

$S^2\text{-KD}$: Semantic-Spectral Knowledge Distillation Spatiotemporal Forecasting

Wenshuo Wang¹, Yaomin Shen², Yingjie Tan³, Yihao Chen^{4*}

¹School of Future Technology, South China University of Technology, Guangzhou, China

²Nanchang Research Institute, Zhejiang University, Nanchang, China

³School of Software, Beihang University, Beijing, China

⁴College of Control Science and Engineering, Zhejiang University, Hangzhou, China

202364870251@mail.scut.edu.cn, coolshennf@gmail.com, tanyingjie@buaa.edu.cn, yihaochen@zju.edu.cn

Abstract

Spatiotemporal forecasting often relies on computationally intensive models to capture complex dynamics. Knowledge distillation (KD) has emerged as a key technique for creating lightweight student models, with recent advances like frequency-aware KD successfully preserving spectral properties (i.e., high-frequency details and low-frequency trends). However, these methods are fundamentally constrained by operating on pixel-level signals, leaving them blind to the rich semantic and causal context behind the visual patterns. To overcome this limitation, we introduce $S^2\text{-KD}$, a novel framework that unifies Semantic priors with Spectral representations for distillation. Our approach begins by training a privileged, multimodal **teacher** model. This teacher leverages textual narratives from a Large Multimodal Model (LMM) to reason about the underlying causes of events, while its architecture simultaneously decouples spectral components in its latent space. The core of our framework is a new distillation objective that transfers this unified semantic-spectral knowledge into a lightweight, **vision-only student**. Consequently, the student learns to make predictions that are not only spectrally accurate but also semantically coherent, without requiring any textual input or architectural overhead at inference. Extensive experiments on benchmarks like WeatherBench and TaxiBJ+ show that $S^2\text{-KD}$ significantly boosts the performance of simple student models, enabling them to outperform state-of-the-art methods, particularly in long-horizon and complex non-stationary scenarios.

Introduction

Spatiotemporal forecasting, which aims to predict future states from historical data sequences (Wu et al. 2024a,d; Gao et al. 2022; Shi et al. 2015), is a cornerstone of decision-making in domains ranging from climate science and meteorology to urban traffic management and autonomous navigation (Wu et al. 2025b; Wang et al. 2020; Gao et al. 2025). The fundamental challenge lies in capturing the intricate coupling between high-frequency (Wu et al. 2025a; Bruna et al. 2013), localized variations (e.g., sudden traffic congestion, turbulent eddies) and low-frequency (Li et al. 2020), global trends (e.g., diurnal traffic patterns, seasonal

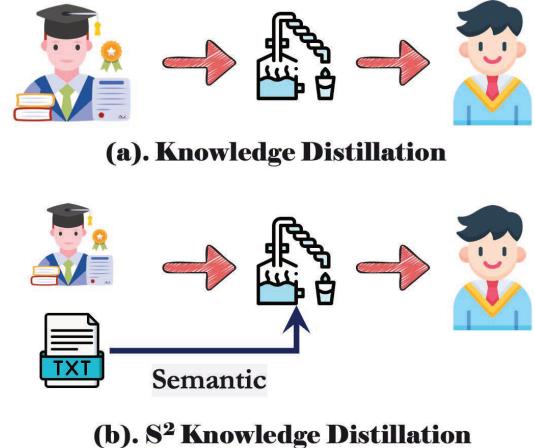


Figure 1: A comparison between traditional Knowledge Distillation and our $S^2\text{-KD}$ framework. (a) Traditional KD directly distills knowledge from a teacher to a student model. (b) Our $S^2\text{-KD}$ framework enriches the distillation process by injecting semantic priors extracted from text, enabling the teacher to transfer a deeper, causal understanding.

climate shifts). This has spurred the development of powerful but computationally expensive models, such as complex CNN-Transformer hybrids, whose demanding resource requirements hinder their deployment in real-world, resource-constrained environments (Wu et al. 2024c).

To address this efficiency-accuracy trade-off, knowledge distillation (KD) (Gou et al. 2021; Park et al. 2019) has emerged as a powerful paradigm for compressing large models into lightweight, efficient students (Chen and Wang 2024; Chen et al. 2022; Wang, Han, and Chen 2025). Recent advances, such as the frequency-aware distillation framework, have made significant strides by ensuring the student model preserves the spectral fidelity both high and low-frequency components of its powerful teacher. **However, these methods, despite their spectral sophistication, are fundamentally operating in a semantic vacuum.** They are adept at mimicking *what* patterns occur, but remain blind to *why* they occur. For instance, they can learn the visual signature of a traffic jam, but cannot distinguish whether it is caused by predictable rush-hour volume or an unpredictable

*Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

traffic accident a distinction crucial for accurate forecasting.

This reveals a critical gap in the literature: the absence of causal and semantic reasoning in current spatiotemporal distillation frameworks. *Our key insight is that natural language can serve as a powerful bridge to this missing semantic layer.* By leveraging large multimodal models (LMM) to generate descriptive narratives of spatiotemporal scenes (e.g., *a strong cold front is approaching, causing a sudden drop in temperature*), we can provide the model with contextual and causal information at the highest level that is simply unavailable in raw pixel data. This linguistic knowledge acts as ‘privileged information’ during training, allowing a teacher model to develop a much deeper and more robust understanding of the underlying physical processes.

To this end, we propose **$S^2\text{-KD}$: Semantic-Spectral Knowledge Distillation**, a novel framework designed to distill both causal understanding and spectral characteristics. $S^2\text{-KD}$ first trains a powerful, multimodal **teacher** that jointly reasons over visual inputs and their corresponding textual narratives. This teacher is architected to be both **semantically-aware** and **spectrally-decoupled**. Subsequently, our tailored distillation process transfers this unified knowledge into a lightweight, **vision-only student**, empowering it to make semantically coherent predictions without needing any text at inference time. The student, therefore, learns to implicitly reason about causes and effects, guided by the teacher’s richer, multimodal wisdom (as illustrated in Figure 1). Our contributions are summarized as follows:

- **A New Paradigm:** We are the first to propose a paradigm for spatiotemporal distillation that enriches representations with semantic knowledge extracted from language, moving beyond pixel-level pattern imitation to a more causal understanding.
- **A Novel Framework:** We design and implement $S^2\text{-KD}$, a concrete framework featuring a multimodal teacher that fuses visual and linguistic information, and a tailored semantic-spectral distillation objective to transfer this unified knowledge to a unimodal student.
- **State-of-the-Art Performance:** We conduct extensive experiments on multiple benchmarks, demonstrating that $S^2\text{-KD}$ significantly boosts the performance of simple, lightweight models, enabling them to achieve new state-of-the-art results for efficient spatiotemporal forecasting.

