UNIGIO: UNIFIED GENERATIVE GLOBAL IN-SITU WEATHER MODELING FROM SPATIOTEMPORAL INCOMPLETE OBSERVATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Global In-situ Observation (GIO) provides fine-scale, pointwise weather measurements beyond gridded satellite products, supporting numerical weather prediction and disaster prevention. However, severe spatiotemporal incompleteness in GIO leads existing methods to rely on AI-ready preprocessing or massive global multi-source data, limiting accurate, timely, and efficient in-situ weather services. To address this gap, we propose UniGIO, a unified generative framework that exploits station-centered observational complementarity and decouples local extremes from steady evolution, unifying forecasting, imputation, and generation directly from incomplete GIO. UniGIO progressively builds complementarity from discrete station interactions to continuous local fields. It captures missing-aware sequential evolution in a shared state space, while separate expert representations support extreme-pattern adaptation and global pattern recognition. These mechanisms form a lightweight CVAE suitable for distributed stations, explicitly modeling missing-state uncertainty for extreme-event assessment and generation. Experiments on Weather-5K, the largest up-to-date GIO dataset, demonstrate SOTA performance, with average gains of 11%, 12%, and 5% in accuracy, fidelity, and extreme-event capture, advancing unified weather modeling from incomplete station observations.

1 INTRODUCTION

Global In-situ Observation (GIO) provides timely, fine-scale point observations of meteorological variables, capturing local weather evolution, volatility, and extreme events while supporting numerical weather prediction, aviation, agriculture, and disaster monitoring (WMO, 2018; Zhu et al., 2026), and has even been applied in key scenarios such as the Winter Olympics (Wu et al., 2023). Unlike heavy grid-based reanalysis and satellite data that rely on interpolation to sub-grid scales and smooth transient and extreme signals (Bi et al., 2023; Lam et al., 2023; Price et al., 2025; Vaughan et al., 2025; Ozdemir et al., 2026), GIO directly records lightweight sub-grid atmospheric dynamics, making it a primary source for local weather services and a critical bridge from global forecasts to fine-scale local modeling (Kalnay, 2003; Bauer et al., 2015; Rasp et al., 2026; Vaughan et al., 2025).

Nevertheless, GIO is inherently spatiotemporally incomplete (Cao et al., 2018; Tashiro et al., 2021). As shown in Fig. 1 (a), Weather-5K (Han et al., 2026) has a 27.5% missing rate across 5k+ stations over 10 years, while only about 25%–10% of 24h–120h windows are AI-ready, due to sensor or transmission failures, uneven station distribution, and heterogeneous temporal resolutions (WMO, 2018; 2021a). Such persistent gaps make GIO poorly compatible with existing AI frameworks.

Existing methods face three limitations: Forecasting models typically assume complete inputs, so separate imputation brings time delay and propagates errors (Wu et al., 2023; Yang et al., 2025a). Generative models support flexible masks, but shared representations for smooth and extreme dynamics and weak observation conditioning can cause over-smoothing or spurious fluctuations (Yuan & Qiao, 2024; Yang et al., 2025b). Observation foundation models support predictions at arbitrary locations, but deriving local outputs from global representations through interpolation can smooth fine-scale variability rather than directly model in-situ dynamics. Their reliance on extensive multi-source inputs further adds data and computational demands, hindering deployment in resource-constrained in-situ services (Gong et al., 2026; Vaughan et al., 2025; Ozdemir et al., 2026; Rasp et al., 2026).

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

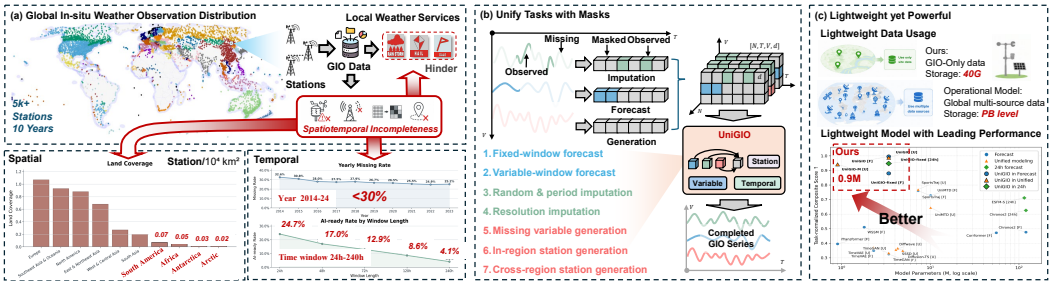

Figure 1: (a) GIO incompleteness: sparse station coverage, high missing rates, and low AI-ready rates (the fraction of fully observed fixed-length windows). (b) Mask-driven generation unifies GIO tasks. (c) UniGIO achieves leading performance with limited training data and few parameters.

Beyond accuracy degradation, incomplete GIO can obscure local extremes, which account for 90% global extreme events and cause 70% of mortality and economic losses (WMO, 2021b; UNDRR, 2022). This risk is particularly severe in less-developed regions such as Africa, where station density is only $\sim 1/84$ of that in Germany and 22% of stations meet the GBON standard (WMO, 2019; WMO; Rogers et al., 2019; WMO, 2021a). Since stations are also more failure-prone during extremes, reliably capturing abrupt events from incomplete observations is critical (WMO, 2024; 2021b).

These challenges motivate UniGIO, a unified framework that directly models distributed station observations without spatial interpolation. It unifies forecasting, imputation, and generation as mask-driven observation-to-missing generation (Fig. 1(b)). To avoid the high data and computational cost of global multi-source complementarity for station-level services, we introduce station-centered complementarity, realized by an Observation Mixer (OM) that progressively organizes fragmented evidence across variables and neighboring stations into continuous local fields to constrain missing states. To resolve the conflicting representation requirements of stable evolution and rare abrupt extremes, we introduce evolution–extreme pattern decoupling, realized by an Adaptive Temporal Mixer (ATM) that combines reliability-weighted sequential modeling with specialized pattern experts for global pattern recognition and local extreme adaptation. For lightweight probabilistic modeling directly at stations, we adopt a CVAE (Sohn et al., 2015) that conditions generation on available observations while explicitly modeling uncertainty at missing positions, supporting extreme-event assessment and controllable generation. UniGIO learns from 40GB of sparse GIO rather than petabyte-scale global observations and can be compressed to fewer than 1M parameters while retaining leading performance (Fig. 1(c)).

We conducted quantitative and qualitative experiments under both native and curated spatiotemporal incompleteness to benchmark the in-situ weather modeling task on the up-to-date largest GIO dataset Weather-5K (Han et al., 2026), recording hourly temperature, dew point, wind direction, wind rate, and sea-level pressure from 5k+ surface weather stations worldwide, covering 6 continents and 10 years. UniGIO uniformly handles 7 tasks, including fix/variable-window forecasting, random&period/resolution imputation, variable/in-region/cross-region station generation, and outperforms baselines by an average 9% across 7 metrics, respectively 11%, 12%, and 5% improvement on accuracy, fidelity, and extreme-event capture. Our contributions are as follows:

- We propose UniGIO, a novel unified framework for versatile and lightweight global in-situ weather modeling from incomplete GIO through an observation-to-missing generation to advance all-time in-situ service, even when real-world stations are absent or inoperable.
- We develop station-centered complementarity and evolution–extreme pattern decoupling for incomplete GIO. OM builds local context around stations to jointly constrain missing states, while ATM separates reliable sequential dynamics from extreme-pattern adaptation to reduce interference. Explicit uncertainty supports extreme-event assessment and generation.
- We benchmark across 5k+ surface weather stations worldwide, covering a 10-year period to establish strong baselines for in-situ weather modeling. Extensive experiments validate our consistent and outstanding performance across tasks, regions, and mask formats.

A detailed discussion of related work is provided in Appendix A.

