], such as ConvLSTM [Shi *et al.*, 2015] and PredRNN v2 [Wang *et al.*, 2022], which capture thermal inertia via recurrent hidden-state updates. However, these methods typically represent time purely as discrete step indices, lacking an explicit notion of absolute physical time. This temporal agnosticism makes it difficult to capture regime shifts driven by production schedules. Consequently, long-horizon forecasts often suffer from phase drift, where predicted trajectories gradually desynchronize from the actual operating cycles.

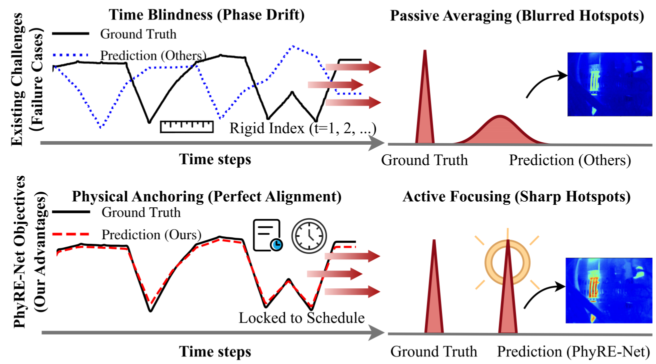


Figure 1: Current Challenges vs. PhyRE-Net Objectives. **Top:** Existing methods suffer from time blindness (phase drift) and passive averaging (blurred hotspots). **Bottom:** PhyRE-Net proposes physical anchoring for precise schedule alignment and active focusing to preserve sharp, safety-critical hotspots.

The second category comprises global regression models built on CNNs and Transformers [Wu, 2017; Vaswani *et al.*, 2017], which abandon recurrence in favor of global receptive fields. For example, CNN-based methods such as SimVP [Gao *et al.*, 2022] employ large-kernel convolutions, while Transformer-based predictors like PatchTST [Nie *et al.*, 2023] and TimeMixer [Wang *et al.*, 2024] leverage long-range token and feature mixing. Despite strong general performance, these models face two limitations in industrial thermal prediction. Temporally, most rely on index-based or relative time representations without an explicit physical time anchor, causing long-horizon predictions to deviate from schedule-driven thermal rhythms. Spatially, their deterministic regression objective is dominated by the background, favoring smoothed estimates that attenuate sparse, safety-critical hotspots and weaken fault localization.

Finally, probabilistic generative models such as DiT [Peebles and Xie, 2023] and Diffusion-TS [Yuan and Qiao, 2024] cast forecasting as conditional generation. They model the conditional distribution for fine-grained thermal textures via iterative denoising. However, stochastic sampling can yield hallucinated hotspots despite sharper high-frequency details. This trade-off favors visual realism over deterministic reliability, complicating safety-critical hotspot verification.

Fundamentally, these limitations stem from a mismatch between generic architectural designs and the unique characteristics of industrial infrared data, as illustrated in Fig. 1. On one hand, thermal evolution is tightly coupled with variable production schedules and non-stationary operational regimes rather than a stationary periodic process. On the other hand, critical anomalies follow an extreme long-tailed spatial distribution, occupying only a small fraction of pixels yet carrying the majority of safety risks. This creates a dual requirement to temporally align with schedule-driven periodicities while spatially preserving sparse hotspots, with deterministic and reliable predictions required for safety-critical monitoring.

To address these requirements, we propose PhyRE-Net, a physical-time-anchored network with active spatial focusing. We devise the Multi-Period Resolution Rotary Positional Embedding (MP-RoPE) in the backbone. Building on RoPE, MP-RoPE decomposes absolute timestamps into multi-resolution frequency bands to capture long-range, phase-consistent temporal dependencies. This design separates environmental drift from operational fluctuations, providing a consistent physical temporal coordinate system for schedule-driven thermal evolution. We then devise Dense Residual PPO (DR-PPO), a reinforcement learning driven spatial gating mechanism for active spatial focusing. DR-PPO adapts PPO to learn pixel-wise continuous gain maps that modulate decoder features, conditioned on historical variance and real-time residuals, thereby shifting computational focus from the dominant background to sparse safety-critical hotspots. With rewards emphasizing extreme accuracy and trend consistency, it supports deterministic inference while mitigating sampling-induced spurious hotspots.

Our main contributions are summarized as follows:

- We propose MP-RoPE, a multi-period resolution rotary positional embedding mechanism that injects absolute timestamps to improve long-horizon phase alignment under schedule-driven, non-stationary regimes.
- We design DR-PPO, a reinforcement learning driven pixel-wise spatial gating mechanism that enables active spatial focusing and improves sparse hotspot preservation with deterministic inference.
- We demonstrate consistent improvements over strong baselines on heterogeneous real-world industrial datasets across all evaluated metrics, particularly on safety-critical top-ρ hotspot metrics.

2 Related Work

2.1 Spatiotemporal Recurrent Neural Networks

Recurrent spatiotemporal predictors capture temporal inertia by iteratively updating hidden states. ConvLSTM [Shi *et al.*, 2015] pioneered this paradigm, followed by variants enhancing motion modeling and spatial coherence, such as E3D-LSTM [Wang *et al.*, 2018b], TrajGRU [Shi *et al.*, 2017], and SA-ConvLSTM [Lin *et al.*, 2020]. In parallel, the PredRNN family [Wang *et al.*, 2017; Wang *et al.*, 2018a; Wang *et al.*, 2022] strengthens long-term dynamics via spatiotemporal memory propagation. MIM [Wang *et al.*, 2019] further targets non-stationarity through differential memory updates, while recent hybrids like SwinLSTM [Tang *et al.*, 2023] integrate attention mechanisms. However, most recurrent formulations parameterize time primarily as discrete step indices rather than absolute physical timestamps, making them susceptible to schedule-driven regime shifts and cumulative phase drift in long-horizon industrial forecasting. In this paper, we explore an absolute-time-aware forecasting approach to improve long-horizon synchronization.