Related Work

Spatiotemporal Forecasting Models aim to capture the dynamics of complex systems. Early deep learning methods, such as ConvLSTM (Shi et al. 2015) and PredRNN (Wang et al. 2022), pioneer the combination of convolution and recurrent networks to model local spatiotemporal correlations. However, they exhibit limitations in handling long-range dependencies and global dynamics (Fan et al. 2020; Fahlman and Fernández 2022; Sorjamaa et al. 2007). To overcome these issues, modern research shifts towards powerful CNN-Transformer hybrid architectures (Chen et al. 2023b; Wu et al. 2024a; Chen et al. 2023a; Bi et al. 2023; Wu et al. 2025b). These models achieve state-of-the-art (SOTA) prediction accuracy on various benchmarks by integrating

the local receptive fields of CNNs with the global modeling capabilities of Transformers. Nevertheless, this superior performance comes at the cost of immense computational and memory overhead. The self-attention mechanism in Transformers introduces quadratic complexity with respect to sequence length (Wu et al. 2025b; Kurth et al. 2023; Guibas et al. 2021), while deep stacks of convolutions also contribute a large number of parameters. This high cost severely hinders their deployment in resource-constrained real-world scenarios, such as autonomous vehicles and edge-based weather stations, thereby highlighting the urgent need for lightweight models.

Knowledge distillation (KD) (Phuong and Lampert 2019; Mirzadeh et al. 2020; Stanton et al. 2021) offers an effective pathway to address the aforementioned model complexity. Classic KD, pioneered by Hinton (Hinton, Vinyals, and Dean 2015), transfers knowledge via soft labels (Zhang et al. 2021), while subsequent methods like FitNets focus on matching intermediate features (Murata et al. 2023). In the spatiotemporal domain, KD has also seen significant progress. In particular, frequency-aware distillation frameworks like SDKD represent a major advancement by preserving the spectral fidelity of the teacher model, including both high and low-frequency components. *However, a fundamental limitation underlies all existing KD approaches, from classic to frequency-aware: they operate exclusively within a unimodal paradigm.* This means they can only distill the visual patterns the teacher model ‘sees’ (*what*) but fail to transfer the underlying causal understanding of these patterns (*why*). They distill ‘appearance’ rather than ‘comprehension’, making the student model vulnerable to dynamic changes caused by unseen factors (e.g., sudden events), even if it is spectrally aligned with the teacher.

Multimodal Learning and Privileged Information. To break the semantic bottleneck of unimodal distillation, our work draws inspiration from multimodal learning and the concept of privileged information (Ramachandram and Taylor 2017; Blikstein 2013). Recent breakthroughs in Large Multimodal Models (LMMs), such as CLIP (Radford et al. 2021) and LLaVA (Liu et al. 2023), successfully bridge the gap between vision and language, enabling the generation of high-level, logically coherent textual descriptions for spatiotemporal scenes. Natural language not only provides a holistic summary of a scene but, more importantly, contains rich causal relationships, object attributes, and common-sense knowledge that is difficult to extract from raw pixel data alone. We frame our approach within Vapnik’s paradigm of Learning Using Privileged Information (LUPI) (Pechyony and Vapnik 2010). In this framework, the textual narratives generated by LMMs serve as the ‘privileged information’, which is available only during the training phase to help the teacher model build a profound understanding of the dynamics. The $S^2\text{-KD}$ framework, therefore, essentially distills this deep understanding gained from privileged information and internalizes it within the parameters of a vision-only student model through a novel semantic-spectral distillation process. Unlike previous works that focus exclusively on knowledge transfer within a single modality, our work is the first to explore how to

distill cross-modal semantic knowledge from a multimodal privileged teacher to a unimodal student, opening up a new avenue for building more intelligent and robust lightweight forecasting models.

Methodology

Problem Formulation

The primary goal of spatiotemporal forecasting is to predict a sequence of future states given a history of observations. Let $\mathcal{X} = \{\mathbf{X}_t \in \mathbb{R}^{H \times W \times C}\}_{t=1}^{T_{in}}$ represent the historical sequence of spatiotemporal data, where H and W are the spatial dimensions, C is the number of channels, and T_{in} is the length of the input sequence. The objective is to learn a mapping function \mathcal{F} that predicts the future sequence $\mathcal{Y} = \{\mathbf{Y}_{t'} \in \mathbb{R}^{H \times W \times C}\}_{t'=1}^{T_{out}}$, where T_{out} is the prediction horizon. A conventional forecasting model is trained by minimizing an objective function, typically the Mean Squared Error (MSE), between the predictions and the ground truth:

$$\mathcal{L}_{pred} = \mathbb{E}_{(\mathcal{X}, \mathcal{Y})} \|\mathcal{F}(\mathcal{X}; \theta_F) - \mathcal{Y}\|_2^2, \quad (1)$$

where θ_F are the parameters of the model \mathcal{F} . Our work extends this formulation by introducing a knowledge distillation framework. We first train a powerful multimodal teacher model \mathcal{T} , which leverages both the visual sequence \mathcal{X} and a corresponding textual description \mathcal{S} generated by a Large Multimodal Model (LMM). This teacher is optimized to produce highly accurate predictions. Subsequently, we aim to train a lightweight, vision-only student model $\mathcal{G}(\mathcal{X}; \theta_G)$ to mimic the teacher's behavior. The student's training objective is a composite loss that includes not only the prediction loss but also a distillation loss, which transfers the semantic-spectral knowledge from the teacher:

$$\min_{\theta_G} \mathcal{L}_{pred}(\mathcal{G}(\mathcal{X}), \mathcal{Y}) + \lambda \mathcal{L}_{distill}(\mathcal{G}(\mathcal{X}), \mathcal{T}(\mathcal{X}, \mathcal{S})), \quad (2)$$

where $\mathcal{L}_{distill}$ measures the discrepancy between the student's and teacher's internal representations, and λ is a hyperparameter balancing the two loss terms. The ultimate goal is to obtain an efficient student model \mathcal{G} that achieves performance comparable to the privileged teacher \mathcal{T} at inference time without requiring the textual input \mathcal{S} .