2 METHOD

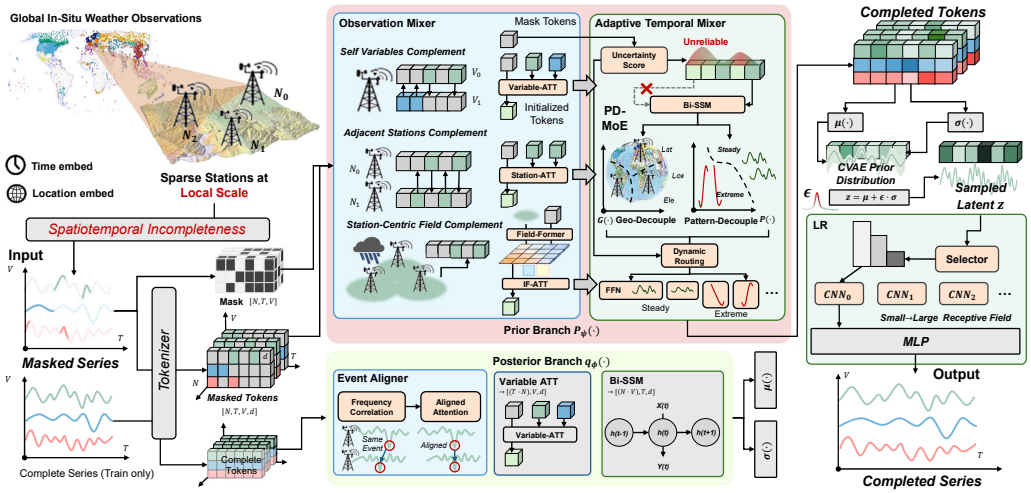

Figure 2: Illustration of UniGIO, with key designs for station-centered complementarity in blue and evolution–extreme decoupling in green, all unified within a light weight CVAE framework.

2.1 TASK FORMULATION AND FRAMEWORK

We represent weather observations from N neighboring stations over T time steps and C variables as $\mathbf{X} \in \mathbb{R}^{N\times T\times C}$. An observation mask $\mathbf{M}$ records whether each entry is available: $M_{n,t,c} = 1$ for observed values and 0 otherwise. Given incomplete input $\mathbf{X}_M$, the mask, and spatiotemporal metadata $\mathbf{I}$, UniGIO generates the missing entries from available observations. Different masks express forecasting, imputation, and variable or station generation within the same conditional modeling problem. The seven task-specific mask constructions are given in Appendix B.

As illustrated in Fig. 2, UniGIO is organized as a conditional variational autoencoder (CVAE) with two main stages. First, the Observation Mixer (OM) propagates evidence across variables, stations, and local fields to initialize unobserved positions. Second, the Adaptive Temporal Mixer (ATM) combines reliability-weighted state-space modeling with pattern-decoupled experts to capture smooth evolution and abrupt changes. The Local Refiner (LR) restores local continuity during reconstruction. An Event Aligner (EA) provides aligned temporal evidence in the posterior branch during training. The conditional prior operates on incomplete observations at inference. The task and branch definitions are detailed in Appendices D.1 and D.2; Algorithm C.1 gives the training and inference pipeline.

2.2 OBSERVATION MIXER AND EVENT ALIGNER

Complementarity across variables, stations, and fields. A missing observation cannot always be recovered from its broken temporal sequence alone. OM first applies self-attention across variables within each station, then distance-based graph attention with A adjacency matrix $\mathbf{A}$ across neighboring stations, producing features $\mathbf{f}^t_{\text{adj}}$ at time t. These operations exchange evidence at discrete locations, but weather processes occupy spatially extended regions: sparse, unevenly distributed stations observe only fragments of the same evolving field. A shared field representation lets these fragments jointly constrain missing locations, rather than relying solely on pairwise station interactions.

$$\mathbf{f}^t_{\text{adj}} = \text{GraphAttention}_{\text{sta}}\big(\text{MHSA}_{\text{var}}(\mathbf{f}^t_{\text{vars}}), \mathbf{A}\big)\,. \tag{1}$$

OM therefore constructs a small local grid of field tokens $\mathbf{z}_{\text{field}}$, whose spatial encoding is shared with station features. Learnable memory tokens $\mathbf{z}_{\text{mem}}$ provide context shared across regional fields. These are latent queries, not interpolated weather measurements on a dense global grid. Let $\mathbf{Z} =$

$[\mathbf{z}_{\text{field}}, \mathbf{z}_{\text{mem}}]$, where brackets concatenate the token sets. Field complementarity is

$$\begin{aligned} \mathbf{Z}' &= \mathrm{Ff}\big(q = \mathbf{Z},\ k/v = \{\mathbf{f}^t_{\text{adj}}\}_{t=1}^{T}\big)\,, \\ \mathbf{f}_{\mathrm{C}} &= \mathrm{ICAtt}(q = \mathbf{f}_{\text{mask}},\ k/v = \mathrm{CNNs}(\mathbf{Z}'))\,. \end{aligned} \tag{2}$$

Here q and k/v denote attention queries and keys/values. Ff is the Field-former cross-attention operator that aggregates station evidence into updated field and memory tokens $\mathbf{Z}'$. CNNs applies convolutional refinement to this representation. ICAtt is inverse cross-attention: feature queries $\mathbf{f}_{\text{mask}}$ at missing positions retrieve the refined field evidence to produce complemented features $\mathbf{f}_{\mathrm{C}}$. This station-to-field-to-station path introduces regional continuity while retaining station-level outputs. Initializations, adjacency, and grid construction appear in Appendix D.3 and Fig. C.1(a).

Aligning inter-station events. The same weather event may reach neighboring stations at different times. Direct aggregation can mix different event phases, so EA estimates the delay between station sequences $\mathbf{X}_i$ and $\mathbf{X}_j$ by

$$D_{ij} = \arg\max_{\ell} \left[\mathcal{F}^{-1}\Big(\mathcal{F}(\mathbf{X}_i) \odot \overline{\mathcal{F}(\mathbf{X}_j)}\Big)\right]_{\ell}. \tag{3}$$

Here ℓ denotes temporal lag, $\mathcal{F}$ and $\mathcal{F}^{-1}$ are the Fourier transform and its inverse, $\bar{\cdot}$ denotes complex conjugation, and $\odot$ denotes elementwise multiplication. EA shifts neighboring features by the estimated delay D_{ij} to align event phases before attention aggregation. Since delay estimation requires complete sequences, EA is used only in the training-time posterior branch, where the aligned posterior guides conditional-prior learning rather than directly constraining inference. Details are provided in Appendix D.4 and Fig. C.1(b).

2.3 Adaptive Temporal Mixer and Local Refiner

Reliability-weighted temporal modeling. ATM combines a Pattern-Decoupled Mixture-of-Experts (PD-MoE) with state-space temporal modeling to separate sequential dependency modeling from global pattern adaptation. Since attention can become dominated by extreme values, we adopt a Mamba-inspired selective state-space module (Gu & Dao, 2024) to compress sequential evidence into a compact latent state, aiming to selectively filter perturbations and propagate informative temporal signals to preserve long-range temporal dependencies. On top of this robust representation, we introduce MoE-based pattern adaptation (Shazeer et al., 2017), where separate expert representations reduce interference between extreme events and smooth evolution.

For sequential dependencies, compression alone cannot prevent uncertain gap reconstructions from propagating into the recurrent state. Temporal features farther from valid observations receive weaker support and are more likely to contain reconstruction uncertainty. ATM therefore weights inputs before state updates using their distance $d_t = \min_{t' \in \mathcal{T}_{\text{obs}}} |t - t'|$ to the nearest valid observation, where $\mathcal{T}_{\text{obs}}$ is the observed temporal index set:

$$\mathbf{U}^t_s = e^{-\,\mathrm{ReLU}(\mathrm{MLP}(d_t))}, \qquad \mathbf{h}_{1:T} = \mathrm{BiMamba}\big(\{\mathbf{U}^t_s \odot \mathbf{f}^t_{\mathrm{C}}\}_{t=1}^{T}\big)\,. \tag{4}$$

Here $\odot$ denotes elementwise multiplication, and smaller reliability weights suppress weakly supported inputs more strongly. By filtering uncertain evidence before state updates, $\mathrm{BiMamba}$ learns more robust temporal representations for pre-expert features $\mathbf{h}_t$. Full operations are in Appendix D.5.