2.2 Spatiotemporal CNNs and Transformers

Global regression models forecast future representations directly without recurrent updates. In vision forecasting, CNN-based predictors such as U-Net variants [Ronneberger *et al.*, 2015] and SimVP [Gao *et al.*, 2022] perform spatiotemporal field regression, while physics-inspired designs such as PhyDNet [Guen and Thome, 2020] incorporate inductive biases. Transformer forecasters for long sequences, including Informer [Zhou *et al.*, 2021], Autoformer [Wu *et al.*, 2021], and FEDformer [Zhou *et al.*, 2022], and high-dimensional variants such as PatchTST [Nie *et al.*, 2023], iTransformer [Liu *et al.*, 2024], TimeMixer [Wang *et al.*, 2024], and TimeMixer++ [Wang *et al.*, 2025], further strengthen long-range dependency modeling. However, two limitations remain critical for industrial thermal forecasting. Temporally, widely adopted encoding schemes, ranging from the standard sinusoidal absolute embedding [Vaswani *et al.*, 2017] and relative position biases [Raffel *et al.*, 2020] to the recent rotary position embedding (RoPE) [Su *et al.*, 2024], rely primarily on dimensionless sequence indices. Although some approaches incorporate learnable time representations [Kazemi *et al.*, 2019; Press *et al.*, 2022], they typically lack an explicit alignment with schedule-informed absolute physical time, leading to phase shifts under non-stationary regimes. Spatially, deterministic MSE-style regression tends to regress toward the mean, smoothing sparse hotspots under background dominance even with static reweighting, and independent patch or token modeling can further weaken spatially coupled thermal continuity [Mathieu *et al.*, 2016; Lin *et al.*, 2017]. In this paper, we explore explicit absolute-time cues and hotspot-aware learning to improve long-horizon alignment and anomaly fidelity.

2.3 Probabilistic Generative Models

Probabilistic generative models formulate forecasting as conditional generation to capture fine-grained details. Foundations include VAEs [Kingma and Welling, 2014] and GANs [Goodfellow *et al.*, 2014], while flow-based architectures like Glow [Kingma and Dhariwal, 2018] and CrevNet [Yu *et al.*, 2020] further improved fidelity. Recently, diffusion paradigms have dominated, encompassing approaches such as Video Diffusion Models [Ho *et al.*, 2022]

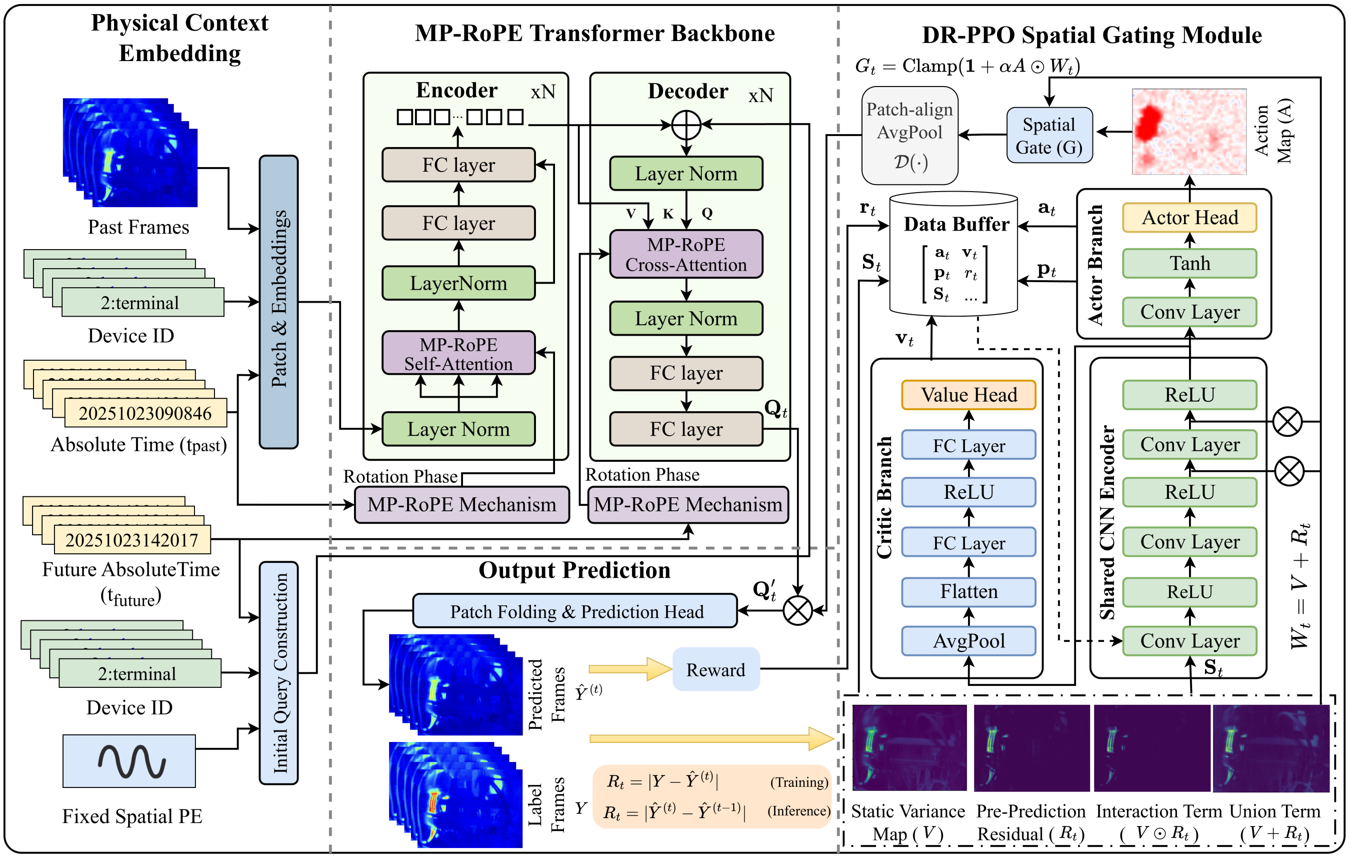


Figure 2: The overall architecture of PhyRE-Net. It adopts a coarse-to-fine synergistic design comprising three integral components: physical context embedding to inject absolute schedule priors, an MP-RoPE transformer backbone for phase-aligned global modeling, and a DR-PPO spatial gating module to actively refine sparse, safety-critical hotspots.

and MCVD [Voleti *et al.*, 2022], alongside Transformer-integrated backbones like DiT [Peebles and Xie, 2023] and Diffusion-TS [Yuan and Qiao, 2024]. However, stochastic sampling risks introducing spurious hotspots and artifacts, rendering them unsuitable for safety-critical monitoring which demands verifiable reliability. In this paper, we propose a deterministic approach with active focusing to eliminate such sampling-induced.

