Spectral-Aware Text-to-Time Series Generation with Billion-Scale Multimodal Meteorological Data

Shijie Zhang
School of Computer Science and Engineering, Northeastern University, Shenyang, China
20235914@stu.neu.edu.cn

***Abstract*—Text-to-time-series generation is particularly important in meteorology, where natural language offers intuitive control over complex, multi-scale atmospheric dynamics. Existing approaches are constrained by the lack of large-scale, physically grounded multimodal datasets and by architectures that overlook the spectral–temporal structure of weather signals. We address these challenges with a unified framework for text-guided meteorological time-series generation. First, we introduce MeteoCap-3B, a billion-scale weather dataset paired with expert-level captions constructed via a Multi-agent Collaborative Captioning (MACC) pipeline, yielding information-dense and physically consistent annotations. Building on this dataset, we propose MTransformer, a diffusion-based model that enables precise semantic control by mapping textual descriptions into multi-band spectral priors through a Spectral Prompt Generator, which guides generation via frequency-aware attention. Extensive experiments on real-world benchmarks demonstrate state-of-the-art generation quality, accurate cross-modal alignment, strong semantic controllability, and substantial gains in downstream forecasting under data-sparse and zero-shot settings. Additional results on general time-series benchmarks indicate that the proposed framework generalizes beyond meteorology.**

***Index Terms*—Time Series Generation, Agentic AI**

I. Introduction

Generative modeling has achieved remarkable success in images [1], audio [2], and text [3], yet its potential for physical time series remains largely underexplored [4], [5]. In meteorology, text-guided generation of weather time series is particularly valuable, as natural language provides an intuitive interface for controlling complex, multi-scale atmospheric dynamics. Such capability enables the simulation of rare extreme events [6]–[8], data augmentation for forecasting in observation-sparse regions [9], and interactive climate analysis. Despite this promise, text-to-time-series generation poses challenges that fundamentally distinguish it from other modalities, arising from both data availability and model design.

A primary bottleneck lies in the absence of large-scale, high-quality datasets that pair time series with information-rich and physically meaningful textual descriptions. Unlike vision–language corpora such as LAION-5B [10], existing time-series datasets (e.g., Time-MMD [11], Time-IMM [12]) largely rely on rigid templates or heuristic rules. Consequently, their captions are often homogeneous and low in semantic density, describing surface-level values rather than underlying physical mechanisms. Such annotations rarely capture domain-specific reasoning, for example, distinguishing structured phenomena like diurnal cycles from stochastic fluctuations or explaining the physical causes of abrupt regime shifts. Without large-scale supervision that reflects diverse physical behaviors and causal semantics, models struggle to learn fine-grained cross-modal alignment between language and complex temporal dynamics.

Equally critical is the mismatch between existing generative architectures and the intrinsic structure of physical time series. Although diffusion models have achieved strong performance in other modalities [13], their direct adaptation to time series often yields degraded results [14]–[16]. Meteorological signals are governed by superimposed frequency components spanning long-term trends, medium-term periodicities, and high-frequency variability [17]–[21]. However, most text-conditioning strategies [22] operate purely in semantic space via cross-attention over raw text embeddings, lacking explicit inductive bias in the frequency domain. As a result, generated sequences may appear semantically plausible yet exhibit spectral distortions, failing to preserve periodicity, volatility, and long-range coherence. Bridging high-level linguistic instructions with low-level spectral–temporal dynamics therefore remains an open challenge.

To address these challenges, we propose a unified framework that advances both data construction and generative modeling for text-guided meteorological time series. First, we introduce **MeteoCap-3B**, a billion-scale weather dataset constructed via a Multi-agent Collaborative Captioning (MACC) pipeline. By coordinating LLM agents as perceptual analyzers, physical reasoners, and consistency critics, MACC produces physically grounded and information-dense captions that substantially surpass template-based annotations in diversity and domain fidelity. Second, we propose **MTransformer**, a diffusion generative model that explicitly incorporates spectral conditioning into the generation process. At its core, a *Spectral Prompt Generator* translates natural language descriptions into multi-band spectral priors, which guide generation through a frequency-aware attention mechanism. This framework provides a new approach that jointly advances large-scale dataset construction and spectral-aware generative modeling for text-guided meteorological time series. Our contributions are:

- We introduce the multi-agent-driven reasoning pipeline for automated weather time-series captioning, producing **MeteoCap-3B**, a large-scale, physically consistent multimodal dataset verified by human experts.
- We propose **MTransformer**, a diffusion-based architecture that injects explicit spectral conditioning into the generation process, enabling precise natural-language

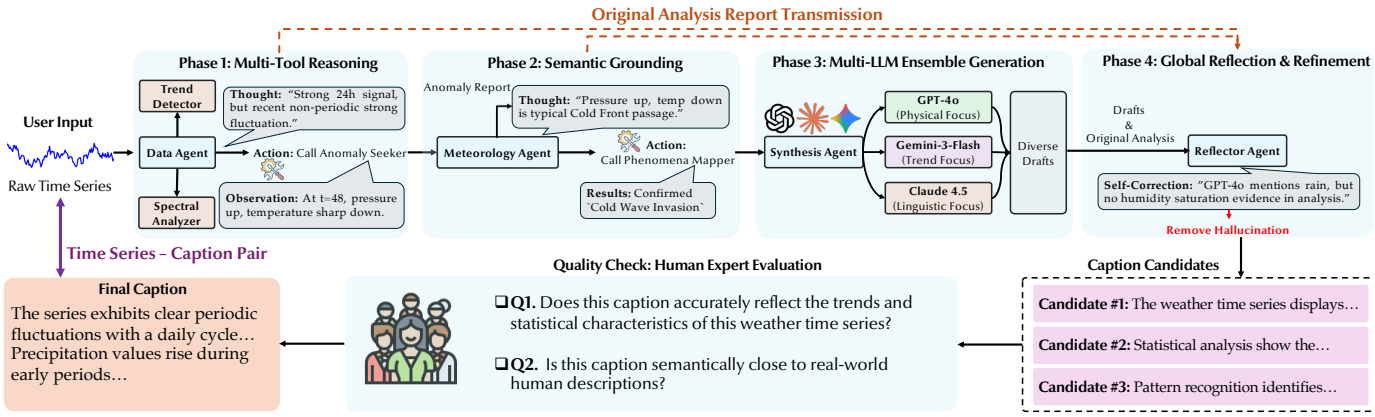

Fig. 1. Overview of our Multi-agent Collaborative Captioning (MACC) pipeline for MeteoCap-3B construction. Note that we used GPT-4o, Claude Sonnet 3.5, and Gemini-3-Flash in Phase 3 and used DeepSeek V3 [23] in Phase 4. Human experts contains eight PhD in meteorology.

control over multi-scale temporal dynamics.

- Extensive experiments demonstrate state-of-the-art generation fidelity, accurate cross-modal alignment, and significant improvements in downstream forecasting under zero-shot and data-sparse settings.