Pattern-decoupled expert adaptation. PD-MoE shares experts across temporal positions and stations to associate globally similar patterns. Routing also incorporates geography: similar fluctuations may require different responses across regions, while the same station may alternate between stable and extreme states. The learned maps $\mathrm{P}(\mathbf{h}_t) \in \mathbb{R}^{N_e \times d_p}$ and $\mathrm{G}(\mathbf{I}) \in \mathbb{R}^{S_e \times d_p}$ encode pattern and geographic affinity, respectively, where N_e is the number of experts, S_e the number of regional subspaces, and d_p their shared dimension:

$$\begin{aligned} [\mathbf{w}_g, \mathbf{b}_g] = \mathcal{R}_g(G(\mathbf{I})), \qquad \mathbf{r}_t = \mathrm{Softmax}\left(\frac{P(\mathbf{h}_t)\mathbf{w}_g}{\sqrt{d_p}} + \mathbf{b}_g\right), \\ E_t = \mathrm{Relative}(\mathbf{U}^t_s), \qquad \mathcal{E}_t = \mathrm{TopK}(\mathbf{r}_t, E_t), \qquad \mathbf{f}^t_{\mathrm{T}} = \sum_{e \in \mathcal{E}_t} r_{t,e}\, \mathrm{MLP}_e(\mathbf{h}_t). \end{aligned} \tag{5}$$

Here $\mathcal{R}_g$ denotes the geographic routing module, which aggregates geographic subspaces and generates the location-conditioned projection $\mathbf{w}_g$ and bias $\mathbf{b}_g$. The resulting $\mathbf{r}_t$ is the expert routing distribution jointly conditioned on temporal patterns and geography. $\mathrm{Relative}$ determines the adaptive expert count E_t from reliability $\mathbf{U}_s^t$, and TopK selects $\mathcal{E}_t$ according to $\mathbf{r}_t$. The selected routing weights $r_{t,e}$ aggregate expert outputs into the adapted feature $\mathbf{f}_{\mathrm{T}}^t$. Routing details are given in Appendix D.6 and Fig. C.2(a).

Local reconstruction. To restore continuity weakened by expert mixing and latent sampling, LR applies convolutions with multiple receptive fields to the sampled latent representation. A selective MLP gate conditioned on temporal features weights the concatenated convolution outputs, which the decoder combines with $\mathbf{f}_{\mathrm{T}}$ to reconstruct $\hat{\mathbf{X}}$. Full equations are in Appendix D.7.

2.4 Training Objective and Inference

UniGIO learns a posterior $q_\phi(\mathbf{z} \mid \mathbf{X}, \mathbf{I})$ and a conditional prior $p_\psi(\mathbf{z} \mid \mathbf{X}_M, \mathbf{M}, \mathbf{I})$. Here $\mathbf{z}$ is the latent variable and ϕ and ψ parameterize the posterior and prior. Its objective combines reconstruction, prior–posterior KL divergence, and orthonormal regularization:

$$\mathcal{L} = \mathcal{L}_{\mathrm{rec}} + \beta \mathcal{L}_{\mathrm{KL}} + \gamma \mathcal{L}_{\mathrm{orth}}. \tag{6}$$

The reconstruction term trains the decoder from posterior samples, the KL term aligns the conditional prior with the posterior, and the orthonormal term constrains the pattern and regional projections and memory-token weights. The coefficients β and γ weight the KL and orthonormal terms. The full objective is retained in Appendix D.8, Eq. D.17. During inference, UniGIO samples from the conditional prior without access to complete targets. Pointwise sampling strength controls latent perturbations, enabling the controllable generation procedure described in Algorithm C.1.

3 Experiments

3.1 Benchmark and Setup

Dataset and settings. We evaluate UniGIO on GIO-U, constructed from Weather-5K (Han et al., 2026), which contains hourly temperature, dew point, wind rate, wind direction, and sea-level pressure from 5,672 global stations during 2014-2024. We adopt a chronological 8/2 years training/testing split. We primarily use 120h windows, with additional 72h/240h evaluations, covering unified modeling, forecasting, imputation, generation, spatial generalization, and ablations.

Curated and native masks. For the masks used, curated masks artificially hide time steps, variables, and stations, whereas native masks reflect original missingness. Since genuinely missing values lack ground truth, we train and evaluate by masking observations, and then evaluate under native masks without further training. Curated masks also enable controlled task and missing-pattern comparisons.

Evaluation metrics. We measure accuracy with MAE/MSE, fidelity with Fréchet Inception Distance (FID), statistical error (STE), and Correlation Score (CroS), and extreme-event capture with F1-SEDI at the 99.5th/90.0th percentiles. F1-SEDI extends the recall-based formulation in prior work through a precision–recall harmonic mean, accounting for misses and false alarms. Circular wind direction is excluded from F1-SEDI.

Baselines. We compare TimeGAN, TimeVAE, DiffWave, SSSD, Diffusion-TS, and interaction-aware UniMTD/SportsTraj adapted to station and variable dependencies, alongside task-specific forecasters and observation/time series foundation models.

Benchmark, baseline, and metric details appear in Appendix E.

3.2 Overall Performance

3.2.1 Quantitative Results

UniGIO achieves state-of-the-art performance across most of the seven metrics and five variables under both curated and native masks (Fig. 3(a); Tables 1–3). At 120h, it outperforms the strongest competing baselines by 12.1%/10.8% in accuracy, 3.5%/6.5% in F1-SEDI, and 11.1%/12.6% in

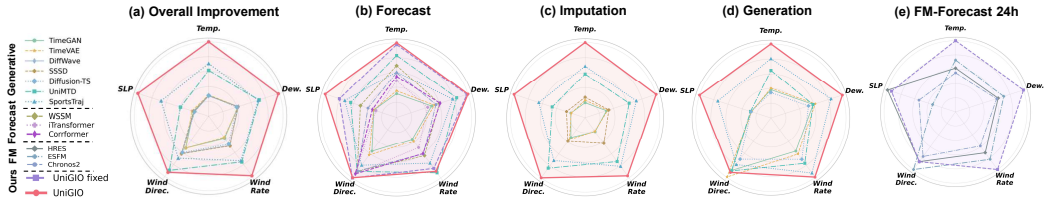

Figure 3: UniGIO and UniGIO-fixed form the outer envelopes of the radar plots, demonstrating SOTA performance in (a) overall across time windows, (b)-(d) forecasting, imputation, and generation, (e) 24-hour forecasting against foundation models. Larger areas indicate better performance.

Table 1: Quantitative Accuracy Comparison on Curated / Native Masks (120h) † means re-implemented

Baselines	Temperature		Dewpoint		Wind Rate		Wind Direc.		Sea-Level Pressure	
	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓ ($\times 10^3$)	MAE↓	MSE↓
Deterministic	4.54 / 4.32	25.5 / 26.8	3.37 / 3.54	16.7 / 17.5	2.10 / 2.16	6.19 / 5.98	92.8 / 97.4	10.5 / 11.0	4.77 / 4.96	34.8 / 33.1
TimeGAN	5.51 / 10.3	37.8 / 76.5	3.91 / 6.98	22.4 / 41.5	2.51 / 4.33	8.31 / 14.7	103 / 147	12.5 / 17.9	5.77 / 8.69	49.8 / 78.5
TimeVAE	4.58 / 5.36	25.6 / 29.9	3.43 / 4.02	16.8 / 19.7	2.09 / 2.45	6.04 / 7.08	96.9 / 114	10.6 / 12.4	5.16 / 6.04	37.5 / 43.9
Diffwave	5.14 / 5.76	34.8 / 38.8	3.74 / 4.17	21.5 / 24.7	2.45 / 2.48	8.84 / 8.76	105 / 119	14.5 / 15.9	4.99 / 5.14	40.6 / 43.9
SSSD	4.98 / 5.39	33.1 / 37.9	3.88 / 4.25	23.9 / 27.0	2.37 / 2.69	8.65 / 9.52	107 / 111	15.4 / 16.1	5.00 / 5.78	41.6 / 48.1
Diffusion-TS	5.34 / 6.11	37.5 / 42.9	4.05 / 4.58	24.9 / 27.9	2.40 / 2.77	8.39 / 9.41	108 / 122	14.7 / 16.3	5.71 / 6.19	53.2 / 59.8
UniMTD †	2.69 / 2.32	15.0 / 13.1	2.70 / 2.30	16.6 / 13.7	1.68 / 1.36	5.98 / 4.82	81.1 / 63.8	10.7 / 7.47	3.69 / 3.03	33.8 / 26.3
SportsTraj †	2.05 / 2.43	7.75 / 9.36	1.97 / 2.31	7.93 / 9.70	1.63 / 1.81	4.92 / 5.32	80.1 / 80.6	10.6 / 8.72	1.44 / 1.61	4.37 / 4.43
UniGIO	1.45 / 1.51	5.35 / 6.01	1.50 / 1.45	5.99 / 6.06	1.27 / 1.22	3.63 / 3.61	64.6 / 60.1	8.45 / 7.04	1.02 / 0.91	3.67 / 1.83

Table 2: Quantitative F1-SEDI Comparison on Curated / Native Masks (120h) † means re-implemented