Overall Architecture of S²-KD

Our S²-KD framework, illustrated in Figure 2, comprises two stages: training a privileged multimodal teacher for knowledge distillation, and deploying a lightweight student for efficient inference. During training, the teacher model \mathcal{T} processes both the visual sequence \mathcal{X} and a textual narrative \mathcal{S} from a Large Multimodal Model (LMM). An internal *Alignment Module* fuses these inputs into a unified, semantically-rich representation, guided by the ground truth \mathcal{Y} . Instead of distilling from final predictions, we extract knowledge directly from this fused intermediate representation. We define this as *semantic-spectral knowledge*, as it combines causal semantics from language with the spectral properties of the visual dynamics. Subsequently, a lightweight student model \mathcal{G} , which only sees \mathcal{X} , is trained to mimic this knowledge via the objective in Equation 2.

This distillation process internalizes the teacher's multimodal reasoning into the student's parameters. As a result, the deployed student performs efficient inference on visual data alone, yet retains the teacher's semantic understanding. *The design of this framework ensures that the complexity of multimodal reasoning is confined to the offline training stage, guaranteeing high efficiency for online deployment.*

Multimodal Privileged Teacher

Our privileged teacher model, \mathcal{T} , is architected to effectively fuse spatiotemporal visual patterns with high-level semantic narratives. It comprises a visual encoder \mathcal{E}_v , a text encoder \mathcal{E}_s , a cross-modal alignment module, and a predictive decoder \mathcal{D}_v . The visual encoder, \mathcal{E}_v , first maps the input visual sequence \mathcal{X} into a sequence of latent feature vectors:

$$\mathbf{Z}_v = \mathcal{E}_v(\mathcal{X}), \quad \text{where } \mathbf{Z}_v \in \mathbb{R}^{L_v \times D}. \quad (3)$$

Simultaneously, the text encoder \mathcal{E}_s , a pre-trained Transformer, processes the textual description \mathcal{S} to produce a sequence of semantic embeddings:

$$\mathbf{Z}_s = \mathcal{E}_s(\mathcal{S}), \quad \text{where } \mathbf{Z}_s \in \mathbb{R}^{L_s \times D}. \quad (4)$$

The core of our teacher is the *Cross-Modal Alignment Module*, which facilitates deep interaction between these two modalities using a stack of N cross-attention layers. For each layer $i \in \{1, \dots, N\}$, we first compute the query (\mathbf{Q}), key (\mathbf{K}), and value (\mathbf{V}) matrices from the visual and text features of the previous layer, $\mathbf{Z}_v^{(i-1)}$ and $\mathbf{Z}_s^{(i-1)}$ (with $\mathbf{Z}_v^{(0)} = \mathbf{Z}_v$, $\mathbf{Z}_s^{(0)} = \mathbf{Z}_s$), using learnable projection matrices:

$$\mathbf{Q}^{(i)} = \mathbf{Z}_v^{(i-1)} \mathbf{W}_Q^{(i)}, \quad (5)$$

$$\mathbf{K}^{(i)} = \mathbf{Z}_s^{(i-1)} \mathbf{W}_K^{(i)}, \quad (6)$$

$$\mathbf{V}^{(i)} = \mathbf{Z}_s^{(i-1)} \mathbf{W}_V^{(i)}, \quad (7)$$

where $\mathbf{W}_Q^{(i)}, \mathbf{W}_K^{(i)}, \mathbf{W}_V^{(i)} \in \mathbb{R}^{D \times D}$ are the projection parameters. The semantically-enhanced features are then computed via multi-head attention (MHA):

$$\text{Attn}(\mathbf{Q}^{(i)}, \mathbf{K}^{(i)}, \mathbf{V}^{(i)}) = \text{Softmax} \left(\frac{(\mathbf{Q}^{(i)} \mathbf{K}^{(i)T})}{\sqrt{d_k}} \right) \mathbf{V}^{(i)}. \quad (8)$$

This operation is followed by a residual connection and layer normalization, forming a complete cross-attention block. The output of the i -th block, $\mathbf{Z}_v^{(i)}$, is then fed into the next. After N such blocks, we obtain the final fused representation $\mathbf{Z}_{\text{fused}} = \mathbf{Z}_v^{(N)}$, which serves as the source for distillation:

$$\mathbf{Z}_{\text{fused}} = \text{LayerNorm} \left(\mathbf{Z}_v^{(i-1)} + \text{MHA}(\mathbf{Q}^{(i)}, \mathbf{K}^{(i)}, \mathbf{V}^{(i)}) \right). \quad (9)$$

Finally, the predictive decoder \mathcal{D}_v takes this fused representation $\mathbf{Z}_{\text{fused}}$ and reconstructs the future spatiotemporal sequence $\hat{\mathcal{Y}} = \mathcal{D}_v(\mathbf{Z}_{\text{fused}})$.

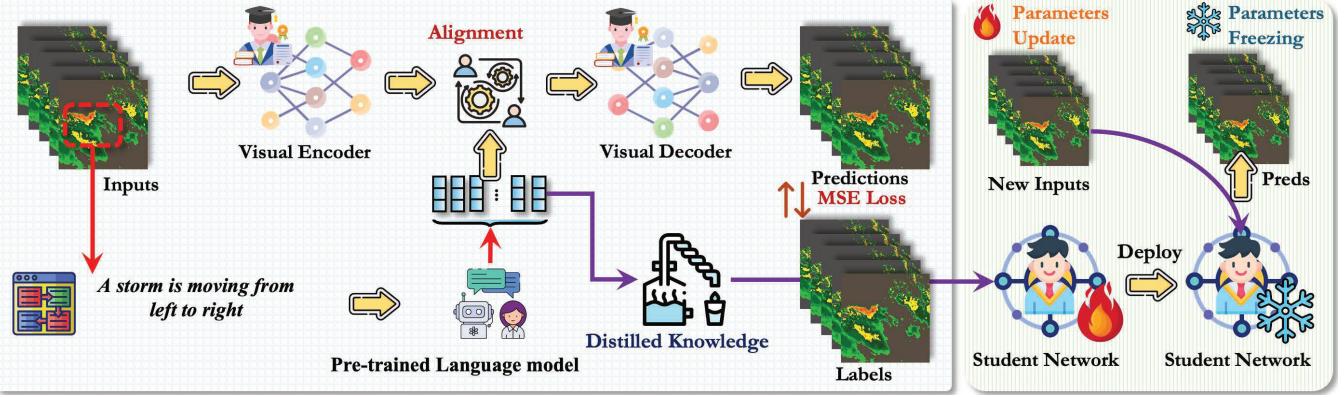


Figure 2: An overview of our proposed S²-KD framework.