II. MeteoCap-3B: Agentic Captioning

To bridge the modality gap between numerical weather observations and natural language, we construct **MeteoCap-3B**, a large-scale multimodal dataset comprising three billion meteorological observations paired with high-quality textual descriptions. The dataset is built based on our proposed *Multi-agent Collaborative Captioning (MACC)* pipeline with human-in-the-loop refinement, ensuring both semantic richness and physical consistency, as shown in **Fig. 1**.

A. Multi-agent Collaborative Captioning

The high-fidelity captioning for weather time series is formulated as a multi-stage collaborative process involving heterogenous LLM agents and specialized meteorological tools.

Phase 1: Multi-Tool Reasoning. The pipeline begins by extracting structured primitives from the raw sequence $\mathbf{x}_{1:T}$ under a ReAct-style workflow [24]. An agent iteratively interacts with a set of analytical tools $\mathcal{T}_{\text{perc}} = \{\text{Trend\_Det}, \text{Anomaly\_Seek}, \text{FFT\_Analyzer}\}$ to characterize temporal trends, abrupt changes, and dominant frequency components. For example, the agent computes the spectral energy distribution $\mathcal{E}_{\text{freq}}$ to identify periodicities and invokes the anomaly detector to locate abrupt shifts $\mathcal{A} = \{t||\nabla x_t| > \delta\}$. The output of this phase is a structured perception vector $\mathbf{v}_{feat} = \text{Agent}(\mathbf{x}_{1:T}; \mathcal{T}_{perc})$ that summarizes the statistical and frequency-domain essence of the series

Phase 2: Semantic Grounding. The extracted primitives $\mathbf{v}_{\text{feat}}$ are then mapped to physically meaningful meteorological concepts. Equipped with a knowledge-retrieval tool $\mathcal{T}_{\text{kb}}$ encoding synoptic climatology rules, the agent performs symbolic reasoning over feature co-occurrences to infer high-level phenomena. This grounding step yields a semantic representation $\mathcal{K} = \text{Ground}(\mathbf{v}_{feat}, \mathcal{T}_{kb})$ containing interpretable tags such as *frontogenesis*, *diurnal thermal inversion*, or *convective instability*, ensuring that subsequent captions reflect physically plausible interpretations rather than surface-level descriptions.

System Prompt: Phase 1

Role: You are an expert meteorological data analyst.
Goal: Analyze the input time series $\mathbf{x}_{1:T}$ using available tools.
Protocol:
1) Reason inside `<think>` tags before taking action.
2) Call tools inside `<tool>` tags.
3) Conclude with `<answer>` containing feature vector.

Input Series: [288.15, 287.50, 285.20, ...] (Temperature K)
Model Generation:
`<think>` The sequence shows a sudden drop. I need to check if this is a trend or a sudden anomaly. First, I will detect the overall trend. `</think>`
`<tool>` call: Trend_Detector(window=24) `</tool>`
[System: Tool Output: Trend=-0.85 (Strong Downward), p-value<0.01]
`<think>` The trend is significant. Now I need to check for periodicity to see if this is just a diurnal cycle. `</think>`
`<tool>` call: FFT_Analyzer(top_k=3) `</tool>` ...

Phase 3: Multi-LLM Ensemble Generation. To capture diverse linguistic styles and focus points, we employ an *Ensemble Generation Agent* that orchestrates multiple LLMs. Each model $\mathcal{M}_m \in \{\mathcal{M}_{GPT}, \mathcal{M}_{Clau}, \mathcal{M}_{Gemini}\}$ is prompted with the grounded knowledge $\mathcal{K}$ and a specific task-role (e.g., "Standard Reporting," "Trend Analysis," or "Casual Human-like Description"). This results in a candidate set of captions $\mathcal{C} = \{c_1, c_2, c_3\}$, where each $c_i = \text{Generate}(\mathcal{K}, \mathbf{x}_{1:T}; \mathcal{M}_i)$. This ensemble approach mitigates the stochastic bias of a single model and enriches the semantic diversity of the dataset.

Phase 4: Global Reflection and Self-Correction. The final phase involves a *Reflector Agent* (powered by DeepSeek-V3) tasked with cross-consistency verification. The reflector takes $\mathcal{C}$ and the original data primitives $\mathbf{v}_{feat}$ as input to perform a "fact-checking" loop. It identifies hallucinations, such as an LLM claiming a "heavy rain" trend when the humidity sensor

System Prompt: Phase 2

Role: You are a senior synoptic meteorologist. **Task:** Interpret the provided numerical features using the retrieved climatological rules. **Input:**

1) Feature Vector: `feature_vector` (e.g., trend=-0.8, fft_peak=24h, anomaly_idx=[48])
2) Retrieved Rules: `retrieved_rules` (e.g., "Sharp temp drop + pressure rise to Cold Front")

Instruction: - Do NOT generate the final caption yet. - Output a list of **Physical Tags** and a brief **Reasoning Chain**. - Verify if the anomaly aligns with the retrieved rules.
Example Output: `<reasoning>` The temperature trend is negative (-0.8), and there is a sharp anomaly at t=48. Combined with the retrieved rule, this strongly suggests a cold front passage disrupting the diurnal cycle. `</reasoning>` `<tags>` [Phenomenon: Cold_Front], [Scale: Synoptic], [State: Transition_Phase] `</tags>`

System Prompt: Phase 4

Role: You are a strict quality assurance auditor for meteorological reports. **Input:**

1) Original Data Facts: `feature_vector` (Ground Truth)
2) Candidate Caption: `candidate_caption`

Checklist:

1) **Hallucination Check:** Does the caption mention events (e.g., "rain", "storm") not supported by the data features?
2) **Trend Consistency:** Does the described direction (up/down) match the numerical trend?
3) **Extrema Accuracy:** Are the mentioned peak times roughly correct?

Instruction: - If errors are found, output `STATUS: REJECT` and provide specific `<feedback>`. - If the caption is factual, output `STATUS: PASS`.
Example Interaction:
Input Caption: "The region experienced heavy rainfall..."
Data Fact: Precipitation variable is consistently 0.
Output: `STATUS: REJECT`
`<feedback>` Hallucination detected: Caption claims rainfall, but data shows zero precipitation. Remove references to rain. `</feedback>`

shows low saturation. The agent provides feedback $\mathcal{F}$ to the generation agents for iterative refinement: $c_i' = \text{Refine}(c_i, \mathcal{F})$. This phase ensures the internal validity of the captions before they are passed to the human-expert quality check stage, significantly reducing the human cognitive load.