Baselines	Temperature		Dewpoint		Wind Rate		Wind Direc.		Sea-Level Pressure	
	99.5th↑	90.0th↑	99.5th↑	90.0th↑	99.5th↑	90.0th↑	99.5th↑	90.0th↑	99.5th↑	90.0th↑
Deterministic	0.07 / 0.07	0.35 / 0.33	0.12 / 0.11	0.50 / 0.53	0.01 / 0.01	0.10 / 0.20	\	\	0.08 / 0.08	0.40 / 0.38
TimeGAN	0.10 / 0.19	0.46 / 0.48	0.21 / 0.28	0.58 / 0.60	0.03 / 0.06	0.19 / 0.26	\	\	0.19 / 0.19	0.50 / 0.51
TimeVAE	0.11 / 0.12	0.47 / 0.47	0.19 / 0.19	0.60 / 0.60	0.02 / 0.02	0.14 / 0.14	\	\	0.13 / 0.13	0.52 / 0.51
Diffwave	0.12 / 0.12	0.39 / 0.40	0.19 / 0.19	0.53 / 0.54	0.10 / 0.10	0.22 / 0.22	\	\	0.15 / 0.15	0.44 / 0.44
SSSD	0.12 / 0.13	0.40 / 0.42	0.20 / 0.21	0.53 / 0.55	0.12 / 0.12	0.25 / 0.26	\	\	0.15 / 0.16	0.44 / 0.46
Diffusion-TS	0.12 / 0.13	0.42 / 0.44	0.20 / 0.21	0.53 / 0.55	0.07 / 0.08	0.21 / 0.22	\	\	0.17 / 0.18	0.45 / 0.48
UniMTD †	0.34 / 0.44	0.66 / 0.72	0.28 / 0.43	0.60 / 0.67	0.12 / 0.21	0.32 / 0.44	\	\	0.35 / 0.47	0.59 / 0.65
SportsTraj †	0.46 / 0.50	0.76 / 0.78	0.41 / 0.47	0.72 / 0.76	0.26 / 0.31	0.38 / 0.50	\	\	0.73 / 0.74	0.85 / 0.86
UniGIO	0.55 / 0.61	0.81 / 0.83	0.46 / 0.56	0.75 / 0.80	0.22 / 0.36	0.39 / 0.54	\	\	0.79 / 0.83	0.88 / 0.91

Table 3: Quantitative Fidelity Comparison on Curated / Native Masks (120h) † means re-implemented

Baselines	Temperature		Dewpoint		Wind Rate		Wind Direc.		Sea-Level Pressure		CroS↓
	FID↓	STE↓	FID↓	STE↓	FID↓	STE↓	FID↓	STE↓ ($\times 10^3$)	FID↓	STE↓	
Deterministic	4.76 / 4.95	6.21 / 6.42	4.73 / 4.50	2.16 / 2.05	4.54 / 4.31	1.66 / 1.74	3.88 / 4.08	3.98 / 4.18	5.40 / 5.13	1.58 / 1.66	3.20 / 3.36
TimeGAN	1.22 / 0.95	5.53 / 5.38	1.11 / 0.89	2.32 / 2.29	1.08 / 0.78	1.50 / 1.52	0.95 / 0.72	3.80 / 3.80	2.97 / 2.51	2.53 / 2.51	3.10 / 3.17
TimeVAE	1.28 / 1.20	6.57 / 6.22	1.88 / 1.78	2.42 / 2.31	1.64 / 1.56	1.80 / 1.71	1.38 / 1.31	4.18 / 3.94	3.37 / 3.20	1.67 / 1.59	4.45 / 4.33
Diffwave	0.92 / 0.87	6.01 / 5.69	0.53 / 0.50	3.04 / 2.89	0.46 / 0.44	1.81 / 1.72	0.67 / 0.75	4.24 / 4.03	2.21 / 2.11	3.02 / 2.87	3.13 / 3.14
SSSD	0.82 / 0.78	6.05 / 5.75	0.51 / 0.48	3.77 / 3.58	0.48 / 0.36	2.01 / 1.91	0.64 / 0.71	4.65 / 4.42	1.28 / 1.20	2.54 / 2.41	3.16 / 3.11
Diffusion-TS	0.78 / 0.73	6.08 / 5.78	0.67 / 0.63	3.26 / 3.07	0.43 / 0.40	1.54 / 1.46	0.65 / 0.79	3.90 / 3.71	1.72 / 1.63	3.67 / 3.44	3.64 / 3.59
UniMTD †	0.50 / 0.27	3.34 / 2.16	0.30 / 0.27	1.64 / 1.23	0.44 / 0.35	1.08 / 0.78	0.52 / 0.68	2.73 / 1.83	0.42 / 0.34	1.38 / 0.96	2.44 / 1.42
SportsTraj †	0.53 / 0.26	3.27 / 2.37	0.77 / 0.39	2.26 / 1.69	0.87 / 0.47	1.33 / 1.00	1.44 / 0.78	3.08 / 2.07	0.33 / 0.25	1.30 / 1.00	1.52 / 0.93
UniGIO	0.17 / 0.13	2.48 / 1.67	0.25 / 0.21	1.45 / 1.02	0.28 / 0.19	1.07 / 0.68	0.60 / 0.52	2.53 / 1.67	0.11 / 0.09	0.79 / 0.51	1.46 / 0.81

Table 4: Parameter Comparison of Unified and Task-specific GIO models.

Method	TimeGAN	TimeVAE	Diffwave	SSSD	Diffusion-TS	UniMTD	SportsTraj	UniGIO-M	UniGIO	ESFM-S	Chronos2	WSSM	iTransformer	Corrformer
Params.	3.4M	2.3M	4.1M	4.3M	4.9M	10.1M	7.3M	0.9M	3.4M	115M	120M	1.8M	0.9M	55.6M

fidelity under curated/native masks, respectively. Consistent gains on native masks without additional training support transfer from constructed missing patterns to real-world incompleteness.

The deterministic variant tends toward mean predictions, while general time series generators exhibit either over-smoothing or spurious fluctuations. Interaction-aware trajectory models provide stronger baselines, and the CVAE-based approach outperforms its diffusion counterpart on F1-SEDI. These comparisons highlight the importance of observation conditioning and station complementarity for preserving both stable dynamics and extremes.

UniGIO also maintains its advantage at 72h and 240h (Tables F.1–F.3), improving accuracy/F1-SEDI/fidelity by 8.8%/13.5%, 3.6%/3.6%, and 7.1%/5.3% at 72h, with larger gains of 11.1%/10.0%, 5.1%/6.0%, and 21.4%/16.4% at 240h under curated/native masks. Although most metrics deteriorate with longer windows, CroS improves, suggesting that richer temporal context helps estimate inter-variable relationships and reduces sensitivity to short-term fluctuations.

Table 5: Quantitative comparison under variable-length, fixed-length, and noisy forecasting settings. † denotes re-implemented methods, and ⋆ denotes forecasting with an imputed look-back window.