Semantic-Spectral Distillation Loss

Having defined the powerful multimodal teacher, the next crucial step is to design a distillation loss, $\mathcal{L}_{distill}$, that effectively transfers its rich, unified knowledge to the lightweight student \mathcal{G} . Our objective is to formulate a loss that preserves both the high-level semantic understanding and the fine-grained spectral characteristics captured in the teacher's fused representation, $\mathbf{Z}_{\text{fused}}^T$, from Equation 9. To achieve this, we design a composite loss consisting of two complementary components: a semantic alignment loss and a spectral alignment loss.

First, we align the intermediate representations of the student and teacher. The student model \mathcal{G} employs a visual encoder $\mathcal{E}_v^{\mathcal{G}}$ to generate its own latent representation $\mathbf{Z}_v^{\mathcal{G}} = \mathcal{E}_v^{\mathcal{G}}(\mathcal{X})$. To ensure dimensional compatibility with the teacher's representation, we apply a linear projection layer P . The core of our distillation is to minimize the discrepancy between the student's projected representation, $P(\mathbf{Z}_v^{\mathcal{G}})$, and the teacher's fused representation, $\mathbf{Z}_{\text{fused}}^T$.

The first component is the **semantic alignment loss**, $\mathcal{L}_{semantic}$, which enforces the student to capture the high-level semantic structure of the teacher's representation. We employ the Mean Squared Error (MSE) for this purpose, as it effectively matches the overall feature distributions:

$$\mathcal{L}_{semantic} = \|P(\mathbf{Z}_v^{\mathcal{G}}) - \mathbf{Z}_{\text{fused}}^T\|_2^2. \quad (10)$$

This loss compels the student to reconstruct the teacher's semantically-informed "thought process," thereby implicitly internalizing the knowledge gained from language.

The second component is the **Spectral Alignment Loss**, $\mathcal{L}_{spectral}$, which inherits the core idea from frequency-aware distillation to preserve the modeling of both high-frequency details and low-frequency trends. We apply the Fast Fourier Transform (\mathcal{F}) to map the features into the frequency domain and compute the L1 loss between their spectral magnitudes:

$$\mathcal{L}_{spectral} = \|\mathcal{F}(P(\mathbf{Z}_v^{\mathcal{G}})) - \mathcal{F}(\mathbf{Z}_{\text{fused}}^T)\|_1. \quad (11)$$

By directly aligning the spectral representations, we explicitly guide the student to learn the same frequency response

as the teacher. Finally, our total distillation loss is a weighted sum of these two components:

$$\mathcal{L}_{distill} = \mathcal{L}_{semantic} + \beta \mathcal{L}_{spectral}, \quad (12)$$

where β is a hyperparameter balancing the importance of semantic and spectral alignment. This composite loss ensures a comprehensive knowledge transfer, encompassing both macroscopic semantic understanding and microscopic dynamic details.

Final Objective and Training Procedure

The training of our S²-KD framework is conducted in two sequential stages: first, pre-training the multimodal privileged teacher, and second, training the lightweight student via our proposed semantic-spectral distillation.

Stage 1: Teacher Model Training. In the first stage, we train the multimodal teacher model \mathcal{T} to learn an effective mapping from the visual sequence \mathcal{X} and the textual narrative \mathcal{S} to the future sequence \mathcal{Y} . The teacher is optimized solely based on the standard predictive loss, which is the Mean Squared Error (MSE) between its predictions and the ground truth:

$$\mathcal{L}_{Teacher} = \mathbb{E}_{(\mathcal{X}, \mathcal{S}, \mathcal{Y})} \|\mathcal{T}(\mathcal{X}, \mathcal{S}; \theta_T) - \mathcal{Y}\|_2^2. \quad (13)$$

Through this process, the teacher learns to leverage semantic information to form a high-quality internal representation $\mathbf{Z}_{\text{fused}}^T$. Upon completion of this stage, the parameters θ_T of the teacher model are frozen.

Stage 2: Student Model Distillation. In the second stage, we train the lightweight, vision-only student model \mathcal{G} . The student is optimized using a composite objective function that combines the predictive loss with our semantic-spectral distillation loss $\mathcal{L}_{distill}$. The final objective for the student model is:

$$\min_{\theta_G} \mathcal{L}_{Student} = \mathcal{L}_{pred}(\mathcal{G}(\mathcal{X}), \mathcal{Y}) + \lambda \mathcal{L}_{distill}, \quad (14)$$

where \mathcal{L}_{pred} is the standard MSE loss for the student's predictions, and $\mathcal{L}_{distill}$ is defined as $\mathcal{L}_{semantic} + \beta \mathcal{L}_{spectral}$. The hyperparameters λ and β balance the contributions of

the predictive task, semantic alignment, and spectral alignment. During this stage, the teacher model operates in evaluation mode solely to provide the target representation $\mathbf{Z}_{\text{fused}}^T$, and no gradients are backpropagated through it. Upon completion of this two-stage procedure, we obtain a lightweight and efficient student model \mathcal{G} that inherits the advanced reasoning capabilities of the privileged teacher, ready for deployment.

Experiment

We aim to answer three key research questions:

- **RQ1.** How does S^2 -KD perform against state-of-the-art forecasting and knowledge distillation methods?
- **RQ2.** What are the individual contributions of the semantic and spectral distillation components?
- **RQ3.** How effective is S^2 -KD in predicting high-impact extreme events?