B. Human Expert Quality Check

To ensure the fidelity and naturalness of the captions, we implement a rigorous HITL consensus mechanism. A panel of human meteorological experts evaluate $\mathcal{C}$ based on two orthogonal dimensions, including **Q1. Physical Alignment (S_{pa}):** Does the caption accurately reflect the trends, extrema, and statistical characteristics of $\mathbf{x}_{1:T}$? **Q2. Semantic Realism (S_{sr}):** Is the caption's linguistic structure consistent with professional human meteorological reporting? The final caption c^* is selected or refined to maximize a joint utility function:

$$c^* = \arg\max_{c \in \mathcal{C}} \left(\alpha \cdot \text{Sim}(\phi(\mathbf{x}_{1:T}), \psi(c)) + \beta \cdot \text{LLM-Eval}(c) \right), \tag{1}$$

where $\phi(\cdot)$ is a pre-trained time-series encoder, $\psi(\cdot)$ is a text embedding Qwen3-Embedding-7B [25]), and $\text{Sim}(\cdot, \cdot)$ denotes the cosine similarity in the joint embedding space.

C. Complexity-based Categorization.

To facilitate robust training and evaluation, we categorize **MeteoCap-3B** into three subsets based on a *Dynamical Complexity Index* ($\mathcal{D}$), computed as the weighted entropy of the signal's Lyapunov exponent and spectral volatility:

- **Meteo-Steady ($\mathcal{D} < \delta_1$):** Sequences with high periodicity and low innovation, representing 60% of the corpus.
- **Meteo-Transitional ($\delta_1 \leq \mathcal{D} < \delta_2$):** Sequences featuring state transitions, representing 30% of the corpus.
- **Meteo-Volatile ($\mathcal{D} \geq \delta_2$):** Sequences containing extreme weather events and stochastic turbulence, representing the remaining 10%, as the most challenging benchmark.

Representative samples in MeteoCap-3B as shown in Fig. 2.

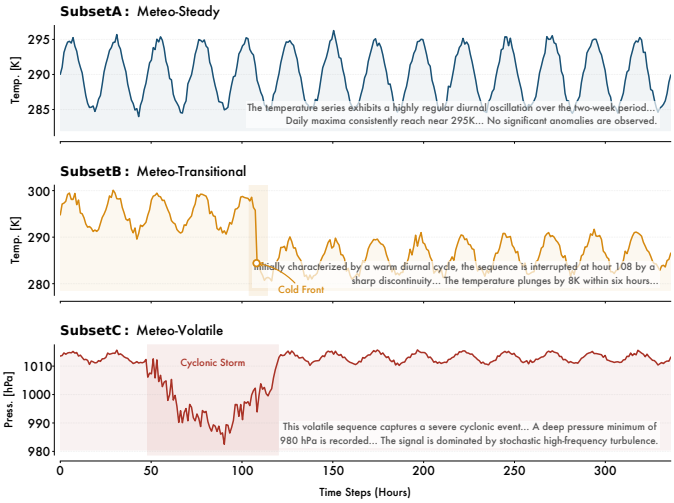

Fig. 2. Representative samples across three subsets of MeteoCap-3B.

D. Statistical Characteristics of MeteoCap-3B

MeteoCap-3B is a large-scale multimodal meteorological corpus containing three billion observations organized into about one hundred million high-quality sequence–caption pairs. The data are curated from the Integrated Surface Database (ISD) [26], covering thirty four years from 1991 to 2025 and more than thirty thousand stations worldwide. Each sequence $\mathbf{x}_{1:T}$ has length up to $T = 512$, corresponding to one week of hourly measurements, and spans seven Köppen–Geiger climate zones. The dataset captures diverse temporal dynamics, including a Meteo-Steady subset with pronounced diurnal periodicity, where the average twenty four hour autocorrelation exceeds 0.85, and a Meteo-Volatile subset containing roughly 1.2 million extreme events such as rapid

pressure drops above 24 hPa within twenty four hours. The accompanying textual modality forms a rich semantic layer, with over fifty thousand unique tokens and dense use of professional meteorological terminology. Each sequence is paired with multiple expert-written captions, and the final descriptions average forty five words with high linguistic complexity, reflected by a Flesch–Kincaid grade level of 10.5.

TABLE I
STATISTICS OF METEOCAP-3B. TEMP: TEMPERATURE, PRESS: PRESSURE, HUMID: HUMIDITY, WIND: WIND SPEED, PRECIP: PRECIPITATION.

Category	Metric	Value
Global Scale	Total observation points	$3.15 \times 10^9+$
	Number of stations	31,400+
	Time span	1991–2025
Series Modality	Variables per step	5 (Temp, Press, Humid, Wind, Precip)
	Sequence length (T)	168 (hourly) / 512 (sub-hourly)
	Steady / Transient / Volatile	60% / 30% / 10%
Text Modality	Total unique captions	Approx. 105,000,000
	Average caption length	45.2 words
	Vocabulary size	52.4K
Alignment	Human QC pass rate	94.2%
	Avg. cosine similarity	0.824

E. Spectral-aware Text-to-Time Series Generation

Directly adapting text-conditioned diffusion models to weather time series often neglects their intrinsic frequency structures, which are critical for physical consistency in meteorological data. To address this limitation, we propose a *Spectral-aware Text-to-Time Series Diffusion* (**Fig. 3**), which injects explicit multi-band spectral priors that are decoded from natural language captions into a Diffusion Transformer (DiT) [27] operating in the latent space of a VQ-VAE [28].

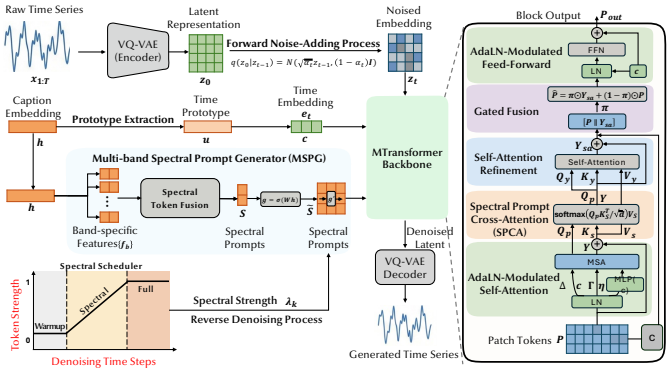

Fig. 3. Our MTransformer for time series generation. *Qwen-3-Embedding-7B* [25] is used to achieve the caption embedding with 1024 dimensions.

Latent Diffusion Process. A raw sequence $\mathbf{x}_{1:T}$ is first encoded into a compact latent representation $\mathbf{z}_0 \in \mathbb{R}^{H \times W}$. Following DDPM [13], the forward process gradually perturbs $\mathbf{z}_0$ with Gaussian noise: $q(\mathbf{z}_t \mid \mathbf{z}_{t-1}) = \mathcal{N}(\sqrt{\alpha_t}\mathbf{z}_{t-1}, (1-\alpha_t)\mathbf{I})$. The denoiser $\boldsymbol{\epsilon}_\theta$ is trained to recover $\mathbf{z}_0$, conditioned on both textual semantics and frequency-specific spectral prompts.