3 Methodology

3.1 Problem Formulation

Let $\mathcal{X}_{\text{in}} \in \mathbb{R}^{T_{\text{in}} \times C \times H \times W}$ denote the historical infrared sequence, where T_{in} is the observation window length, C is the channel dimension, and $H \times W$ is the spatial resolution. We aim to forecast the future thermal sequence $\mathcal{Y}_{\text{out}} \in \mathbb{R}^{T_{\text{out}} \times C \times H \times W}$. In industrial monitoring, thermal evolution is closely tied to absolute time (e.g., production schedules) and device-specific operating regimes, which are not explicitly represented by index-based predictors. We therefore introduce a physical context tuple $\mathcal{C} = (\mathbf{t}_{\text{abs}}, d_{\text{id}})$, where $\mathbf{t}_{\text{abs}} \in \mathbb{R}^{T_{\text{in}}+T_{\text{out}}}$ denotes continuous absolute timestamps normalized on a schedule-consistent scale, and $d_{\text{id}} \in \{1, \dots, N_D\}$ indicates the device identity. Our goal is to learn a mapping $\mathcal{F}_\theta$ that predicts $\hat{\mathcal{Y}}_{\text{out}}$ conditioned on both the infrared observations and the physical context:

$$\hat{\mathcal{Y}}_{\text{out}} = \mathcal{F}_\theta(\mathcal{X}_{\text{in}}, \mathbf{t}_{\text{abs}}, d_{\text{id}}). \quad (1)$$

This setting requires maintaining global fidelity while improving the accuracy of sparse, safety-critical hotspots.

3.2 Overall Architecture

Fig. 2 provides an overview of PhyRE-Net, which integrates physical context embedding, an MP-RoPE Transformer backbone, and a DR-PPO spatial gating module for active spatial focusing in long-horizon infrared thermal forecasting.

Given historical infrared frames $\mathcal{X}_{\text{in}}$, we patchify each frame into visual tokens e_{vis}. We inject physical context by adding a learnable device embedding e_{dev} indexed by d_{id} and an absolute-time embedding e_{time} derived from $\mathbf{t}_{\text{abs}}$, yielding context-aware token representations. These tokens are then processed by an encoder–decoder Transformer backbone to model spatiotemporal dependencies. To incorporate schedule-informed absolute time beyond index-based encodings, MP-RoPE maps timestamps into multi-period rotation phases and applies them to the Q/K in attention, enabling phase-consistent interactions across disjoint time windows.

To improve fidelity on sparse, safety-critical hotspots under background dominance, we introduce a DR-PPO spatial

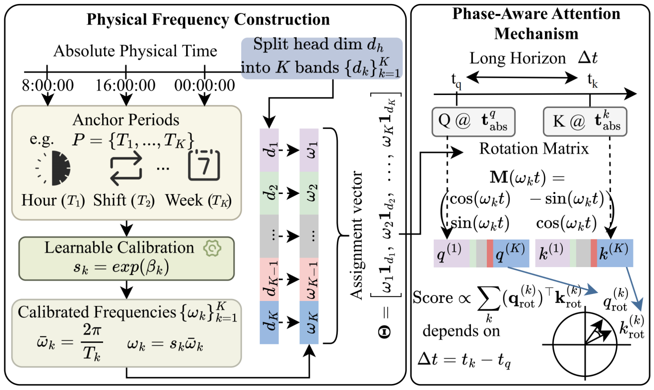


Figure 3: Illustration of MP-RoPE. It decomposes absolute time into multi-resolution periods to drive rotary phase injection in attention, enabling coherent alignment via long-term phase resonance.

gating module to modulate the decoder. It constructs a structured difficulty state from a static variance map V and a residual proxy R, together with their interaction terms, and outputs a continuous spatial gate G. The gate is designed to operate without access to ground truth and remains applicable at inference time. After alignment to the patch grid, it is applied to decoder query tokens as $Q' = Q \odot G$, reallocating representational capacity toward hard regions while retaining stable global structure. Details are in Sec. 3.3 and Sec. 3.4.

3.3 The MP-RoPE Mechanism

Industrial thermal dynamics exhibit multi-scale regularities from routines and environmental cycles. To encode them, we propose Multi-Period Resolution Rotary Positional Embedding (MP-RoPE), as shown in Fig. 3. Unlike index-based encodings, MP-RoPE uses absolute timestamps $\mathbf{t}_{\text{abs}}$, schedule-informed anchor periods $\mathcal{P}$, and data-driven frequency calibration. It partitions each head into $K = |\mathcal{P}|$ subspaces and applies absolute-time rotary shifts, yielding a phase-consistent coordinate system for long-horizon modeling.

Physical Frequency Construction

MP-RoPE builds a multi-resolution frequency basis from anchor periods $\mathcal{P} = \{T_k\}_{k=1}^{K}$ and calibrates them in training. We define anchor angular frequencies $\bar{\omega}_k = 2\pi/T_k$ and use a learnable positive scale $s_k = \exp(\beta_k)$ to obtain:

$$\omega_k = s_k \bar{\omega}_k, \quad k = 1, \dots, K. \tag{2}$$

The parameters $\{\beta_k\}$ are optimized jointly with the backbone. When schedule information is available, $\mathcal{P}$ can be initialized from routine time scales such as intra-shift, shift-level, and daily cycles, and the calibration adapts them. Given per-head dimension d_h, we partition it into K contiguous bands, using equal sizes when d_h is divisible by K and assigning the remaining channels to the last band otherwise. Each band is paired into 2D rotation units and assigned one ω_k. The resulting assignment vector $\mathbf{\Theta} \in \mathbb{R}^{d_h}$ is

$$\mathbf{\Theta} = [\omega_1 \mathbf{1}_{d_1},\ \omega_2 \mathbf{1}_{d_2},\ \dots,\ \omega_K \mathbf{1}_{d_K}], \tag{3}$$

which parameterizes rotary phases for Queries and Keys. This retains interpretable anchors without hand-tuned cycles.