Baselines	Temperature		Dewpoint		Wind Rate		Wind Direc.		Sea-Level Pressure	
	MAE↓/99.5th↑	MSE↓/90.0th↑	MAE↓/99.5th↑	MSE↓/90.0th↑	MAE↓/99.5th↑	MSE↓/90.0th↑	MAE↓/99.5th↑	MSE↓ ($\times 10^3$)/90.0th↑	MAE↓/99.5th↑	MSE↓/90.0th↑
Variable-Length Forecasting										
TimeGAN	7.40 / 0.11	53.7 / 0.46	5.15 / 0.19	30.7 / 0.59	3.23 / 0.02	10.8 / 0.18	130.6 /	16.00 /	7.45 / 0.17	65.8 / 0.50
TimeVAE	5.85 / 0.10	32.77 / 0.47	4.38 / 0.17	21.51 / 0.61	2.65 / 0.02	7.56 / 0.13	123.9 /	13.6 /	6.52 / 0.12	46.9 / 0.51
UniMTD †	2.35 / 0.40	11.3 / 0.70	2.27 / 0.35	11.7 / 0.65	1.59 / 0.22	5.29 / 0.35	76.3 /	10.7 /	2.79 / 0.41	19.5 / 0.67
SportsTraj †	3.32 / 0.30	17.2 / 0.55	2.61 / 0.32	12.1 / 0.64	1.94 / 0.17	6.14 / 0.30	88.9 /	11.7 /	2.59 / 0.42	12.9 / 0.71
UniGIO	1.88 / 0.45	7.85 / 0.75	1.87 / 0.38	8.23 / 0.71	1.35 / 0.14	4.35 / 0.33	70.6 /	9.27 /	1.71 / 0.61	8.28 / 0.78
Fixed-Length Forecasting										
WSSM	2.65 / 0.33	15.5 / 0.62	2.89 / 0.21	18.1 / 0.53	1.68 / 0.09	5.78 / 0.20	75.9 /	10.80 /	3.91 / 0.36	28.2 / 0.55
Chronos2	2.83 / 0.23	17.5 / 0.58	2.95 / 0.19	22.3 / 0.50	1.55 / 0.05	5.02 / 0.16	75.9 /	10.9 /	4.17 / 0.24	39.5 / 0.46
iTransformer	3.29 / 0.15	21.98 / 0.52	3.54 / 0.10	25.91 / 0.45	1.70 / 0.01	6.17 / 0.16	77.83 /	9.64 /	5.18 / 0.13	58.49 / 0.37
Corrformer	3.00 / 0.17	17.72 / 0.56	3.16 / 0.28	20.88 / 0.51	1.50 / 0.04	4.44 / 0.12	76.56 /	9.74 /	4.30 / 0.15	40.91 / 0.40
UniGIO	1.96 / 0.45	8.12 / 0.73	1.92 / 0.38	8.28 / 0.69	1.43 / 0.12	4.31 / 0.31	70.4 /	9.24 /	2.23 / 0.49	12.1 / 0.74
Forecasting with Incomplete Look-Back Window										
WSSM ⋆	3.23 / 0.13	20.1 / 0.49	3.26 / 0.11	21.84 / 0.47	1.63 / 0.01	5.53 / 0.10	74.7 /	9.15 /	5.23 / 0.13	56.1 / 0.38
Chronos2 ⋆	3.37 / 0.12	21.9 / 0.36	2.06 / 0.13	13.1 / 0.46	1.78 / 0.02	6.19 / 0.11	76.5 /	9.58 /	4.75 / 0.10	54.5 / 0.32
iTransformer ⋆	3.55 / 0.07	25.8 / 0.37	3.19 / 0.06	21.0 / 0.38	1.77 / 0.01	6.70 / 0.09	77.83 /	9.62 /	5.58 / 0.08	68.49 / 0.35
Corrformer ⋆	3.32 / 0.13	20.1 / 0.41	3.05 / 0.12	19.2 / 0.48	1.64 / 0.01	5.36 / 0.09	75.34 /	9.08 /	4.83 / 0.14	49.48 / 0.39
UniGIO	2.06 / 0.40	8.73 / 0.72	2.02 / 0.33	8.98 / 0.68	1.48 / 0.12	4.56 / 0.29	73.2 /	9.64 /	2.31 / 0.48	12.6 / 0.73

Table 6: 24h Forecast Comparison with Finetuned Observation and Time Series Foundation models

Baselines	Temperature		Dewpoint		Wind Rate		Wind Direc.		Sea-Level Pressure	
	MAE↓/99.5th↑	MSE↓/90.0th↑	MAE↓/99.5th↑	MSE↓/90.0th↑	MAE↓/99.5th↑	MSE↓/90.0th↑	MAE↓/99.5th↑	MSE↓ ($\times 10^3$)/90.0th↑	MAE↓/99.5th↑	MSE↓/90.0th↑
HRES	1.76 / -	7.39 / -	1.85 / -	7.94 / -	1.48 / -	4.53 / -	63.8 /	7.15 /	0.86 / -	2.68 / -
ESFM	1.76 / -	- / -	1.86 / -	- / -	1.40 / -	- / -	46.3 /	- /	2.55 / -	- / -
Chronos2	2.05 / 0.30	9.5 / 0.58	2.06 / 0.31	12.3 / 0.65	1.47 / 0.08	4.51 / 0.24	64.4 /	8.65 /	2.13 / 0.52	11.3 / 0.72
UniGIO fixed	1.28 / 0.63	3.68 / 0.83	1.32 / 0.52	4.29 / 0.78	1.15 / 0.20	2.88 / 0.45	58.6 /	7.63 /	1.05 / 0.76	2.82 / 0.87

Table 7: Quantitative Comparison on Imputation Task † means re-implemented

Baselines	Temperature		Dewpoint		Wind Rate		Wind Direc.		Sea-Level Pressure	
	MAE↓/99.5th↑	MSE↓/90.0th↑	MAE↓/99.5th↑	MSE↓/90.0th↑	MAE↓/99.5th↑	MSE↓/90.0th↑	MAE↓/99.5th↑	MSE↓ ($\times 10^3$)/90.0th↑	MAE↓/99.5th↑	MSE↓/90.0th↑
TimeGAN	8.84 / 0.14	64.84 / 0.42	6.10 / 0.23	36.4 / 0.57	3.74 / 0.03	12.48 / 0.20	152 /	18.5 /	8.94 / 0.20	80.6 / 0.49
TimeVAE	6.77 / 0.16	37.7 / 0.47	5.14 / 0.22	25.3 / 0.60	3.10 / 0.02	8.87 / 0.14	145 /	15.9 /	7.71 / 0.14	56.3 / 0.52
UniMTD †	1.78 / 0.51	6.69 / 0.77	1.61 / 0.49	6.20 / 0.75	1.21 / 0.19	3.13 / 0.42	59.5 /	7.01 /	1.89 / 0.62	9.23 / 0.78
SportsTraj †	1.74 / 0.60	3.80 / 0.84	1.65 / 0.58	3.88 / 0.82	1.52 / 0.30	3.17 / 0.49	73.9 /	7.87 /	1.25 / 0.82	1.93 / 0.91
SSSD	5.49 / 0.22	35.21 / 0.51	4.51 / 0.26	28.23 / 0.59	2.52 / 0.17	8.85 / 0.27	116.82 /	16.46 /	5.67 / 0.31	47.54 / 0.57
UniGIO	0.91 / 0.74	1.80 / 0.89	0.90 / 0.69	2.12 / 0.86	1.00 / 0.26	2.09 / 0.53	49.0 /	6.01 /	0.55 / 0.92	0.61 / 0.95

Table 8: Quantitative Comparison on Generation Task † means re-implemented

Baselines	Temperature		Dewpoint		Wind Rate		Wind Direc.		Sea-Level Pressure	
	MAE↓/99.5th↑	MSE↓/90.0th↑	MAE↓/99.5th↑	MSE↓/90.0th↑	MAE↓/99.5th↑	MSE↓/90.0th↑	MAE↓/99.5th↑	MSE↓ ($\times 10^3$)/90.0th↑	MAE↓/99.5th↑	MSE↓/90.0th↑
TimeGAN	3.81 / 0.12	23.9 / 0.47	2.80 / 0.20	15.2 / 0.59	1.87 / 0.03	6.15 / 0.20	77.7 /	9.46 /	4.19 / 0.19	34.5 / 0.51
TimeVAE	3.46 / 0.11	19.3 / 0.47	2.57 / 0.18	12.5 / 0.60	1.59 / 0.02	4.63 / 0.15	72.7 /	7.96 /	3.89 / 0.13	28.2 / 0.52
UniMTD †	2.60 / 0.33	14.15 / 0.67	2.83 / 0.23	17.2 / 0.57	1.57 / 0.10	5.31 / 0.27	75.2 /	9.58 /	4.39 / 0.24	43.2 / 0.50
SportsTraj †	2.10 / 0.41	9.06 / 0.73	2.04 / 0.34	9.35 / 0.69	1.63 / 0.20	5.55 / 0.35	80.8 /	11.6 /	1.46 / 0.71	5.19 / 0.84
Diffusion-TS	4.23 / 0.13	33.06 / 0.42	3.06 / 0.19	20.44 / 0.54	1.94 / 0.15	7.80 / 0.25	88.58 /	13.59 /	4.22 / 0.21	38.68 / 0.49
UniGIO	1.61 / 0.49	6.06 / 0.77	1.67 / 0.40	7.01 / 0.73	1.48 / 0.20	4.7 / 0.36	71.9 /	9.50 /	0.94 / 0.79	2.48 / 0.88

Table 9: Quantitative Evaluation on cross-region (South America / Africa) Generation

Settings	Temperature		Wind Rate		Wind Direc.		Temperature		Wind Rate	
	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓ ($\times 10^3$)	99.5th↑	90.0th↑	99.5th↑	90.0th↑
Worldwide	1.88	8.03	1.52	5.13	75.3	10.4	0.44	0.75	0.16	0.34
All stations	3.14 / 2.22	20.5 / 11.3	1.81 / 1.62	6.91 / 5.77	80.9 / 84.9	11.7 / 12.2	0.13 / 0.29	0.46 / 0.64	0.08 / 0.12	0.20 / 0.23
90% stations	2.28 / 2.07	15.9 / 9.67	1.83 / 1.52	7.48 / 4.77	74.0 / 82.1	10.2 / 11.4	0.22 / 0.27	0.60 / 0.68	0.08 / 0.08	0.25 / 0.24
60% stations	2.06 / 1.84	8.76 / 7.99	1.68 / 1.36	6.11 / 3.78	73.4 / 80.9	10.1 / 11.1	0.32 / 0.30	0.64 / 0.71	0.14 / 0.10	0.27 / 0.25
30% stations	2.04 / 1.70	8.20 / 7.17	1.65 / 1.35	5.76 / 3.05	72.7 / 79.8	10.4 / 10.7	0.36 / 0.33	0.70 / 0.72	0.13 / 0.16	0.28 / 0.27

Table 4 shows that UniGIO is the most lightweight models among all compared foundation and even time series models, with UniGIO-M being further compressed to $< 1M$. Its gains therefore suggest effective modeling rather than parameter scaling.