Multi-band Spectral Prompt Generator (MSPG). To explicitly align textual descriptions with frequency-domain structures, we introduce the *MSPG*. Given a caption embedding $\mathbf{h}$, MSPG projects $\mathbf{h}$ into B frequency-specific subspaces:

$$\mathbf{f}_b = W_b\mathbf{h}, \quad b = \{1, \ldots, B\}, \tag{2}$$

where $\{W_b\}$ are learnable projections corresponding to distinct frequency bands. Unlike fixed octave or predefined bands, these bands are *learned end-to-end*, allowing the model to adaptively capture domain-specific temporal scales (e.g., diurnal, synoptic, or high-frequency variability) from data. For each band, a lightweight generator produces M spectral tokens as $\mathbf{S}_b = \mathrm{Gen}(\mathbf{f}_b) \in \mathbb{R}^{M \times d}$, which are then concatenated and linearly fused: $\mathbf{S} = \mathrm{Fuse}(\{\mathbf{S}_b\}_{b=1}^{B}) \in \mathbb{R}^{M \times d}$. MSPG is trained jointly with the diffusion model via the standard denoising objective, without additional supervision, which encourages the generated spectral prompts to encode frequency priors that are most useful for latent denoising.

Adaptive Spectral Gating. To allow selective activation of frequency bands based on textual semantics, we introduce an adaptive gating mechanism as follows:

$$\mathbf{g} = \sigma(W_g\mathbf{h}) \in \mathbb{R}^M, \tag{3}$$

where each element of $\mathbf{g}$ modulates the importance of a corresponding spectral token. The gated spectral prompts are given by $\tilde{\mathbf{S}} = \mathbf{S} \odot \mathbf{g}$. Semantically, this enables captions emphasizing, for example, "volatile fluctuations" or "smooth trends" to dynamically amplify high- or low-frequency components.

Denoising with Spectral-Prompt Cross-Attention. The denoiser is a Transformer. Each block first applies AdaLN-modulated self-attention to temporal patch tokens $\mathbf{P}$, followed by our proposed *Spectral-Prompt Cross-Attention (SPCA)*:

$$\bar{\mathbf{Y}} = \mathrm{Softmax}\left(\frac{(PW_Q)(\tilde{S}W_K)^\top}{\sqrt{d}}\right)(\tilde{S}W_V),$$

where $\mathbf{P}$ serves as Query and $\tilde{\mathbf{S}}$ as Key and Value. This design enforces frequency-aware refinement of the denoising trajectory, directly conditioning temporal updates on text-derived spectral priors. To prevent overly strong spectral constraints at early diffusion steps, we introduce a *Spectral Scheduler* λ_k, which linearly increases from 0 to 1 over full training iterations. This weak-to-strong curriculum stabilizes optimization while progressively strengthening spectral guidance.

III. EXPERIMENTS AND RESULTS

We design experiments to answer the following questions:

- **RQ1:** How well does MTransformer generate high-quality time series compared with existing generative baselines across weather and general domains?
- **RQ2:** How does MTransformer scale with respect to model size, data scale, and sequence length?
- **RQ3:** Do the generated time series improve performance on downstream weather time series forecasting tasks?
- **RQ4:** Does MeteoCap-3B enable cross-modal alignment?

Implementation. During training, MeteoCap-3B is split into train, val, and test sets with a 7:2:1. The VQ-VAE is pretrained for 5k steps with 4 residual layers (hidden size 512, embedding

TABLE II

PERFORMANCE COMPARISON ON TSFRAGMENT-600K DATASETS AND OUR PROPOSED **METEOCAP-3B**. NOTE THAT FOR METEO DATASETS, WE EVALUATE ON LONGER SEQUENCES (UP TO 336) TO TEST LONG-RANGE CONSISTENCY. **BOLD**: THE BEST.