Experimental Setup

Datasets. We evaluate S^2 -KD on three benchmarks to assess its performance and generalization capabilities. ①. **Prometheus** is a large-scale fire simulation dataset designed for out-of-distribution (OOD) fluid dynamics. It includes two scenarios, Tunnel Fire (Prometheus-T) (Wu et al. 2024b) and Pool Fire (Prometheus-P), with distribution shifts created by varying physical parameters like heat release rate. Models are trained on seen environments and tested on unseen ones to evaluate OOD generalization. ②.

WeatherBench (ERA5) (Rasp et al. 2020) is a scientific benchmark for global weather forecasting. We use Geopotential (Z500) and Temperature (T850) variables to capture large-scale, slowly-varying atmospheric dynamics. ③. **TaxiBJ+** (Wu et al. 2023) is an urban traffic flow dataset from Beijing. It represents a highly non-stationary system, challenging models to capture both periodic patterns and stochastic events.

Model Selection. To demonstrate the versatility and effectiveness of our S^2 -KD framework, we adopt a domain-specific teacher and general-purpose student strategy. For each benchmark, we select a powerful, state-of-the-art model from its respective domain as the **teacher**: **Triton** (Wu et al. 2025b) for weather forecasting on WeatherBench, **EarthFarseer** (Wu et al. 2023) for urban dynamics on TaxiBJ+, and a deep variant of **SimVP** (Gao et al. 2022) for fluid dynamics on Prometheus. This ensures that the distilled knowledge originates from a top-performing specialist. For semantic extraction, we primarily use the open-source **DeepSeek-VL** (Lu et al. 2024), with other LLMs explored in our ablation studies. For the **student**, we employ a diverse set of lightweight architectures, including **U-Net** (Ronneberger, Fischer, and Brox 2015), a shallow **ResNet** (He et al. 2016), and an **MLP-Mixer** (Tolstikhin et al. 2021), to validate that the benefits of S^2 -KD are architecture-agnostic. As for distillation baselines, we compare against the classic **Standard KD** (Hinton, Vinyals, and Dean 2015) and feature-based **FitNet** (Chen et al. 2021).

Implementation Details. All experiments are conducted on a server equipped with four NVIDIA A100 (80GB) GPUs, using PyTorch 2.1 and CUDA 12.0. We employ the Adam optimizer (Kingma and Ba 2014) for all model training. For the teacher models, we largely follow the optimal hyperparameter configurations reported in their original papers. For student model training via distillation, we set an initial learning rate of 1×10^{-3} , which is reduced by a factor of 10 if the validation loss plateaus for 5 consecutive epochs. The batch size is set to 16. All models are trained for a maximum of 100 epochs with an early stopping mechanism based on the validation set performance to prevent overfitting. For our S^2 -KD framework, the distillation loss weight λ is set to 1.0, and the spectral alignment weight β is set to 0.5, determined via a grid search on a validation subset. For the LMM-based semantic extraction, we generate a single descriptive caption for each input sequence using a standardized prompt template. Unless otherwise specified, DeepSeek-VL is used as the default LLM. To ensure reproducibility, we set the global random seed to 42 for all experiments.

Model	Method	Params (M) ↓	Latency (ms) ↓	MSE ↓	MAE ↓	SSIM ↑
Teacher (Triton)	-	150.2	85.6	0.0683	0.7287	0.9493
U-Net	Baseline	5.1	12.3	0.0831	0.9822	0.8635
	+ FitNet	5.1	12.3	0.0765	0.9150	0.8712
	+ S^2 -KD (Ours)	5.1	12.3	0.0698	0.8104	0.9012
ResNet	Baseline	4.5	10.8	0.0876	1.0210	0.8590
	+ FitNet	4.5	10.8	0.0801	0.9433	0.8705
	+ S^2 -KD (Ours)	4.5	10.8	0.0753	0.8819	0.8831
MLP-Mixer	Baseline	6.2	15.1	0.0953	1.1527	0.8421
	+ FitNet	6.2	15.1	0.0925	1.1098	0.8490
	+ S^2 -KD (Ours)	6.2	15.1	0.0895	1.0436	0.8615

Table 1: Performance on **WeatherBench**. Teacher model (Triton) results are provided as an upper bound. Our S^2 -KD consistently enables lightweight students to approach the teacher’s performance more closely than other methods, including the classic feature-based baseline (FitNet).

Model	Method	Params (M) ↓	Latency (ms) ↓	MSE ↓	MAE ↓	SSIM ↑
Teacher (EarthFarseer)	-	125.8	72.4	0.1172	0.9701	0.9810
U-Net	Baseline	5.1	12.3	0.1354	1.1032	0.9532
	+ FitNet	5.1	12.3	0.1298	1.0760	0.9580
	+ S^2 -KD (Ours)	5.1	12.3	0.1180	0.9855	0.9754
ResNet	Baseline	4.5	10.8	0.1402	1.1450	0.9499
	+ FitNet	4.5	10.8	0.1331	1.1027	0.9556
	+ S^2 -KD (Ours)	4.5	10.8	0.1231	1.0189	0.9678
MLP-Mixer	Baseline	6.2	15.1	0.1520	1.2345	0.9380
	+ FitNet	6.2	15.1	0.1485	1.2010	0.9413
	+ S^2 -KD (Ours)	6.2	15.1	0.1399	1.1401	0.9523

Table 2: Performance on **TaxiBJ+**. For this non-stationary urban dynamics task, our S^2 -KD not only surpasses other methods but also brings the lightweight students remarkably close to the performance of the large EarthFarseer teacher model.

Main results (RQ1.)