Phase-Aware Attention Mechanism

With band-wise frequencies assigned to the per-head feature space, MP-RoPE rotates Queries and Keys using absolute timestamps. Let $\mathbf{q}^{(k)}, \mathbf{k}^{(k)} \in \mathbb{R}^{d_k}$ denote the k-th band sub-vectors of a head, where d_k is even after 2D pairing. For a calibrated frequency ω_k, we rotate every 2D pair by:

$$\mathbf{M}(\omega_k t) = \begin{pmatrix} \cos(\omega_k t) & -\sin(\omega_k t) \\ \sin(\omega_k t) & \cos(\omega_k t) \end{pmatrix}. \tag{4}$$

Let $\mathrm{RoT}_{\omega_k}(\cdot, t)$ denote this pairwise rotation. Then

$$\mathbf{q}_{\text{rot}}^{(k)} = \mathrm{RoT}_{\omega_k}(\mathbf{q}^{(k)}, t_q), \quad \mathbf{k}_{\text{rot}}^{(k)} = \mathrm{RoT}_{\omega_k}(\mathbf{k}^{(k)}, t_k). \tag{5}$$

Since the rotation is orthogonal, the bandwise inner product can be rewritten as a relative-phase interaction

$$(\mathbf{q}_{\text{rot}}^{(k)})^\top \mathbf{k}_{\text{rot}}^{(k)} = (\mathbf{q}^{(k)})^\top \mathbf{R}_k(t_k - t_q)\, \mathbf{k}^{(k)}, \tag{6}$$

where $\mathbf{R}_k(\Delta t) \in \mathbb{R}^{d_k \times d_k}$ is a block-diagonal matrix composed of repeated $\mathbf{M}(\omega_k \Delta t)$ blocks. Summing over all bands yields the attention score

$$\mathrm{Score}(t_q, t_k) = \sum_{k=1}^{K} (\mathbf{q}_{\text{rot}}^{(k)})^\top \mathbf{k}_{\text{rot}}^{(k)}, \tag{7}$$

followed by the standard scaling by $1/\sqrt{d_h}$ and Softmax.

Long-term Resonance Property

A direct implication of the relative-phase form in Eq. (7) is that the attention affinity is not necessarily monotonic in temporal distance, but periodic in Δt within each frequency band. For band k, the interaction is governed by $\mathbf{R}_k(\Delta t)$ whose 2D blocks are $\mathbf{M}(\omega_k \Delta t)$. When $\Delta t = n\tilde{T}_k$ with $\tilde{T}_k = 2\pi/\omega_k$ and $n \in \mathbb{Z}$, we have the resonance condition

$$\begin{aligned} \Delta t = n\tilde{T}_k\ (\tilde{T}_k = 2\pi/\omega_k) &\Rightarrow\ \omega_k \Delta t = 2\pi n \\ &\Rightarrow\ \mathbf{R}_k(\Delta t) = \mathbf{I}. \end{aligned} \tag{8}$$

Consequently, the k-th band reduces to a zero-distance alignment, making its contribution identical to that at $\Delta t = 0$. Aggregating multiple bands yields a resonance-like matching effect, enabling tokens with similar operational phase to maintain strong affinity even over long horizons and thereby mitigating the schedule-induced phase drift that index- or distance-driven encodings tend to accumulate.

3.4 The DR-PPO Mechanism

Building on the temporal anchoring of MP-RoPE, we design DR-PPO for active spatial focusing under deterministic forecasting (see Fig. 2). Industrial infrared anomalies are spatially sparse and prone to attenuation by background-dominated MSE regression. DR-PPO learns a dense spatial gate from uncertainty cues to modulate decoder representations, enhancing hotspot fidelity while preserving global stability without stochastic sampling.

Uncertainty-Aware State Space Construction

We construct state $S_t \in \mathbb{R}^{4 \times H \times W}$ to guide refinement:

$$S_t = \mathrm{concat}\big(V,\ R_t,\ V \odot R_t,\ V + R_t\big), \tag{9}$$

where V is a time-invariant variance prior from historical statistics and R_t is a residual-based uncertainty cue.

During training, $R_t = |Y_t - \hat{Y}_t^{(0)}|$ with $\hat{Y}_t^{(0)}$ the ungated backbone prediction. During inference, we use a deterministic self-consistency residual

$$R_t = \big|\hat{Y}_t^{(1)} - \hat{Y}_t^{(0)}\big|, \tag{10}$$

where $\hat{Y}_t^{(0)} = \mathcal{F}_\theta(\mathcal{X}_{in}, t_{abs}, d_{id})$, $\hat{Y}_t^{(1)} = \mathcal{F}_\theta(\tilde{\mathcal{X}}_{in}, t_{abs}, d_{id})$ are two forward passes of the same backbone, and $\tilde{\mathcal{X}}_{in} = \mathcal{A}(\mathcal{X}_{in})$ applies a fixed, reproducible patch-wise masking operator $\mathcal{A}$. The interaction terms emphasize high-V and high-R_t regions, yielding a compact difficulty-aware map for hotspot-focused gating.

Dense Actor-Critic Gating

Given the spatial state S_t, we train a dense actor-critic agent with PPO [Schulman *et al.*, 2017] to optimize hotspot- and trend-aware rewards and produce a spatial gate for decoder refinement. The agent encodes S_t with a lightweight CNN into shared features, followed by actor and critic heads. The actor outputs a bounded action map $A_{\mathrm{raw}} \in [-1, 1]^{H\times W}$ through a $\tanh$ activation, enabling localized and amplitude-controlled refinement. The critic applies global pooling to the shared features and predicts a scalar value $v_\phi(S_t)$, and we adopt GAE [Schulman *et al.*, 2016] to compute advantages for stable updates. During training, on-policy transition tuples are collected in a buffer and used to optimize the clipped actor-critic objective in DR-PPO. At inference, the policy is deterministic given S_t, with R_t from Eq. (10).