			MTransformer (Ours)			T2S [22] (IJCAI 2025)			DiffusionTS [14] (ICLR 2024)			TimeVAE [29] (arXiv 2025)			GPT-4o-mini [30] (2024)			Llama3.1-8b [31] (2023)			Gemini-3 (2025)		
Group	Datasets	Length	WAPE ↓	MSE ↓	MRR@10 ↑	WAPE ↓	MSE ↓	MRR@10 ↑	WAPE ↓	MSE ↓	MRR@10 ↑	WAPE ↓	MSE ↓	MRR@10 ↑	WAPE ↓	MSE ↓	MRR@10 ↑	WAPE ↓	MSE ↓	MRR@10 ↑	WAPE ↓	MSE ↓	MRR@10 ↑
TSFragment-600K [22]	ETTh1	24	**0.162**	0.009	0.279	0.183	**0.008**	**0.283**	0.793	0.077	0.267	0.666	0.055	0.211	0.264	0.041	0.104	0.883	0.663	0.097	0.245	0.035	0.112
		48	**0.216**	**0.014**	0.300	0.234	**0.013**	0.289	1.207	0.120	0.298	0.647	0.055	0.286	0.414	0.080	0.100	0.923	1.260	0.086	0.385	0.072	0.108
		96	**0.218**	0.012	**0.296**	0.229	**0.011**	0.291	0.498	0.028	0.214	0.643	0.055	0.286	0.500	0.118	0.096	0.949	1.748	0.056	0.466	0.105	0.099
	ETTm1	24	0.376	0.032	**0.287**	0.426	0.033	0.286	0.604	0.040	0.251	0.666	0.048	0.219	**0.244**	**0.031**	0.101	1.134	0.798	0.099	0.238	0.030	0.108
		48	**0.368**	0.054	0.280	0.53	0.053	0.283	1.119	0.100	**0.285**	0.636	**0.051**	0.217	0.453	0.112	0.097	1.074	1.496	0.079	0.425	0.101	0.103
		96	**0.389**	0.036	**0.302**	0.414	0.041	0.299	0.546	**0.031**	0.293	0.664	0.057	0.208	0.706	0.395	0.091	1.079	1.761	0.057	0.655	0.310	0.095
	Electricity	24	**0.128**	0.011	**0.282**	0.135	**0.010**	0.28	0.617	0.041	0.253	0.207	0.016	0.213	0.734	0.592	0.092	0.926	1.140	0.064	0.685	0.510	0.098
		48	**0.142**	**0.012**	**0.249**	0.155	0.013	0.244	1.128	0.102	0.227	0.208	0.017	0.216	1.014	1.065	0.068	1.038	1.416	0.054	0.950	0.980	0.072
		96	**0.203**	0.026	**0.314**	0.238	0.031	0.318	0.545	0.032	0.247	0.213	**0.018**	0.257	1.024	1.210	0.059	1.085	1.740	0.034	0.985	1.105	0.062
	Exchange Rate	24	**0.260**	**0.031**	**0.331**	0.292	0.033	0.334	0.791	0.077	0.272	1.165	0.105	0.252	1.072	2.060	0.052	1.258	2.052	0.045	0.995	1.850	0.058
		48	**0.237**	**0.029**	**0.319**	0.259	0.033	0.315	1.217	0.122	0.298	1.064	0.106	0.306	0.933	1.074	0.082	1.562	2.125	0.051	0.880	0.950	0.088
		96	**0.454**	**0.043**	**0.312**	0.48	0.047	0.31	0.504	0.048	0.216	0.977	0.106	0.274	1.141	1.625	0.054	1.433	1.892	0.055	1.050	1.450	0.060
	Air Quality	24	0.672	**0.019**	**0.313**	0.884	0.02	0.304	0.806	0.078	0.265	2.303	0.022	0.302	**0.557**	0.379	0.093	0.878	0.697	0.085	0.580	0.355	0.098
		48	1.245	**0.044**	**0.312**	1.295	0.044	0.297	1.439	0.120	0.221	1.648	0.023	0.271	**0.791**	0.715	0.08	1.141	1.642	0.046	0.810	0.680	0.085
		96	1.348	0.043	**0.351**	1.377	0.049	0.34	**0.508**	**0.028**	0.304	1.270	0.024	0.301	0.928	1.127	0.061	1.085	1.551	0.050	0.955	1.050	0.065
	Traffic	24	**0.331**	0.006	0.213	0.353	**0.005**	0.201	0.795	0.077	0.220	0.544	0.008	0.233	1.260	1.912	0.020	1.144	1.938	0.022	1.150	1.850	0.025
		48	**0.462**	**0.007**	**0.221**	0.506	0.008	0.219	1.202	0.120	0.188	0.594	0.011	0.211	1.189	1.928	0.011	1.138	1.988	0.004	1.120	1.880	0.015
		96	**0.512**	**0.009**	0.259	0.543	0.01	**0.262**	0.509	0.028	0.171	0.641	0.013	0.207	1.18	2.093	0.010	1.107	1.994	0.001	1.115	1.950	0.012
MeteoCap-3B	**Meteo Steady**	96	**0.092**	**0.005**	**0.345**	0.108	0.007	0.330	0.450	0.025	0.240	0.135	0.012	0.310	0.350	0.065	0.125	0.850	0.620	0.110	0.320	0.055	0.135
		192	**0.115**	**0.008**	**0.328**	0.142	0.010	0.305	0.580	0.038	0.210	0.168	0.015	0.285	0.420	0.095	0.110	0.980	0.850	0.095	0.405	0.088	0.118
		336	**0.138**	**0.011**	**0.312**	0.175	0.014	0.288	0.650	0.045	0.190	0.210	0.020	0.250	0.550	0.145	0.095	1.150	1.100	0.080	0.510	0.130	0.105
	Meteo Trans.	96	**0.155**	**0.012**	**0.305**	0.185	0.016	0.290	0.520	0.035	0.220	0.340	0.040	0.240	0.480	0.120	0.105	1.050	0.950	0.090	0.450	0.105	0.115
		192	**0.188**	**0.018**	**0.292**	0.230	0.024	0.275	0.680	0.055	0.195	0.420	0.065	0.210	0.650	0.210	0.090	1.250	1.350	0.075	0.610	0.190	0.098
		336	**0.225**	**0.025**	**0.278**	0.285	0.032	0.255	0.820	0.085	0.170	0.550	0.095	0.180	0.850	0.350	0.075	1.450	1.850	0.060	0.780	0.320	0.085
	Meteo Volatile	96	**0.265**	**0.028**	**0.285**	0.310	0.036	0.265	0.750	0.065	0.200	0.650	0.080	0.190	0.720	0.420	0.085	1.350	1.550	0.070	0.680	0.380	0.095
		192	**0.310**	**0.038**	**0.270**	0.385	0.048	0.245	0.920	0.110	0.180	0.820	0.120	0.160	0.950	0.650	0.070	1.650	2.100	0.050	0.890	0.580	0.078
		336	**0.365**	**0.052**	**0.255**	0.460	0.065	0.225	1.150	0.150	0.160	1.050	0.180	0.140	1.250	0.950	0.055	1.950	2.550	0.040	1.150	0.880	0.065
	1st count	-	**23**	**17**	**22**	0	5	2	1	2	1	0	2	0	3	1	0	0	0	0	0	0	0

size 128). The diffusion model is trained for 80k steps with spectral token length 6, text embedding size 1024, and text prototype size 128. The spectral training scheduler follows three phases: warmup (0–$20k$), spectral ($20k$–$40k$), and full ($40k$–$80k$). Our MTransformer uses a 12-layer Transformer with a hidden dimension of 768 and 12 attention heads, a feed-forward dimension of 3072, a temporal patch size of 16, and spectral prompts with 8 frequency bands and 16 tokens per band. Optimization is performed with AdamW using an initial learning rate of $1e^{-4}$ and a batch size of 256. The sampling steps are set to $1k$ unless otherwise specified. All implementations are conducted on 2 × NVIDIA A100 80GB.

Baseline and Evaluation Metrics. For time series generation, following [22], we compare MTransformer with representative generative baselines, including T2S [22], DiffusionTS [14], TimeVAE [29], and LLMs such as GPT-4o-mini [30], LLaMA-3.1 [31], and Gemini-3-Flash [22]. All methods are evaluated using three metrics: MSE, WAPE, and MRR@10. Experiments are conducted on two benchmarks: (i) TSFragment-600K [22], which spans seven domains with 600K samples, and (ii) our `MeteoCap-3B`, comprising 3B high-quality samples organized into three subsets of increasing difficulty.

A. RQ1: Evaluation on Text-to-Time Series Generation

Main Results. Table II summarizes the main results. MTransformer achieves the best performance in most settings, showing strong and stable generation quality. It consistently outperforms existing baselines on TSFragment-600K in WAPE and MRR@10 while maintaining competitive or best MSE, and exhibits a clearer advantage on `MeteoCap-3B` with lower errors across Steady, Transitional, and Volatile subsets, particularly under volatile conditions. These results demonstrate that MTransformer effectively handles long-range generation and challenging dynamics, outperforming both specialized time-series generators and general-purpose LLMs.

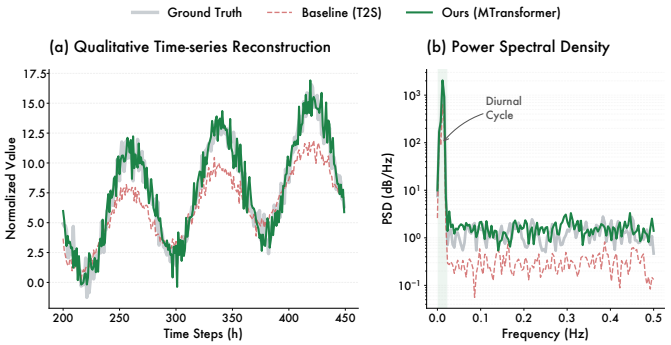

Fig. 4. Qualitative analysis. **Left:** Qualitative time series reconstruction. **Right:** Power spectral density of real and generated time series.

Qualitative Analysis. Fig. 4 compares generation with the ground truth in time and frequency domains. MTransformer closely preserves trends and fluctuations in the time domain and accurately matches the ground-truth power spectral density across both low and high frequencies, while T2S smooths fine-grained dynamics and exhibits clear spectral distortion.