This section addresses our first research question (RQ1): *How does S^2 -KD perform compared to the baseline and*

Model	Method	Params (M) ↓	Latency (ms) ↓	MSE ↓	MAE ↓	SSIM ↑
Teacher (SimVP-Deep)	-	80.5	55.1	0.0210	0.1805	0.8521
U-Net	Baseline	5.1	12.3	0.0295	0.2510	0.7811
	+ FitNet	5.1	12.3	0.0268	0.2243	0.8034
	+ $S^2\text{-KD}$ (Ours)	5.1	12.3	0.0219	0.1895	0.8416
ResNet	Baseline	4.5	10.8	0.0312	0.2680	0.7725
	+ FitNet	4.5	10.8	0.0280	0.2415	0.7992
	+ $S^2\text{-KD}$ (Ours)	4.5	10.8	0.0240	0.2033	0.8291
MLP-Mixer	Baseline	6.2	15.1	0.0350	0.2915	0.7504
	+ FitNet	6.2	15.1	0.0321	0.2706	0.7810
	+ $S^2\text{-KD}$ (Ours)	6.2	15.1	0.0286	0.2407	0.8077

Table 3: Performance on the **Prometheus** (OOD) dataset. The results show that our $S^2\text{-KD}$ provides the best generalization to unseen physical conditions, achieving performance closest to the teacher model with a fraction of the computational cost.

classic knowledge distillation methods? We evaluate our framework through extensive experiments on three benchmark datasets with diverse dynamics, WeatherBench, TaxiBJ+ and Prometheus, with the performance of teacher models serving as a reference upper bound.

The detailed results in Tables 1, 2, and 3 clearly demonstrate the effectiveness of our approach. While classic knowledge distillation methods like FitNet consistently improve upon the baseline models, our $S^2\text{-KD}$ framework achieves a far more substantial leap in performance. This superiority is evident across all datasets; for instance, the $S^2\text{-KD}$ empowered U-Net shows a 9.0% MSE improvement over FitNet on WeatherBench. This suggests that for complex spatiotemporal dynamics, simple feature mimicry is suboptimal, whereas the structured semantic and spectral knowledge provided by $S^2\text{-KD}$ offers a more potent guidance signal. Moreover, this performance advantage is not confined to a single architecture. The superiority of our framework holds consistently across diverse student models, including the convolution-based U-Net and ResNet, and the MLP-based Mixer, robustly demonstrating that $S^2\text{-KD}$ is a general, architecture-agnostic framework. Ultimately, $S^2\text{-KD}$ enables lightweight student models with only a few million parameters to achieve performance remarkably close to their massive teacher counterparts. On the TaxiBJ+ dataset, the distilled U-Net (MSE of 0.1180) nearly matches the performance of the 125.8M-parameter teacher (MSE of 0.1172), highlighting the framework’s exceptional capability in balancing high performance with computational efficiency. In summary, the experimental results provide compelling evidence for the superiority of the $S^2\text{-KD}$ framework, establishing it as a powerful and versatile solution that not only surpasses traditional distillation methods but also consistently elevates the performance of lightweight models to new heights across various tasks and architectures.

Ablation Study (RQ2.)

To answer our second research question (RQ2): *What are the individual contributions of the semantic and spectral distillation components?* we conduct a series of ablation experiments to quantitatively dissect the effectiveness of each core component within our $S^2\text{-KD}$ framework. This study is

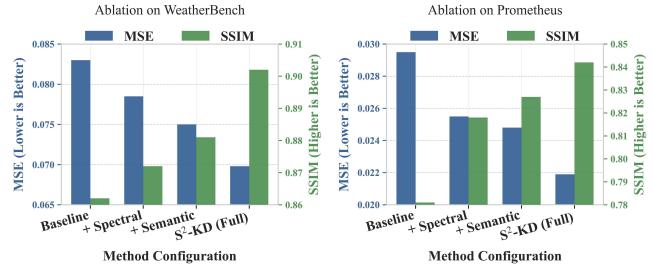


Figure 3: Ablation study of $S^2\text{-KD}$ on (a) WeatherBench and (b) Prometheus. Both MSE (blue, left axis) and SSIM (green, right axis) metrics demonstrate that while each component is individually beneficial, their combination in the full $S^2\text{-KD}$ framework yields the best performance. This validates the synergistic effect of integrating semantic and spectral knowledge.

performed with the U-Net as the student model across three datasets with distinct characteristics.

The experimental results are clearly presented in Table 4 and Figure 3, from which a highly consistent trend can be observed.

First, introducing either the spectral distillation loss (+ Spectral) or the semantic distillation loss (+ Semantic) alone improves performance over the baseline model across all metrics. For instance, on the TaxiBJ+ dataset (Table 4), semantic distillation alone reduces the MSE from 0.1354 to 0.1261. This provides strong evidence for the individual effectiveness of our framework’s two core components.

Second, the full $S^2\text{-KD}$ framework, which integrates both components, achieves the best performance in all tested scenarios. As illustrated in Figure 3, on both WeatherBench and Prometheus, the full method not only secures the lowest MSE but also attains the highest SSIM. This result compellingly demonstrates a significant synergistic effect between the semantic and spectral knowledge. They are not merely additive but complementary, and their combination is essential for maximizing the student model’s performance, thus validating the rationale behind our method’s design.

Baseline	Method Components			Performance Metrics		
	+ \mathcal{L}_{pred}	+ $\mathcal{L}_{spectral}$	+ $\mathcal{L}_{semantic}$	MSE ↓	MAE ↓	SSIM ↑
✓	✓			0.1354	1.1032	0.9532
✓		✓		0.1280	1.0617	0.9591
✓		✓	✓	0.1261	1.0455	0.9610
✓	✓	✓	✓	0.1180	0.9855	0.9754

Table 4: Ablation study of $S^2\text{-KD}$ components on the TaxiBJ+ dataset with a U-Net student. The results demonstrate that both semantic and spectral distillation are beneficial, and their combination yields a synergistic effect, leading to the best overall performance.

Analysis on Different Large Language Models

A core hypothesis of our $S^2\text{-KD}$ framework is that high-quality textual narratives provide invaluable semantic and causal knowledge to the teacher model. To validate this

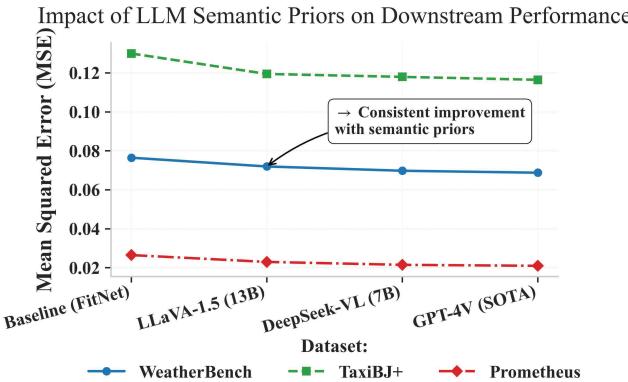


Figure 4: Impact of different LMMs on the performance of the U-Net student across three datasets. The results show a positive correlation between LMM capability and student performance (MSE). Meanwhile, the largest performance leap occurs when moving from no semantic prior (Baseline) to using any LMM, demonstrating the robustness and practicality of our framework.

and investigate the framework’s sensitivity to the quality of the language model, we conduct a comparative experiment. We select three representative Large Multimodal Models (LMMs) to generate textual descriptions and compare their results against a baseline (FitNet) that uses no semantic priors.