Instead of predicting an unconstrained gate, we adopt a residual gating formulation with an explicit baseline and bounded corrections. Specifically, we construct

$$G_t = \mathrm{Clamp}\Big(\mathbf{1} + \alpha\, A_{\mathrm{raw}} \odot \big(V + R_t\big),\ \epsilon_{\min},\ \epsilon_{\max}\Big), \tag{11}$$

where $\mathbf{1} \in \mathbb{R}^{H\times W}$ is all-ones. The term $\mathbf{1}$ provides a baseline pass-through path, while $\alpha\, A_{\mathrm{raw}} \odot (V + R_t)$ adds bounded, state-dependent corrections. The uncertainty weight $(V+R_t)$ serves as a prior-guided mask so that modulation concentrates on volatile or unreliable regions and remains small in low-variance, low-residual background areas, improving global structural stability. We set $\epsilon_{\min} > 0$ to keep the gate strictly nonnegative and use $\epsilon_{\max}$ to prevent over-amplification.

To apply G_t in the decoder, we align it to the patch-token grid via a patch-aligned mapping $\mathcal{D}(\cdot)$ implemented by average pooling over each patch cell. Let the decoder query tokens at time t be Q_t, we obtain the gated queries as

$$Q'_t = Q_t \odot \mathcal{D}(G_t), \tag{12}$$

so the decoder focuses correction on difficult regions while remaining deterministic.

Safety-Aware Multi-Objective Reward Design

We train the gating policy with a safety-aware reward balancing global accuracy and hotspot fidelity, used only in training where the ground truth is available. At each step t, we define the total reward as:

$$r_t = -\Big(\lambda_1 \mathcal{L}_{\mathrm{mse}}(t) + \lambda_2 \mathcal{L}_{\mathrm{top}\rho}(t) + \lambda_3 \mathcal{L}_{\mathrm{trend}}(t) + \lambda_4 \mathcal{L}_{\mathrm{reg}}(t)\Big). \tag{13}$$

Global reconstruction. We compute the global reconstruction error between the gated prediction $\hat{Y}_t$ and the ground truth Y_t as:

$$\mathcal{L}_{\mathrm{mse}}(t) = \frac{1}{HW}\sum_{i,j} \big\|Y_t^{(i,j)} - \hat{Y}_t^{(i,j)}\big\|_2^2. \tag{14}$$

Hotspot prioritization. Let $\Omega_{\mathrm{top}\rho}$ denote the top-ρ pixels of the variance prior V. We compute

$$\mathcal{L}_{\mathrm{top}\rho}(t) = \frac{1}{|\Omega_{\mathrm{top}\rho}|}\sum_{(i,j)\in\Omega_{\mathrm{top}\rho}} \big\|Y_t^{(i,j)} - \hat{Y}_t^{(i,j)}\big\|_2^2. \tag{15}$$

Trend consistency. We encourage consistent increments using a dead-zone sign $\mathrm{sgn}_\tau(x) = 0$ if $|x| < \tau$ and $\mathrm{sgn}_\tau(x) = \mathrm{sgn}(x)$ otherwise. Let $\Delta Y_t = Y_t - Y_{t-1}$ and $\Delta\hat{Y}_t = \hat{Y}_t - \hat{Y}_{t-1}$, then

$$\mathcal{L}_{\mathrm{trend}}(t) = \frac{1}{HW}\sum_{i,j}\frac{1}{2}\Big(1 - \mathrm{sgn}_\tau(\Delta Y_t^{(i,j)}) \cdot \mathrm{sgn}_\tau(\Delta\hat{Y}_t^{(i,j)})\Big). \tag{16}$$

Gate regularization. To avoid over-modulation and keep structure stable, we penalize deviations from the baseline:

$$\mathcal{L}_{\mathrm{reg}}(t) = \frac{1}{HW}\sum_{i,j} \big|G_t^{(i,j)} - \mathbf{1}^{(i,j)}\big|, \tag{17}$$

where $\mathbf{1} \in \mathbb{R}^{H\times W}$ is the all-ones tensor.

4 Experiments

4.1 Experimental Settings

Dataset and Preprocessing

As no public dataset is available for this task, we collected real-world infrared thermal sequences from control cabinets in three factories at 1 frame/min, covering N_D=7 device categories and $\sim 1.4 \times 10^5$ frames. To avoid temporal leakage, we split data chronologically per device into train/val/test with a 7:2:1 ratio, and generate all forecasting samples after splitting. We standardize each device category using training statistics only and report physical metrics in the original temperature domain after inverse transformation. We also compute a device-specific static variance prior $V \in \mathbb{R}^{H\times W}$ from the training split, which is reused for Dyn-MSE/Top-ρ evaluation and DR-PPO state construction. Detailed dataset statistics and access instructions are provided in Appendix A.

Implementation Details

We implement PhyRE-Net in PyTorch and train on 4×RTX 3090 GPUs for 80 epochs in two stages, with 10 epochs backbone pre-training followed by 70 epochs training with DR-PPO enabled. We use AdamW [Loshchilov and Hutter, 2019] (lr = 10^{-3}) for the backbone and Adam [Kingma and Ba, 2015] (lr = 3×10^{-4}) for DR-PPO, with cosine annealing on the backbone lr. We use a global batch size of 16 and set the clip ratio to $\epsilon_{\mathrm{clip}} = 0.2$ with GAE. Unless otherwise specified, we evaluate the long-gap protocol with $(T_{\mathrm{in}}, \mathrm{gap}, T_{\mathrm{out}}) = (4, 100, 20)$. We use a stage-wise curriculum on $\{\lambda_1, \ldots, \lambda_4\}$ that gradually increases the weight on Top-ρ fidelity and trend consistency. Detailed hyperparameter settings are provided in Appendix B.