B. RQ2: Evaluation on Model Scaling

To further assess scalability, we analyze MTransformer along three dimensions using MSE as a unified metric: model size, data scale, and sequence length, as shown in Figure 5. For model scaling, MTransformer exhibits consistent performance gains as the parameter count increases from 10M to 1B, without clear saturation, in contrast to the standard DiT. For data scaling, MTransformer effectively exploits the large-scale MeteoCap-3B corpus, with a steeper performance

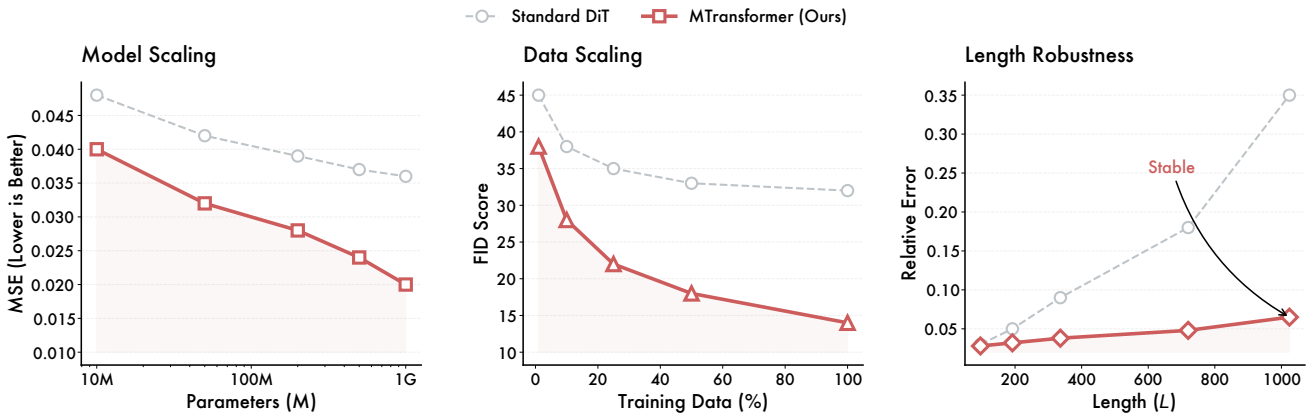

Fig. 5. Scaling behavior of MTransformer with respect to model size (length of 96), data scale, and sequence length, evaluated on the Meteo-Volatile subset.

improvement curve indicating higher data efficiency and lower MSE at comparable data budgets. Finally, when extending the sequence length from 96 up to 1024, standard models suffer from severe error accumulation, whereas MTransformer maintains stable and low MSE, demonstrating that the introduced spectral priors robustly anchor long-horizon generation. These demonstrate the superior scalability of our proposed MTransformer and provide insights into developing more scalable generative models for time series generation.

C. RQ3: Evaluation on Downstream Weather Forecasting

To assess the practical utility of the generated data, we conduct downstream forecasting experiments on the held-out ISD-Test dataset using TimesNet [32] and a standard Transformer [33] across four horizons $H \in \{96, 192, 336, 720\}$. The experiments evaluate both data augmentation and pre-training scenarios, using a 1:1 mix of synthetic and real data for augmentation. To prevent data leakage, all synthetic series are generated using only training-split metadata and temporal indices, with no access to test stations or future horizons.

TABLE III
DOWNSTREAM FORECASTING PERFORMANCE (MSE ↓) ON ISD-TEST UNDER THE FEW-SHOT (10% REAL DATA) SETTING WITH TWO DISTINCT BACKBONES: A STANDARD TRANSFORMER AND TIMESNET.

Training Data Source	Backbone: Transformer				Backbone: TimesNet			
	96	192	336	720	96	192	336	720
Real Only (10% data)	0.342	0.385	0.441	0.512	0.285	0.324	0.388	0.442
+ TimeVAE [29]	0.335	0.372	0.430	0.498	0.278	0.315	0.375	0.430
+ Diffusion-TS [14]	0.318	0.355	0.412	0.475	0.262	0.301	0.358	0.412
+ T2S [22]	0.305	0.340	0.395	0.455	0.255	0.292	0.345	0.398
+ MTransformer (Ours)	**0.272**	**0.305**	**0.352**	**0.408**	**0.238**	**0.274**	**0.322**	**0.375**
Reference: Real (100%)	*0.255*	*0.288*	*0.335*	*0.390*	*0.215*	*0.252*	*0.304*	*0.358*

Augmentation for Data-Sparse Forecasting. Table III reports forecasting MSE under a few-shot setting using only 10% of real data. MTransformer consistently outperforms all competitive generative baselines across horizons and backbones. The advantage becomes more pronounced for long-term forecasting, indicating that the spectral-aware design effectively preserves low-frequency global trends that are often degraded in diffusion-based samples. MTransformer also surpasses T2S, demonstrating that explicit frequency constraints provide more reliable physical consistency than implicit text embeddings.

Scalable Pre-training for Foundation Models. We further evaluate MeteoCap-3B as a pre-training corpus. As shown in Table IV, we pre-train an Transformer on the synthetic data and evaluate it under both Zero-Shot Transfer to unseen stations and Full Fine-Tuning. The multimodal pre-training strategy using aligned series–caption pairs achieves a 22.8% reduction in MSE compared to training from scratch, demonstrating substantially improved cross-domain generalization. In contrast, unimodal pre-training yields only marginal gains, indicating that textual semantics provide effective regularization and that the synergy between text and time series is crucial for learning transferable representations.

TABLE IV
IMPACT OF PRE-TRAINING MODALITY ON DOWNSTREAM PERFORMANCE UNDER ZERO-SHOT TRANSFER AND FULL FINE-TUNING (FT).

Pre-training	Zero-Shot		Full FT		Avg. Gain	
	MSE	MAE	MSE	MAE	ZS	FT
Scratch (Random Init.)	0.612	0.525	0.285	0.318	-	-
Unimodal Pre-train (Series Only)	0.545	0.482	0.264	0.295	+10.1%	+7.3%
Multimodal Pre-train (Ours)	**0.465**	**0.412**	**0.235**	**0.268**	**+22.8%**	**+16.4%**

D. Framework Analysis

Semantic Controllability Evaluation. Fig. 6 reports counterfactual prompting results on 500 test cases, evaluating semantic controllability beyond distribution memorization. Using Attribute Alignment Accuracy (AAA) that verified via trend slopes, FFT peak shifts, and variance changes, MTransformer consistently achieves high alignment across all attributes, while T2S shows low sensitivity to textual control and GPT-4o struggles with precise long-horizon periodicity. These show that spectral prompting reliably translates frequency-level semantic instructions into corresponding temporal behaviors.