3.2.2 Qualitative Results

Fig. 4 compares forecasting, imputation, and generation results. UniGIO preserves both temporal trends and local variations, particularly for temperature and sea-level pressure. SportsTraj produces spurious fluctuations that may resemble wind-direction variability but disrupt smoother variables and imputation continuity. UniMTD over-smooths sequences, losing small-scale pressure patterns, while TimeGAN generates high-frequency fluctuations with incorrect trends. These differences highlight the need to accommodate evolution–extreme dynamics. The visualizations confirm the superiority of our method in quantitative results and also demonstrate heterogeneity among variables.

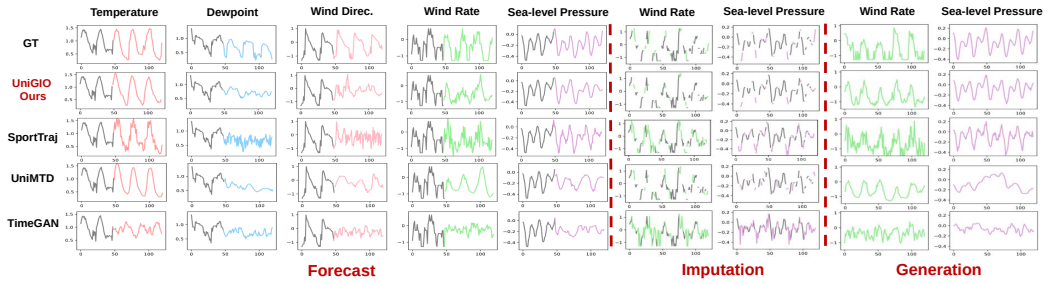

Figure 4: Visualized comparison. Due to space constraints, only wind rate and sea-level pressure are displayed for imputation and generation.

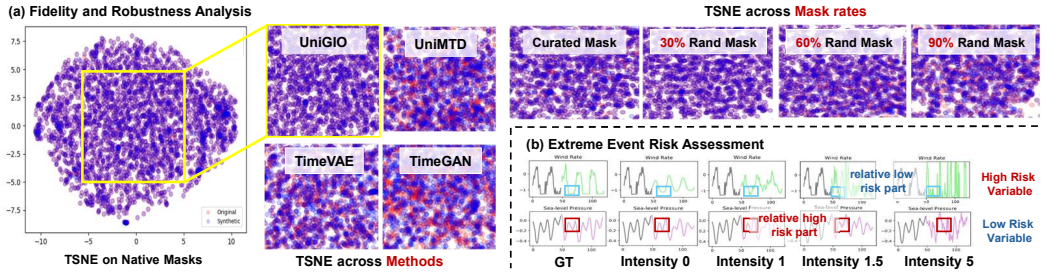

Figure 5: (a) t-SNE visualization of generated samples and ground truth across methods and mask settings; less exposed red indicates greater coverage.TSNE plots look unclustered because GIO is largely continuous with no 'category'. (b) Responses to sampling noise intensity for wind rate with extremes and stable sea-level pressure, revealing variable- and time-dependent event sensitivity.

3.3 Task-specific Analysis

To evaluate UniGIO as a versatile framework across tasks, in Fig. 3 (b)-(d), we verified that across seven forecasting, imputation, and generation tasks within 120h windows, UniGIO achieves the best results on 92% of 108 metrics. Accuracy/F1-SEDI gains reach 18.5%/8.13% for forecasting, 9.4%/6.0% for imputation, and 6.8%/4.4% for generation, outperforming specialized baselines.

We further examine forecasting, a central task in operational weather prediction. UniGIO surpasses WSSM, iTransformer, Corrformer, and Chronos2 in the 48h–72h fixed-window setting. With incomplete look-back observations, it improves accuracy/F1-SEDI by 24.5%/23.3% over imputation-then-forecasting approaches, supporting direct joint modeling of missing observations and future states. Under the ESFM 24h forecasting protocol, it also outperforms observation and time series foundation model ESFM, Chronos2, and operational HRES, with an improvement of 14.7% as shown in Fig. 3 (e). These results verify that UniGIO, as a lightweight unified framework, is broadly effective across different sub-tasks. It can even surpass task-specific models and, despite model and data volume, outperform foundation models and operational NWP that rely on global multi-source observations or generic time series. (Tables 5, 6, 7, and 8).

3.4 Spatial Generalization Analysis

To evaluate spatial generalization, we conduct 120h cross-region generation in South America and Africa, focusing on temperature, wind rate, and wind direction (Table 9; full results in Appendix G.4). Unlike random station splits, cross-region evaluation excludes target regions during training and zero-shot and fine-tuning with 10%, 40%, and 70% regional stations. With only 10% regional data, fine-tuning improves zero-shot performance; with 70% data, it surpasses the worldwide baseline on 35% metrics, while extreme-event capture remains about 10% lower, indicating that UniGIO's ability to adapt across unseen region and spatial adaptation is easier than rare-event transfer.

3.5 Fidelity and Robustness Analysis

Fig. 5 (a) visualizes generated and ground-truth samples through t-SNE, where exposed red regions indicate uncovered ground-truth samples. UniGIO shows greater overlap than competing baselines

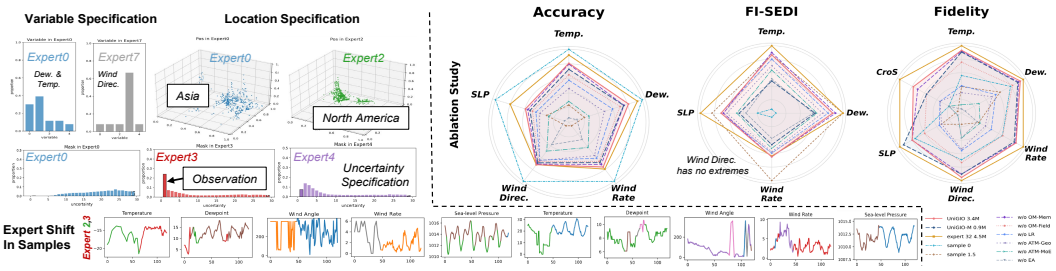

Figure 6: Ablation study results (right) and the representative expert specializations across variables, locations, uncertainties, and temporal patterns (left), different experts are denoted by colors.

under both curated and native masks, consistent with its quantitative fidelity results. Coverage declines as missing rates increase from 30% to 90%, but substantial overlap remains across mask types and rates. These observations support robustness to varying missing patterns.

3.6 Extreme Event Risk Analysis

Fig. 5 (b) examines sampling sensitivity for wind speed with multiple extremes and relatively stable sea-level pressure. Zero noise produces smooth mean trends; increasing noise first amplifies wind-speed fluctuations while leaving pressure relatively stable, whereas excessive noise destabilizes both. Responses also vary within each sequence, indicating temporally localized sensitivity rather than uniform uncertainty at the variable level. These observations suggest that latent variance may indicate extreme-event risk and support pointwise control of generated extremes (Fig. C.2(b)).

3.7 Model Structure Analysis

Fig. 6 (right part) and Tables H.1 and H.2 summarize ablations on 120h windows. Increasing experts from 16 to 32 improves performance, while UniGIO-M reduces parameters by 73.5% with only a 5% average performance loss. Lower sampling intensity favors accuracy over F1-SEDI, whereas higher intensity reverses this trade-off; both reduce fidelity relative to the default setting.

Component ablations support the complementary roles of the modules. OM improves accuracy and fidelity, consistent with modeling weather as spatial fields rather than isolated station sequences. MoE improves overall performance, particularly extreme-event capture, while LR strengthens local reconstruction and fidelity. EA benefits variables with distinct periodicity, supporting the value of aligning inter-station events.