Category	Methods	Year	Venue	Std-MSE ↓	Dyn-MSE ↓	Real-RMSE ↓	Top-ρ MSE ↓		
							$\rho = 0.1$	$\rho = 0.05$	$\rho = 0.01$
RNNs	ConvLSTM [Shi *et al.*, 2015]	2015	NeurIPS	0.1138	0.1645	1.4448	0.2856	0.3490	0.3982
	SAConvLSTM [Lin *et al.*, 2020]	2020	AAAI	0.0256	0.0389	0.6419	0.1083	0.1462	0.2193
	PredRNN v2 [Wang *et al.*, 2022]	2022	TPAMI	0.0543	0.0915	0.8366	0.1706	0.2568	0.3171
CNNs/ Transformers	SimVP [Gao *et al.*, 2022]	2022	CVPR	0.0315	0.0507	0.6616	0.0650	0.1038	0.1863
	PatchTST [Nie *et al.*, 2023]	2023	ICLR	0.0231	0.0303	0.6347	0.0488	0.0885	0.1431
	iTransformer [Liu *et al.*, 2024]	2024	ICLR	0.1351	0.1605	1.5472	0.1827	0.2462	0.3279
	TimeMixer [Wang *et al.*, 2024]	2024	ICLR	0.1073	0.1639	1.0919	0.1645	0.2165	0.3027
	TimeMixer++ [Wang *et al.*, 2025]	2025	ICLR	0.0237	0.0343	0.6461	0.0643	0.0874	0.1438
Generative	CrevNet [Yu *et al.*, 2020]	2020	ICLR	0.0314	0.0360	0.5891	0.0956	0.1411	0.2976
	DiT [Peebles and Xie, 2023]	2023	ICCV	0.1013	0.1640	0.7498	0.1564	0.2134	0.3108
	Diffusion-TS [Yuan and Qiao, 2024]	2024	ICLR	0.1495	0.1980	1.3635	0.1807	0.2671	0.3747
Ours	**PhyRE-Net**	–	–	**0.0179**	**0.0206**	**0.5422**	**0.0349**	**0.0544**	**0.1079**

Table 1: Quantitative comparison on long-gap forecasting ($T_{\text{in}} = 4, \text{gap} = 100, T_{\text{out}} = 20$) under the same device and timestamp context. Best results are shown in **bold**. Metrics other than Real-RMSE are computed in the standardized space.

Evaluation Metrics

We evaluate forecasting quality in terms of global accuracy, dynamic fidelity, and hotspot fidelity, computed per frame and averaged over the evaluation horizon.

Global and Physical Accuracy. We report **Std-MSE** in the standardized space to reflect optimization convergence, and **Real-RMSE** (in °C) in the original temperature domain for practical interpretability [Ioffe and Szegedy, 2015; Hyndman and Koehler, 2006]. In long-gap forecasting, phase drift appears as higher Std-MSE/Real-RMSE.

Dynamic Fidelity (Dyn-MSE). To reduce background dominance, we use the device-specific static variance prior $V \in \mathbb{R}^{H\times W}$ as a spatial weight. For a given time step t, the weighted error is defined as

$$\text{Dyn-MSE}_t = \frac{\sum_{i,j} V^{(i,j)} \big(Y^{(i,j)}_{t,\text{std}} - \hat{Y}^{(i,j)}_{t,\text{std}}\big)^2}{\sum_{i,j} V^{(i,j)}}. \tag{18}$$

Safety-Critical Hotspot Error (Top-ρ MSE). We focus on safety-relevant pixels with large intrinsic fluctuations. Let $\Omega_{\text{top}\rho} = \text{Top-}\rho(V)$ be the high-risk set with $\rho \in \{0.1, 0.05, 0.01\}$. The error on these hotspots is measured in the standardized space as

$$\text{Top-}\rho\ \text{MSE}_t = \frac{1}{|\Omega_{\text{top}\rho}|} \sum_{(i,j)\in\Omega_{\text{top}\rho}} \big(Y^{(i,j)}_{t,\text{std}} - \hat{Y}^{(i,j)}_{t,\text{std}}\big)^2. \tag{19}$$

This metric explicitly quantifies the model's reliability on sparse, safety-critical anomalies.

4.2 Main Results

Table 1 presents the comparison under the challenging long-gap setting. PhyRE-Net consistently outperforms all baselines across global and local metrics. Specifically, for global fidelity, it reduces the Std-MSE of the strongest baseline (PatchTST) by 22.5% and achieves the lowest Real-RMSE (0.5422), outperforming CrevNet by 8.0%. Regarding dynamics-aware reconstruction, PhyRE-Net surpasses PatchTST by 32.0% in Dyn-MSE. Crucially, for safety-critical hotspot monitoring, our method improves Top-ρ MSE (ρ=1%) by 24.5%. The varying best baselines highlight the complementary strengths of existing families, whereas PhyRE-Net achieves uniform dominance. Overall, the improvements align with the our designs. MP-RoPE anchors attention to physical time and learns calibrated multi-period phase alignment for long-gap stability, while DR-PPO concentrates deterministic decoder refinement on sparse hotspots to improve Top-ρ reliability without stochastic sampling.

Variant	Real-RMSE ↓	Dyn-MSE ↓	Top-ρ MSE ↓
Baseline	0.8310	0.0391	0.3084
+ MP-RoPE	0.5850	0.0287	0.2208
+ DR-PPO	0.6920	0.0251	0.1196
Ours	**0.5422**	**0.0206**	**0.1079**

Table 2: Component-wise ablations.

4.3 Ablation Study

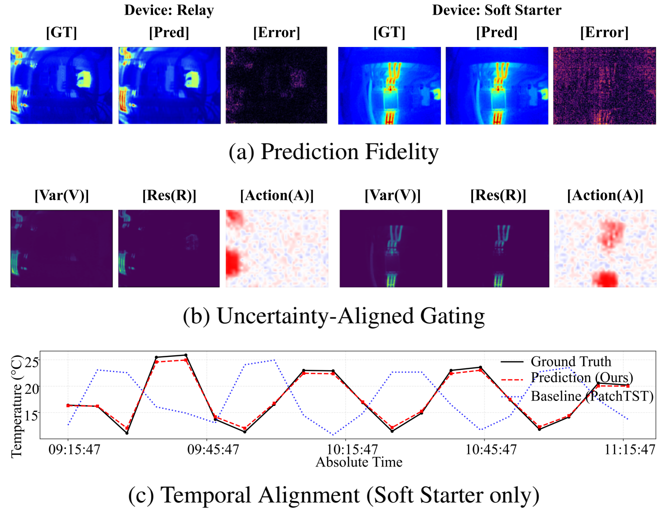
We conduct ablations under the same long-gap protocol $(T_{\text{in}}, \text{gap}, T_{\text{out}}) = (4, 100, 20)$. All metrics are computed per frame and averaged over the evaluation horizon, and Top-ρ uses ρ=1% unless otherwise stated.