Ablation Study. Table V (Part A) shows that each component is critical for performance on the Meteo-Volatile subset. Removing MSPG causes the largest degradation, which indicates that raw text embeddings fail to capture high-frequency variability without explicit spectral inductive bias. Replacing SPCA with standard cross-attention further increases

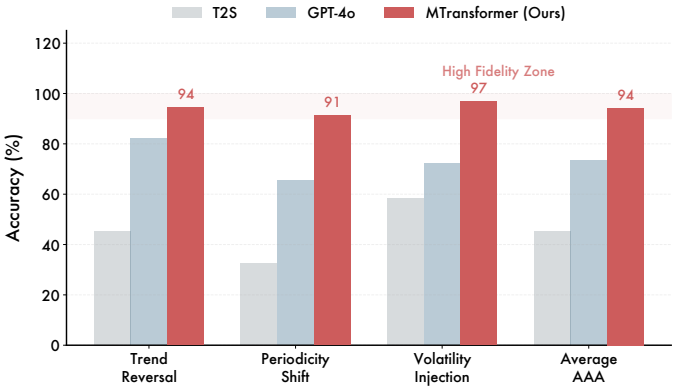

Fig. 6. Controllability of MTransformer generated sequences.

MSE, highlighting its role in aligning temporal tokens with frequency-domain prompts. Disabling the scheduler leads to a smaller but consistent drop, suggesting that the weak-to-strong spectral curriculum stabilizes early-stage optimization.

Hyperparameter Sensitivity. Part B of Table V examines the effect of spectral granularity. Increasing the number of frequency bands from $B = 1$ to $B = 8$ consistently improves performance, indicating better disentanglement of multi-scale weather patterns, while further increasing B to 16 yields no additional benefit and likely introduces redundant high-frequency noise. Similarly, performance saturates at $M = 16$ spectral tokens, with $M = 32$ providing only marginal WAPE improvement at a higher computational cost. Based on this trade-off, we adopt $B = 8$ and $M = 16$ as the default setting.

TABLE V
ABLATION STUDY AND HYPERPARAMETER SENSITIVITY ANALYSIS ON THE METEO-VOLATILE SUBSET ($L = 96$, MSE, WAPE REPORT).

	Model Variants	MSE ↓	WAPE ↓
Part A: Ablation	**Full Model (Ours)**	**0.028**	**0.265**
	w/o Spectral Scheduler (Fixed $\lambda = 1$)	0.030	0.278
	w/o SPAC (Standard Cross-Attn)	0.033	0.292
	w/o MSPG (Raw Text Embed.)	0.038	0.325
	w/o Text Cond. (Base DiT)	0.045	0.368
	Parameter Settings	**MSE ↓**	**WAPE ↓**
Part B: Hyperparameter	*Number of Frequency Bands (B)*		
	$B = 1$ (Global)	0.035	0.312
	$B = 4$	0.029	0.275
	$B = 8$ (Default)	**0.028**	**0.265**
	$B = 16$	0.029	0.268
	Spectral Tokens per Band (M)		
	$M = 4$	0.032	0.288
	$M = 8$	0.029	0.271
	$M = 16$ (Default)	**0.028**	**0.265**
	$M = 32$	0.028	0.264

E. Validation of Captioning Pipeline

To validate the effectiveness of our captioning pipeline, we conduct a controlled human study on 400 randomly sampled instances. Captions generated by GPT-4o, Claude-4.5, and our MACC system are independently evaluated in a blind setting by eight PhD-level meteorology experts using three criteria: Hallucination Rate, measuring factual contradictions; Physical Consistency, rated on a 1–5 scale; and Informativeness, also rated on a 1–5 scale. MACC reduces hallucinations to 3.8% while achieving near-perfect physical consistency and high informativeness, substantially outperforming all baselines. These results confirm that MACC delivers genuine gains in factual reliability and physical validity beyond simple API stacking.

TABLE VI
HUMAN EXPERT EVALUATION OF CAPTIONING QUALITY. VALUES ARE REPORTED AS MEAN ± STD.

Method	Hallucination Rate (↓)	Phys. Consistency (↑)	Informativeness (↑)
GPT-4o	24.5%	3.12 ± 0.45	3.45 ± 0.38
Claude-4.5	18.2%	3.35 ± 0.41	3.68 ± 0.42
MACC (Ours)	**3.8%**	**4.82 ± 0.15**	**4.65 ± 0.22**

F. RQ4: Cross-modal Alignment and Retrieval Analysis

To verify whether MeteoCap-3B and our MTransformer effectively bridge the semantic gap between meteorological signals and natural language, we conduct bi-directional cross-modal retrieval experiments. This task evaluates the model's ability to identify fine-grained correspondences between temporal dynamics and textual descriptions.

Setup and Baselines. We conduct retrieval on the held-out test split of MeteoCap-3B under two settings: **(1) Text-to-Series**, where a caption query retrieves time series via cosine similarity in the latent space, and **(2) Series-to-Text**, which performs the inverse retrieval. We report Recall@K ($K = 1, 5$) and MRR, and further evaluate on Meteo-Volatile to assess robustness under extreme conditions. We compare against three categories: **(1) Heuristics**, including *Stat-Feature Matching* (Euclidean distance on statistical moments) and *DTW-Matching* [34], which computes DTW distance between the query and a GPT-4o-generated pseudo-series; **(2) Trained Baselines**, including *Time-MMD* [11] (transfer) and *TS-CLIP* [35]; and **(3) LLM Adapters**, represented by *LLM-Adapter* built on frozen Llama-3 [31] with adapters [36].

Results. Table VII reports the results. MTransformer achieves the best performance across all metrics, outperforming the strongest baseline, LLM-Adapter, by **+10.0% in R@1**, with larger gains on **Meteo-Volatile**. TS-CLIP trained on MeteoScript-3B attains moderate performance while heuristic methods remain limited, highlighting the inadequacy of purely statistical or shape-based matching for meteorological seman-

TABLE VII
BI-DIRECTIONAL CROSS-MODAL RETRIEVAL PERFORMANCE ON METEOCAP-3B TEST SET. WE COMPARE AGAINST HEURISTIC METHODS, TRAINED BASELINES, AND LLM-BASED ADAPTERS. METEO-VOLATILE SUBSET TESTS ALIGNMENT UNDER EXTREME CONDITIONS.

Method	Overall Test Set (All Subsets)						Meteo-Volatile Subset (Extreme Events)					
	Text-to-Series			Series-to-Text			Text-to-Series			Series-to-Text		
	R@1	R@5	MRR	R@1	R@5	MRR	R@1	R@5	MRR	R@1	R@5	MRR
Heuristic Baselines												
Random Guess	0.01	0.05	0.001	0.01	0.05	0.001	0.01	0.05	0.001	0.01	0.05	0.001
Stat-Feature Matching	8.5	22.1	0.145	8.2	21.5	0.142	4.2	12.8	0.085	3.8	11.5	0.080
DTW-Matching	12.4	26.8	0.185	11.5	25.2	0.178	6.8	15.5	0.110	6.2	14.2	0.105
Neural Baselines												
Time-MMD	18.5	41.2	0.285	17.6	39.5	0.274	12.4	28.5	0.195	11.8	27.2	0.188
TS-CLIP	42.5	68.4	0.535	41.2	66.8	0.522	35.6	60.5	0.455	34.2	58.8	0.442
LLM-Adapter	48.2	73.5	0.585	47.5	71.8	0.572	40.5	65.2	0.495	39.2	63.8	0.482
MTransformer (Ours)	**58.2**	**81.4**	**0.675**	**56.9**	**79.8**	**0.662**	**52.5**	**76.2**	**0.615**	**50.8**	**74.5**	**0.598**
w/o Spectral Prompt	45.8	72.1	0.565	44.2	70.5	0.552	38.5	64.2	0.485	37.2	62.5	0.468

tics. The notable degradation without Spectral Prompts further underscores the importance of frequency-domain alignment.