The results, illustrated in Figure 4, reveal two important conclusions. First, the final performance of the student model positively correlates with the capability of the LMM providing the text. Across all three datasets, we observe a clear downward trend: as the LMM progresses from the open-source LLaVA-1.5 and DeepSeek-VL to the state-of-the-art GPT-4V, the student model’s MSE consistently decreases. This provides strong evidence that higher-quality, more insightful textual descriptions indeed translate into more effective knowledge, thereby enhancing student performance.

Second, and equally important, the most significant performance gain occurs in the leap from having no semantic prior (Baseline) to having any semantic prior (LLaVA-1.5). This indicates that our framework is robust and not fragilely dependent on a single, top-tier model. Even with moderately-sized, open-source models, S^2 -KD delivers substantial benefits far exceeding traditional methods. This analysis not only confirms the importance of semantic knowledge quality but also demonstrates the practicality and adaptability of the S^2 -KD framework as a general approach.

Qualitative Analysis (RQ3.)

To provide a more intuitive understanding of the benefits of our proposed S^2 -KD framework, we present a qualitative comparison of prediction results in Figure 5. The figure displays three panels: the ground-truth weather state, the prediction from a lightweight U-Net student distilled with S^2 -KD, and the prediction from an identical U-Net trained with



Figure 5: Qualitative comparison of prediction results on the WeatherBench dataset.

a standard baseline method.

The ground-truth image (left panel) exhibits a high degree of complexity, characterized by intricate, fine-grained vortex structures and sharp gradients, particularly in the high-latitude region highlighted by the red dashed box. These features represent the high-frequency components of the atmospheric dynamics, which are notoriously difficult for compact models to capture.

The prediction from the baseline U-Net (right panel) starkly illustrates the challenge. The result is overly smooth and blurry, indicating a significant loss of high-frequency spectral information. The detailed structures within the highlighted box are smeared into an almost uniform, indistinct patch. Furthermore, the image is plagued by visible artifacts, such as unnatural blockiness and horizontal banding, which betray a superficial, pixel-level pattern imitation rather than a genuine understanding of the underlying physical processes. This outcome is a clear manifestation of the “semantic vacuum” we identified: the model learns *what* the general pattern looks like (cold poles, warm equator) but remains blind to *why* and *how* the specific, coherent structures form.

In stark contrast, the prediction from our S^2 -KD-enhanced student (middle panel) demonstrates a remarkable improvement. The overall clarity and sharpness are significantly closer to the ground truth. Crucially, within the highlighted region, the model successfully reconstructs the complex vortex structures with impressive fidelity. This visual evidence validates the dual-component design of our S^2 -KD loss. The **Spectral Alignment Loss** (L_{spectral}) has effectively forced the student to preserve high-frequency details, preventing the blurry output seen in the baseline. More profoundly, the **Semantic Alignment Loss** (L_{semantic}) has endowed the student with the causal and structural knowledge distilled from the privileged, text-informed teacher. As a result, the student’s prediction is not merely a collection of accurate pixels but a **semantically coherent** whole, representing a physically plausible weather state.

Conclusion

In this work, we introduced S^2 -KD, a novel knowledge distillation framework that enriches lightweight spatiotemporal forecasting models with semantic and causal understanding, moving beyond simple pixel-level mimicry. By distilling unified semantic-spectral knowledge from a privileged, text-informed multimodal teacher into a vision-only student, S^2 -KD significantly boosts prediction performance on diverse benchmarks, enabling simple models to approach the accuracy of massive, state-of-the-art counterparts.