Component ablation. Table 2 isolates the effects of the two core designs. MP-RoPE improves long-gap global fidelity, reducing Real-RMSE from 0.8310 to 0.5850 and lowering Dyn-MSE. In contrast, DR-PPO mainly improves hotspot fidelity, reducing Top-ρ MSE from 0.3084 to 0.1196 while keeping a higher Real-RMSE than MP-RoPE, consistent with targeted focusing. Combining both yields the best overall performance, validating that global alignment provides a stable context for the agent to perform precise local refinement.

Training and design ablation. Table 3 validates our optimization and design. For optimization, pre-training with fixed reward weights improves over joint training (Top-ρ 0.1398 vs. 0.1430), while curriculum further reduces it to 0.1079, showing the need to shift from global structure to local refinement. For state construction, our self-consistency residual matches the Oracle (GT) baseline closely (0.1079 ≈ 0.1065), validating it as a reliable proxy for difficulty during inference; in contrast, removing this cue (Variance-only)

Setting	Real-RMSE ↓	Dyn-MSE ↓	Top MSE ↓
Optimization / reward			
Joint-from-scratch	0.5890	0.0247	0.1430
Pre-train + fixed-λ	0.5610	0.0229	0.1398
w/o Top-ρ reward	0.5480	0.0210	0.1784
w/o MSE + reg	0.5560	0.0222	0.1190
Ours	**0.5422**	**0.0206**	**0.1079**
Residual cue in S_t			
Oracle residual	0.5408	0.0201	0.1065
Variance-only	0.5489	0.0212	0.1862
Ours	**0.5422**	**0.0206**	**0.1079**
Gating form			
Direct gate	0.5616	0.0251	0.1793
Differentiable Gate	0.5465	0.0218	0.1620
Ours	**0.5422**	**0.0206**	**0.1079**

Table 3: Training and design ablations. **Ours** denotes the default setting (pre-train + curriculum, self-consistency residual, and residual gating). "Oracle residual" uses GT and is analysis-only.



(a) Prediction Fidelity

(b) Uncertainty-Aligned Gating

(c) Temporal Alignment (Soft Starter only)

Figure 4: Qualitative analysis on heterogeneous devices. We visualize (a) spatial reconstruction quality, (b) internal difficulty-aware gating logic, and (c) temporal synchronization against baselines.

causes a sharp degradation (0.1862). Finally, regarding gating forms, Direct Gate predicts the spatial gate under supervised regression and struggles to converge (0.1793), while Differentiable Gate is trained through the same reconstruction objective and tends to favor smooth average correction (0.1620), confirming the necessity of our residual RL formulation for explicit hotspot prioritization.

4.4 Qualitative Results

Figure 4 visualizes prediction fidelity, gate localization, and long-gap temporal alignment on a Relay and a Soft Starter. Additional device visualizations are provided in Appendix C.

Prediction Fidelity. In Fig. 4a, PhyRE-Net preserves hotspot cores and sharp boundaries with negligible errors, overcoming baseline over-smoothing and validating gains in Top-ρ.

Uncertainty-Aligned Gating. Fig. 4b displays V, R^{vis} (GT for visualization), and A_{raw}. The learned action map explicitly highlights volatile regions while suppressing stable backgrounds, confirming a difficulty-aware focusing strategy.

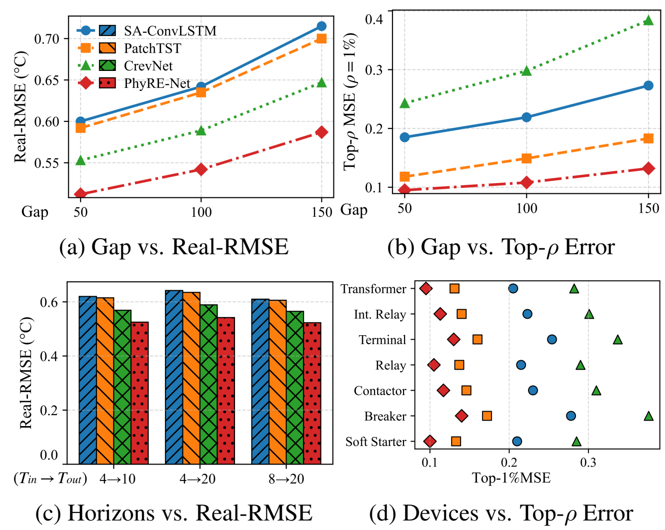
Temporal Alignment. In Fig. 4c, the baseline exhibits severe cumulative phase drift under the non-stationary schedule. In contrast, PhyRE-Net tightly locks onto the thermal rhythm, validating the absolute temporal anchoring of MP-RoPE.

4.5 Generalization across Deployment Shifts

We evaluate model performance under varying settings.

Sensitivity to Prediction Gap. In Fig. 5a and 5b, PhyRE-Net consistently achieves the lowest Real-RMSE and Top-ρ error, while exhibiting the smallest rise as the prediction gap expands (50→150), yielding significantly flatter curves that attest to superior long-gap stability.

Deployment Adaptability. Fig. 5c shows PhyRE-Net maintains the lowest Real-RMSE across varying $(T_{\text{in}}, T_{\text{out}})$ settings. Furthermore, in Fig. 5d, it consistently achieves the lowest Top-ρ error across all device categories, validating agile adaptation to diverse thermal signatures.



(a) Gap vs. Real-RMSE

(b) Gap vs. Top-ρ Error

(c) Horizons vs. Real-RMSE

(d) Devices vs. Top-ρ Error

Figure 5: Generalization under deployment shifts in gap (a,b), horizon (c), and device category (d). More details see Appendix C.