IV. Conclusion

In conclusion, we present a unified framework for text-guided meteorological time-series generation, comprising the **MeteoCap-3B** dataset constructed via a multi-agent collaborative pipeline and the MTransformer model. By integrating physically grounded multimodal supervision with explicit spectral conditioning, our approach achieves high-fidelity generation, accurate cross-modal alignment, and strong semantic controllability, while consistently improving downstream forecasting performance under data-sparse and zero-shot settings.

References

[1] Z. Wang, J. Bao, S. Gu, D. Chen, W. Zhou, and H. Li, "Designdiffusion: High-quality text-to-design image generation with diffusion models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 20 906–20 915.

[2] Y. Jia, Y. Chen, J. Zhao, S. Zhao, W. Zeng, Y. Chen, and Y. Qin, "Audioeditor: A training-free diffusion-based audio editing framework," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[3] J. Zheng, S. Qiu, C. Shi, and Q. Ma, "Towards lifelong learning of large language models: A survey," *ACM Computing Surveys*, vol. 57, no. 8, pp. 1–35, 2025.

[4] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng, and S. Z. Li, "A survey on generative diffusion models," *IEEE transactions on knowledge and data engineering*, vol. 36, no. 7, pp. 2814–2830, 2024.

[5] S. Chen, G. Long, J. Jiang, and C. Zhang, "Federated foundation models on heterogeneous time series," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 15, 2025, pp. 15 839–15 847.

[6] S. Chen, G. Long, T. Shen, and J. Jiang, "Prompt federated learning for weather forecasting: toward foundation models on meteorological data," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 2023, pp. 3532–3540.

[7] S. Chen, G. Long, J. Jiang, and C. Zhang, "Personalized adapter for large meteorology model on devices: Towards weather foundation models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 84 897–84 943, 2024.

[8] S. Chen, G. Long, T. Shen, J. Jiang, and C. Zhang, "Federated prompt learning for weather foundation models on devices," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 2024, pp. 5772–5780.

[9] M. A. Morid, O. R. L. Sheng, and J. Dunbar, "Time series prediction using deep learning methods in healthcare," *ACM Transactions on Management Information Systems*, vol. 14, no. 1, pp. 1–29, 2023.

[10] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in neural information processing systems*, vol. 35, pp. 25 278–25 294, 2022.

[11] H. Liu, S. Xu, Z. Zhao, L. Kong, H. Prabhakar Kamarthi, A. Sasanur, M. Sharma, J. Cui, Q. Wen, C. Zhang *et al.*, "Time-mmd: Multi-domain multimodal dataset for time series analysis," *Advances in Neural Information Processing Systems*, vol. 37, pp. 77 888–77 933, 2024.

[12] C. Chang, J. Hwang, Y. Shi, H. Wang, W.-C. Peng, T.-F. Chen, and W. Wang, "Time-imm: A dataset and benchmark for irregular multimodal multivariate time series," *arXiv preprint arXiv:2506.10412*, 2025.

[13] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[14] X. Yuan and Y. Qiao, "Diffusion-ts: Interpretable diffusion for general time series generation," *arXiv preprint arXiv:2403.01742*, 2024.

[15] L. Lin, Z. Li, R. Li, X. Li, and J. Gao, "Diffusion models for time-series applications: a survey," *Frontiers of Information Technology & Electronic Engineering*, vol. 25, no. 1, pp. 19–41, 2024.

[16] H. Li, Y. Huang, C. Xu, V. Schlegel, R. Jiang, R. Batista-Navarro, G. Nenadic, and J. Bian, "Bridge: Bootstrapping text to control time-series generation via multi-agent iterative optimization and diffusion modelling," *arXiv preprint arXiv:2503.02445*, 2025.

[17] S. Chen, T. Shu, H. Zhao, G. Zhong, and X. Chen, "Tempee: Temporal–spatial parallel transformer for radar echo extrapolation beyond autoregression," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–14, 2023.

[18] K. Yi, J. Fei, Q. Zhang, H. He, S. Hao, D. Lian, and W. Fan, "Filternet: Harnessing frequency filters for time series forecasting," *Advances in Neural Information Processing Systems*, vol. 37, pp. 55 115–55 140, 2024.

[19] S. Chen, G. Long, J. Jiang, D. Liu, and C. Zhang, "Foundation models for weather and climate data understanding: A comprehensive survey," *arXiv preprint arXiv:2312.03014*, 2023.

[20] S. Chen, G. Long, M. Blumenstein, and J. Jiang, "Fedal: Federated dataset learning for general time series foundation models," in *The Fourteenth International Conference on Learning Representations*.

[21] S. Chen, T. Shu, H. Zhao, Q. Wan, J. Huang, and C. Li, "Dynamic multiscale fusion generative adversarial network for radar image extrapolation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–11, 2022.

[22] Y. Ge, J. Li, Y. Zhao, H. Wen, Z. Li, M. Qiu, H. Li, M. Jin, and S. Pan, "T2s: High-resolution time series generation with text-to-series diffusion models," *arXiv preprint arXiv:2505.02417*, 2025.

[23] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan *et al.*, "Deepseek-v3 technical report," *arXiv preprint arXiv:2412.19437*, 2024.

[24] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," in *The eleventh international conference on learning representations*, 2022.

[25] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[26] A. Smith, N. Lott, and R. Vose, "The integrated surface database: Recent developments and partnerships," *Bulletin of the American Meteorological Society*, vol. 92, no. 6, pp. 704–708, 2011.

[27] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4195–4205.

[28] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[29] A. Desai, C. Freeman, Z. Wang, and I. Beaver, "Timevae: A variational auto-encoder for multivariate time series generation," *arXiv preprint arXiv:2111.08095*, 2021.

[30] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[31] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[32] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "Timesnet: Temporal 2d-variation modeling for general time series analysis," *arXiv preprint arXiv:2210.02186*, 2022.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[34] L. Wang and P. Koniusz, "Uncertainty-dtw for time series and sequences," in *European Conference on Computer Vision*. Springer, 2022, pp. 176–195.

[35] Z. Chen, X. Zhang, and M. Zhu, "Ts-clip: Time series understanding by clip," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 4646–4664.

[36] Z. Hu, L. Wang, Y. Lan, W. Xu, E.-P. Lim, L. Bing, X. Xu, S. Poria, and R. Lee, "Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models," in *Proceedings of the 2023 conference on empirical methods in natural language processing*, 2023, pp. 5254–5276.