Expert routing further reveals specialization across variables, regions, and uncertainty levels (Fig. 6 (left part)). Expert 0 favors temperature and dew point, Asian stations, and distant high-uncertainty tokens; Expert 7 favors wind direction, and Expert 2 concentrates on North America and Europe. Experts 4 and 3 favor nearby low-uncertainty tokens and observed values, respectively. Despite token-level routing, expert assignments remain coherent across temporal segments, suggesting reliability-aware temporal representation preserves sequential structure.

4 Conclusion

In this work, we address global in-situ weather modeling under incomplete observations and establish a benchmark for accuracy, fidelity, and extreme-event capture. We further propose UniGIO, a unified framework for forecasting, imputation, and generation directly on interpolation-free distributed stations, emphasizing station-centered complementarity, evolution–extreme decoupling, and lightweight design. OM and EA progressively capture station-to-region, discrete-to-continuous complementarity; ATM decouples stable evolution from chaotic extremes with shared and separate representations, and LR restores local continuity, while explicit uncertainty modeling enables extreme-event assessment and generation. Across curated and native masks over seven tasks, UniGIO achieves SOTA performance, with average gains of 11%, 12%, and 5% in accuracy, fidelity, and extreme-event capture, respectively, demonstrating its potential for future distributed climate foundation models.

5 AI USE STATEMENT

AI tools were used solely to improve formatting and polish the language of the manuscript. All research ideas, methodological designs, experiments, and analyses were developed and conducted by the authors, who reviewed all AI-assisted edits and take full responsibility for the final content.

6 REPRODUCIBILITY STATEMENT

We provide detailed mathematical formulations of UniGIO and its components in the main paper and Appendix D. Appendix B specifies the task-specific mask construction for forecasting, imputation, and generation, while Appendix C provides the complete training and inference pipeline. Dataset construction and temporal splits, curated and native missingness settings, baseline adaptations, evaluation protocols, and metric definitions are documented in Appendix E. Additional evaluations across temporal windows, tasks, regions, and missingness settings are provided in the supplementary appendices. The source code and the mask configurations used in our experiments will be made publicly available soon.

REFERENCES

Juan Lopez Alcaraz and Nils Strodthoff. Diffusion-based time series imputation and forecasting with structured state space models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=hHiIbk7ApW.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksii Shchur, Syama Sundar Rangapuram, Sebastian Arango, Shubham Kapoor, Danielle C Maddix, et al. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

Peter Bauer, Alan Thorpe, and Gilbert Brunet. The quiet revolution of numerical weather prediction. *Nature*, 525:47–55, 2015. doi: 10.1038/nature14956.

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Accurate medium-range global weather forecasting with 3d neural networks. *Nature*, 619:533–538, 2023. doi: 10.1038/s41586-023-06185-3.

Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G. Willcocks. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7327–7347, 2022.

Lane F Burgette and Jerome P Reiter. Multiple imputation for missing data via sequential regression trees. *American journal of epidemiology*, 172(9):1070–1076, 2010.

Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: Bidirectional recurrent imputation for time series. In *Advances in Neural Information Processing Systems*, volume 31, pp. 6776–6786, 2018.

Lei Chen, Xingchao Zhong, Feng Zhang, Yuzhong Cheng, Yunjie Xu, Yu Qi, Hao Li, Min Tang, Ru Gao, Min Wang, et al. Fuxi: A cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, 6(1):190, 2023.

Abhyuday Desai, Cynthia Freeman, Zuhui Wang, and Ian Beaver. Timevae: A variational auto-encoder for multivariate time series generation. *arXiv preprint arXiv:2111.08095*, 2021.

European Centre for Medium-Range Weather Forecasts. IFS Documentation CY49R1 – Part V: Ensemble Prediction System. https://www.ecmwf.int/en/elibrary/81373-ifs-documentation-cy49r1-part-v-ensemble-prediction-system, 2024.

Zhangyang Gao, Xun Shi, Hao Wang, Yuchen Zhu, Yuyang Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. In *Advances in Neural Information Processing Systems*, 2022.

Junchao Gong, Kaiyi Xu, Wangxu Wei, Siwei Tu, Jingyi Xu, Zili Liu, Hang Fan, Zhiwang Zhou, Tao Han, Yi Xiao, et al. Earth-o1: A grid-free observation-native atmospheric world model. *arXiv preprint arXiv:2605.06337*, 2026.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *International Conference on Learning Representations*, 2024.

Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4189–4210, 2023.

Tao Han, Zhibin Wen, Zhenghao Chen, Dazhao Du, Song Guo, and Lei Bai. Benchmarking physics-informed time-series models for operational global station weather forecasting. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL `https://arxiv.org/abs/2406.14399`.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6629–6640, 2017.

Hansika Hewamalage, Christoph Bergmeir, and Kasun Bandara. Recurrent neural networks for time series forecasting: Current status and future directions. *International Journal of Forecasting*, 37 (1):388–427, 2021.

Yiheng Huang, Xiaowei Mao, Shengnan Guo, Yubin Chen, Junfeng Shen, Tiankuo Li, Youfang Lin, and Huaiyu Wan. Std-plm: Understanding both spatial and temporal properties of spatial-temporal data with plm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 11817–11825, 2025.

Eugenia Kalnay. *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 2003.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *International Conference on Learning Representations*, 2021.

Remi Lam, Alvaro Sanchez-Gonzalez, Matthew Willson, Peter Wirnsberger, Meire Fortunato, Ferran Alet, Suman Ravuri, Timo Ewalds, Zach Eaton-Rosen, Weihua Hu, et al. Learning skillful medium-range global weather forecasting. *Science*, 382(6677):1416–1421, 2023.

Bryan Lim and Stefan Zohren. Time-series forecasting with deep learning: a survey. *Philosophical transactions of the royal society a: mathematical, physical and engineering sciences*, 379(2194), 2021.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. itransformer: Inverted transformers are effective for time series forecasting. In *The Twelfth International Conference on Learning Representations*, 2024.

Yukai Liu, Rose Yu, Stephan Zheng, Eric Zhan, and Yisong Yue. Naomi: Non-autoregressive multiresolution sequence imputation. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021. doi: 10.1109/ICCV48922.2021.00986.

Xiaowen Ma, Zhenliang Ni, Shuai Xiao, and Xinghao Chen. Timepro: Efficient multivariate long-term time series forecasting with variable-and time-aware hyper-state. *arXiv preprint arXiv:2505.20774*, 2025.

Tung Nguyen, Ruchi Shah, Hritik Bansal, Tommy Arcomano, Srilakshmi Madireddy, Romit Maulik, Karthik Kashinath, et al. Climax: A foundation model for weather and climate. In *International Conference on Machine Learning*, 2023.

Norazian Mohamed Noor, Mohd Mustafa Al Bakri Abdullah, Ahmad Shukri Yahaya, and Nor Azam Ramli. Comparison of linear interpolation method and mean method to replace the missing values in environmental data set. In *Materials science forum*, volume 803, pp. 278–281. Trans Tech Publ, 2015.

Firat Ozdemir, Yun Cheng, Salman Mohebi, Fanny Lehmann, Simon Adamov, Zhenyi Zhang, Leonardo Trentini, Dana Grund, Oliver Fuhrer, Torsten Hoefler, et al. Earth system foundation model (esfm): A unified framework for heterogeneous data integration and forecasting. *arXiv preprint arXiv:2605.00850*, 2026.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010. doi: 10.1109/TKDE.2009.191.

Ilan Price, Alvaro Sanchez-Gonzalez, Ferran Alet, Tom R Andersson, Andrew El-Kadi, Dominic Masters, Timo Ewalds, Jacklynn Stott, Shakir Mohamed, Peter Battaglia, et al. Probabilistic weather forecasting with machine learning. *Nature*, 637(8044):84–90, 2025.

Md Geaur Rahman and Md Zahidul Islam. Missing value imputation using a fuzzy clustering-based em approach. *Knowledge and Information Systems*, 46(2):389–422, 2016.

Stephan Rasp, Boris Babenko, Dominic Masters, Andrew El-Kadi, Samier Merchant, Guy Shalev, Ilan Price, Fred Zyda, Remi Lam, Sasha Shysheya, et al. Weathernext 3: Increasing resolution and performance of global weather models with raw observations. *arXiv preprint arXiv:2609.03582*, 2026.

David P. Rogers, Vladimir V. Tsirkunov, Haleh Kootval, Alice Soares, Alisher Khamidov, Michael Staudinger, and Mohamed Kadi. *Weathering the Change: How to Improve Hydromet Services in Developing Countries*. World Bank, Washington, DC, 2019. doi: 10.1596/31507. URL https://www.gfdrr.org/en/publication/weathering-change-how-improve-hydromet-services-developing-countries.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Sparsely-gated mixture-of-experts layers. *arXiv preprint arXiv:1701.06538*, 2017.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24804–24816, 2021.