5 Conclusion

We proposed PhyRE-Net for industrial thermal forecasting under non-stationary schedules with long-tailed hotspot distributions. PhyRE-Net combines MP-RoPE, which anchors attention to physical time for long-gap temporal alignment, with DR-PPO, which enables deterministic active spatial gating to better preserve sparse, safety-critical hotspots. Experiments on real-world multi-device data demonstrate consistent gains across gaps, horizons, and device categories, with clear improvements on both global accuracy and Top-ρ hotspot metrics. Future work will explore multi-modal extension and domain generalization across larger-scale deployments.

References

[Gao *et al.*, 2022] Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z. Li. SimVP: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3170–3180, 2022.

[Goodfellow *et al.*, 2014] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

[Guen and Thome, 2020] Vincent Le Guen and Nicolas Thome. Disentangling physical dynamics from unknown factors for unsupervised video prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11474–11484, 2020.

[Ho *et al.*, 2022] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

[Hyndman and Koehler, 2006] Rob J. Hyndman and Anne B. Koehler. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688, 2006.

[Ioffe and Szegedy, 2015] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning*, 2015.

[Kazemi *et al.*, 2019] Seyed Mehran Kazemi, Rishab Goel, Sepehr Eghbali, Janahan Ramanan, Jaspreet Sahota, Sanjay Thakur, Stella Wu, Cathal Smyth, Pascal Poupart, and Marcus A. Brubaker. Time2Vec: Learning a vector representation of time. *CoRR*, abs/1907.05321, 2019.

[Kim *et al.*, 2021] Ju Sik Kim, Kyu Nam Choi, and Sung Woo Kang. Infrared thermal image-based sustainable fault detection for electrical facilities. *Sustainability*, 13(2):557, 2021.

[Kingma and Ba, 2015] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.

[Kingma and Dhariwal, 2018] Durk P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in Neural Information Processing Systems*, 31, 2018.

[Kingma and Welling, 2014] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations*, 2014.

[Lin *et al.*, 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[Lin *et al.*, 2020] Zhihui Lin, Maomao Li, Zhuobin Zheng, Yangyang Cheng, and Chun Yuan. Self-attention ConvLSTM for spatiotemporal prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11531–11538, 2020.

[Lipton *et al.*, 2015] Zachary C. Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*, 2015.

[Liu *et al.*, 2024] Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.

[Loshchilov and Hutter, 2019] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations*, 2019.

[Mathieu *et al.*, 2016] Michaël Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. In *Proceedings of the International Conference on Learning Representations*, 2016.

[Nie *et al.*, 2023] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *Proceedings of the Eleventh International Conference on Learning Representations*, 2023.

[Peebles and Xie, 2023] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[Press *et al.*, 2022] Ofir Press, Noah A. Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. In *Proceedings of the International Conference on Learning Representations*, 2022.

[Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015.

[Schulman *et al.*, 2016] John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *Proceedings of the International Conference on Learning Representations*, 2016.

[Schulman *et al.*, 2017] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

[Shi *et al.*, 2015] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang chun Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems*, 28, 2015.

[Shi *et al.*, 2017] Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang chun Woo. Deep learning for precipitation nowcasting: A benchmark and a new model. *Advances in Neural Information Processing Systems*, 30, 2017.

[Su *et al.*, 2024] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[Tang *et al.*, 2023] Song Tang, Chuang Li, Pu Zhang, and RongNian Tang. SwinLSTM: Improving spatiotemporal prediction accuracy using swin transformer and LSTM. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13470–13479, 2023.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[Voleti *et al.*, 2022] Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. MCVD: Masked conditional video diffusion for prediction, generation, and interpolation. *Advances in Neural Information Processing Systems*, 35:23371–23385, 2022.

[Wang *et al.*, 2017] Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and Philip S. Yu. PredRNN: Recurrent neural networks for predictive learning using spatiotemporal LSTMs. *Advances in Neural Information Processing Systems*, 30, 2017.

[Wang *et al.*, 2018a] Yunbo Wang, Zhifeng Gao, Mingsheng Long, Jianmin Wang, and Philip S. Yu. PredRNN++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In *Proceedings of the International Conference on Machine Learning*, pages 5123–5132, 2018.

[Wang *et al.*, 2018b] Yunbo Wang, Lu Jiang, Ming-Hsuan Yang, Li-Jia Li, Mingsheng Long, and Li Fei-Fei. Eidetic 3D LSTM: A model for video prediction and beyond. In *Proceedings of the International Conference on Learning Representations*, 2018.

[Wang *et al.*, 2019] Yunbo Wang, Jianjin Zhang, Hongyu Zhu, Mingsheng Long, Jianmin Wang, and Philip S. Yu. Memory in memory: A predictive neural network for learning higher-order non-stationarity from spatiotemporal dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9154–9162, 2019.

[Wang *et al.*, 2022] Yunbo Wang, Haixu Wu, Jianjin Zhang, Zhifeng Gao, Jianmin Wang, Philip S. Yu, and Mingsheng Long. PredRNN: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2208–2225, 2022.

[Wang *et al.*, 2024] Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.

[Wang *et al.*, 2025] Shiyu Wang, Jiawei Li, Xiaoming Shi, Zhou Ye, Baichuan Mo, Wenze Lin, Shengtong Ju, Zhixuan Chu, and Ming Jin. TimeMixer++: A general time series pattern machine for universal predictive analysis. In *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025.

[Wu *et al.*, 2021] Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems*, 34:22419–22430, 2021.

[Wu, 2017] Jianxin Wu. Introduction to convolutional neural networks. *National Key Lab for Novel Software Technology, Nanjing University, China*, 5(23):495, 2017.

[Yu *et al.*, 2020] Wei Yu, Yichao Lu, Steve Easterbrook, and Sanja Fidler. Efficient and information-preserving future frame prediction and beyond. In *Proceedings of the International Conference on Learning Representations*, 2020.

[Yuan and Qiao, 2024] Xinyu Yuan and Yan Qiao. Diffusion-TS: Interpretable diffusion for general time series generation. In *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.

[Zhou *et al.*, 2021] Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11106–11115, 2021.

[Zhou *et al.*, 2022] Tian Zhou, Ziqing Ma, Qingsong Wen, Xue Wang, Liang Sun, and Rong Jin. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the International Conference on Machine Learning*, pages 27268–27286, 2022.