References

- Bi, K.; Xie, L.; Zhang, H.; Chen, X.; Gu, X.; and Tian, Q. 2023. Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, 619(7970): 533–538.
- Blikstein, P. 2013. Multimodal learning analytics. In *Proceedings of the third international conference on learning analytics and knowledge*, 102–106.
- Bruna, J.; Zaremba, W.; Szlam, A.; and LeCun, Y. 2013. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*.
- Chen, L.; Zhong, X.; Zhang, F.; Cheng, Y.; Xu, Y.; Qi, Y.; and Li, H. 2023a. FuXi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj climate and atmospheric science*, 6(1): 190.
- Chen, P.; Liu, S.; Zhao, H.; and Jia, J. 2021. Distilling knowledge via knowledge review. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5008–5017.
- Chen, Y.; Li, Z.; Yang, Y.; Xie, L.; Liu, Y.; Ma, L.; Liu, S.; and Tian, G. 2022. CICC: Channel Pruning via the Concentration of Information and Contributions of Channels. In *BMVC*, 243.
- Chen, Y.; Ren, K.; Wang, Y.; Fang, Y.; Sun, W.; and Li, D. 2023b. ContiFormer: Continuous-time transformer for irregular time series modeling. In *NeurIPS*.
- Chen, Y.; and Wang, Z. 2024. An effective information theoretic framework for channel pruning. *arXiv preprint arXiv:2408.16772*.
- Fahlman, S.; and Fernández, R. 2022. Long-term 3D MHD simulations of black hole accretion discs formed in neutron star mergers. *Monthly Notices of the Royal Astronomical Society*, 513(2): 2689–2707.
- Fan, H.; Jiang, J.; Zhang, C.; Wang, X.; and Lai, Y.-C. 2020. Long-term prediction of chaotic systems with machine learning. *Physical Review Research*, 2(1): 012080.
- Gao, Y.; Wu, H.; Shu, R.; Dong, H.; Xu, F.; Chen, R.; Yan, Y.; Wen, Q.; Hu, X.; Wang, K.; et al. 2025. OneForecast: A Universal Framework for Global and Regional Weather Forecasting. *arXiv preprint arXiv:2502.00338*.
- Gao, Z.; Tan, C.; Wu, L.; and Li, S. Z. 2022. Simvp: Simpler yet better video prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3170–3180.
- Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6): 1789–1819.
- Guibas, J.; Mardani, M.; Li, Z.; Tao, A.; Anandkumar, A.; and Catanzaro, B. 2021. Adaptive fourier neural operators: Efficient token mixers for transformers. *arXiv preprint arXiv:2111.13587*.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kurth, T.; Subramanian, S.; Harrington, P.; Pathak, J.; Mardani, M.; Hall, D.; Miele, A.; Kashinath, K.; and Anandkumar, A. 2023. Fourcastnet: Accelerating global high-resolution weather forecasting using adaptive fourier neural operators. In *Proceedings of the platform for advanced scientific computing conference*, 1–11.
- Li, Z.; Kovachki, N.; Azizzadenesheli, K.; Liu, B.; Bhattacharya, K.; Stuart, A.; and Anandkumar, A. 2020. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.
- Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Yang, H.; et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.
- Mirzadeh, S. I.; Farajtabar, M.; Li, A.; Levine, N.; Matsukawa, A.; and Ghasemzadeh, H. 2020. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 5191–5198.
- Murata, R.; Okubo, F.; Minematsu, T.; Taniguchi, Y.; and Shimada, A. 2023. Recurrent neural network-fitnets: improving early prediction of student performance by time-series knowledge distillation. *Journal of Educational Computing Research*, 61(3): 639–670.
- Park, W.; Kim, D.; Lu, Y.; and Cho, M. 2019. Relational knowledge distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3967–3976.
- Pechyonny, D.; and Vapnik, V. 2010. On the theory of learning with privileged information. *Advances in neural information processing systems*, 23.
- Phuong, M.; and Lampert, C. 2019. Towards understanding knowledge distillation. In *International conference on machine learning*, 5142–5151. PMLR.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Ramachandram, D.; and Taylor, G. W. 2017. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6): 96–108.
- Rasp, S.; Dueben, P. D.; Scher, S.; Weyn, J. A.; Mouatadid, S.; and Thuerey, N. 2020. WeatherBench: a benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11): e2020MS002203.

- Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, proceedings, part III* 18, 234–241. Springer.
- Shi, X.; Chen, Z.; Wang, H.; Yeung, D.-Y.; Wong, W.-K.; and Woo, W.-c. 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28.
- Sorjamaa, A.; Hao, J.; Reyhani, N.; Ji, Y.; and Lendasse, A. 2007. Methodology for long-term prediction of time series. *Neurocomputing*, 70(16–18): 2861–2869.
- Stanton, S.; Izmailov, P.; Kirichenko, P.; Alemi, A. A.; and Wilson, A. G. 2021. Does knowledge distillation really work? *Advances in neural information processing systems*, 34: 6906–6919.
- Tolstikhin, I. O.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Steiner, A.; Keysers, D.; Uszkoreit, J.; et al. 2021. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34: 24261–24272.
- Wang, X.; Ma, Y.; Wang, Y.; Jin, W.; Wang, X.; Tang, J.; Jia, C.; and Yu, J. 2020. Traffic flow prediction via spatial-temporal graph neural network. In *TheWebConf*, 1082–1092.
- Wang, Y.; Han, R.; and Chen, Y. 2025. TCFI: Topology-Consistent Pruning with Fisher Information for Efficient Medical Image Segmentation. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6.
- Wang, Y.; Wu, H.; Zhang, J.; Gao, Z.; Wang, J.; Philip, S. Y.; and Long, M. 2022. Predrnn: A recurrent neural network for spatiotemporal predictive learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2): 2208–2225.
- Wu, H.; Gao, Y.; Shu, R.; Han, Z.; Xu, F.; Zhu, Z.; Wen, Q.; Wu, X.; Wang, K.; and Huang, X. 2025a. Turb-L1: Achieving Long-term Turbulence Tracing By Tackling Spectral Bias. *arXiv preprint arXiv:2505.19038*.
- Wu, H.; Gao, Y.; Shu, R.; Wang, K.; Gou, R.; Wu, C.; Liu, X.; He, J.; Cao, S.; Fang, J.; Shi, X.; Tao, F.; Song, Q.; Ji, S.; Xiang, Y.; Sun, Y.; Li, J.; Xu, F.; Dong, H.; Wang, H.; Zhang, F.; Zhao, P.; Wu, X.; Wen, Q.; Chen, D.; and Huang, X. 2025b. Advanced long-term earth system forecasting by learning the small-scale nature. *arXiv preprint arXiv:2505.19432*.
- Wu, H.; Liang, Y.; Xiong, W.; Zhou, Z.; Huang, W.; Wang, S.; and Wang, K. 2024a. Earthfarsser: Versatile Spatio-Temporal Dynamical Systems Modeling in One Model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15906–15914.
- Wu, H.; Wang, H.; Wang, K.; Wang, W.; Tao, Y.; Chen, C.; Hua, X.-S.; Luo, X.; et al. 2024b. Prometheus: Out-of-distribution fluid dynamics modeling with disentangled graph ode. In *Forty-first International Conference on Machine Learning*.
- Wu, H.; Wang, S.; Liang, Y.; Zhou, Z.; Huang, W.; Xiong, W.; and Wang, K. 2023. Earthfarsser: Versatile Spatio-Temporal Dynamical Systems Modeling in One Model. *AAAI2024*.
- Wu, H.; Wen, H.; Zhang, G.; Xia, Y.; Liang, Y.; Zheng, Y.; Wen, Q.; and Wang, K. 2024c. Dynst: Dynamic sparse training for resource-constrained spatio-temporal forecasting. *arXiv preprint arXiv:2403.02914*.
- Wu, H.; Xu, F.; Chen, C.; Hua, X.-S.; Luo, X.; and Wang, H. 2024d. Pastnet: Introducing physical inductive biases for spatio-temporal video prediction. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2917–2926.
- Zhang, S.; Liu, Y.; Sun, Y.; and Shah, N. 2021. Graph-less neural networks: Teaching old mlps new tricks via distillation. *arXiv preprint arXiv:2110.08727*.