UNDRR. Global assessment report on disaster risk reduction 2022: Our world at risk: Transforming governance for a resilient future, 2022. URL https://www.undrr.org/gar/gar2022-our-world-risk. Referenced for extensive risk and high-frequency, low-severity event statistics.

Anna Vaughan, Stratis Markou, Will Tebbutt, James Requeima, Wessel P Bruinsma, Tom R Andersson, Michael Herzog, Nicholas D Lane, Matthew Chantry, J Scott Hosking, et al. End-to-end data-driven weather prediction. *Nature*, 641:1172–1179, 2025. doi: 10.1038/s41586-025-08897-0.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, 2018. doi: 10.1109/CVPR.2018.00813.

Yihang Wang, Yuying Qiu, Peng Chen, Yang Shu, Zhongwen Rao, Lujia Pan, Bin Yang, and Chenjuan Guo. Lightgts: A lightweight general time series forecasting model. *arXiv preprint arXiv:2506.06005*, 2025.

WMO. Systematic observations financing facility (SOFF). URL https://wmo.int/activities/systematic-observations-financing-facility-soff.

WMO. *Manual on the Global Observing System*. World Meteorological Organization, Geneva, Switzerland, 2018. URL https://library.wmo.int/doc_num.php?explnum_id=10000000000067.

WMO. WMO Strategic Plan 2020–2023. Technical Report WMO-No. 1225, World Meteorological Organization, Geneva, Switzerland, 2019. URL https://etrp.wmo.int/pluginfile.php/25842/mod_resource/content/1/WMO%20Strategic%20Plan%202020-2023%20%28WMO-No.1225%29.pdf.

WMO. *Technical Regulations for the Global Basic Observing Network*. World Meteorological Organization, Geneva, Switzerland, 2021a.

WMO. Wmo atlas of mortality and economic losses from weather, climate and water extremes (1970–2019). Technical Report WMO-No. 1267, World Meteorological Organization, Geneva, Switzerland, 2021b.

WMO. *Guide to Instruments and Methods of Observation: Volume I – Measurement of Meteorological Variables*. World Meteorological Organization, Geneva, Switzerland, 2024.

Haixu Wu, Hang Zhou, Mingsheng Long, and Jianmin Wang. Interpretable weather forecasting for worldwide stations with a unified deep model. *Nature Machine Intelligence*, 5(6):602–611, 2023.

Yi Xu and Yun Fu. Sports-traj: A unified trajectory generation model for multi-agent movement in sports. In *The Thirteenth International Conference on Learning Representations*, 2025.

Yi Xu, Armin Bazarjani, Hyung-gun Chi, Chiho Choi, and Yun Fu. Uncovering the missing pattern: Unified framework towards trajectory imputation and prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9632–9643, 2023.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (12):15245–15268, 2023.

Songru Yang, Zili Liu, Zhenwei Shi, and Zhengxia Zou. Wssm: Geographic-enhanced hierarchical state-space model for global station weather forecast. *arXiv preprint arXiv:2501.11238*, 2025a.

Songru Yang, Zhenwei Shi, and Zhengxia Zou. Unified multi-agent trajectory modeling with masked trajectory diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27563–27574, 2025b.

Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar. Time-series generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Chengqing Yu, Fei Wang, Chuanguang Yang, Zezhi Shao, Tao Sun, Tangwen Qian, Wei Wei, Zhulin An, and Yongjun Xu. Merlin: Multi-view representation learning for robust multivariate time series forecasting with unfixed missing rates. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 3633–3644, 2025.

Xinyu Yuan and Yan Qiao. Diffusion-ts: Interpretable diffusion for general time series generation. In *The Twelfth International Conference on Learning Representations*, 2024.

Xiao Xiang Zhu, Zhitong Xiong, Yi Wang, Adam J Stewart, Konrad Heidler, Yuanyuan Wang, Zhenghang Yuan, Thomas Dujardin, Qingsong Xu, and Yilei Shi. On the foundations of earth foundation models. *Communications Earth & Environment*, 7(1):103, 2026. doi: 10.1038/s43247-025-03127-x.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=gZ9hCDWe6ke.

A Related Work

A.1 Data-driven Weather Prediction

Data-driven weather prediction has attracted growing interest, with models such as Pangu-Weather (Bi et al., 2023), GraphCast (Lam et al., 2023), GenCast (Price et al., 2025), and FuXi (Chen et al., 2023) operating on gridded reanalysis data (e.g., 0.25° and 0.09° resolution). In the case that these large-scale, region-averaged methods may not match the station level weather systems and can ignore highly destructive local extreme weather, some initial attempts, such as Corrformer (Wu et al., 2023) and WSSM (Yang et al., 2025a), have treated station weather forecasting as an independent task and achieved promising results. However, these methods operate on spatiotemporally complete data, leaving a gap for incomplete GIO, which greatly limits their real-world applications.

Recently, spatiotemporal weather foundation models (Gao et al., 2022; Nguyen et al., 2023) and observation foundation models (Gong et al., 2026; Vaughan et al., 2025; Ozdemir et al., 2026) have shown the potential to unify station and grid weather modeling at arbitrary locations. Nevertheless, this paradigm remains poorly suited to GIO. It relies on massive multi-source data and intensive computation for global-scale autoregressive forecasting, which differs from the lightweight, station-focused setting considered here. Also, although they implicitly incorporate data assimilation into grid-like structured latent spaces to support observation inputs, their GIO forecasts are still derived from interpolation over smoothed latent fields. As a result, they remain fundamentally close to grid-based forecasting and inherit its key limitations. These challenges indicate that a unified weather modeling framework tailored to GIO is not only indispensable for practical in-situ services, but also essential for bridging the gap in the grid-station and global-local weather modeling.

A.2 Multivariate Time Series Forecasting and Imputation

Forecasting and imputation are two fundamental tasks in multivariate time series modeling. Forecasting aims to infer future dynamics from historical observations, while imputation focuses on recovering missing values from partially observed sequences. In early studies, CNN-based methods (Lim & Zohren, 2021) were adopted to extract local temporal patterns, and RNN-based methods (Hewamalage et al., 2021) were widely used to model sequential dependencies. Recently, Transformer-based architectures (Wang et al., 2025) have enabled more flexible long-range correlation modeling; this trend is consistent with non-local attention (Wang et al., 2018), hierarchical vision transformers (Liu et al., 2021), and deformable attention (Zhu et al., 2021). state-space models (Ma et al., 2025) further introduce temporal inductive biases and recurrent-like computation to construct efficient and robust long-term dependencies. Imputation is often regarded as a preprocessing step for slightly incomplete data. Classical non-learnable methods, such as interpolation (Noor et al., 2015), regression (Burgette & Reiter, 2010), and EM algorithms (Rahman & Islam, 2016), are simple and efficient, but they usually rely on smoothness or distributional assumptions and struggle to recover complex non-stationary sequences. To improve imputation quality, recent learning-based frameworks (Liu et al., 2019; Xu et al., 2023) introduce neural structures such as recurrent networks, attention mechanisms, and graph neural networks to exploit temporal continuity and cross-variable dependencies.

More recent efforts attempt to perform forecasting under missing observations (Yu et al., 2025; Huang et al., 2025), indicating the growing need for models that are robust to incomplete inputs. However, most existing methods still treat forecasting and imputation as separate tasks, where imputation is either performed as an independent preprocessing stage or forecasting is designed for a specific missing pattern. This separation limits their applicability to Global In-situ Observation (GIO), where missingness is not slight or fixed but arbitrarily coexists across variables, stations, regions, and forecast horizons. From a broader learning perspective, current data synthesis and reconstruction methods have shown that imperfect observations and distributions can substantially impair generalization and reliability (Pan & Yang, 2010). To solve these problems, our work formulates these incomplete patterns with a unified observation mask and models forecasting, imputation, and generation within a single framework.

A.3 Generative Time Series Modeling

Generative time series modeling aims to synthesize realistic time series by learning the underlying data distribution, often generating sequences from random noise. Early studies mainly explored GAN-

based (Yoon et al., 2019) and VAE-based (Desai et al., 2021) frameworks, which respectively model temporal distributions through adversarial learning and latent-variable reconstruction. More recently, diffusion models (Yang et al., 2023) have become a dominant paradigm due to their stable training and strong distribution modeling capability in related applications. These generative paradigms (Gui et al., 2023; Bond-Taylor et al., 2022) focusing on data distribution learning, laid the foundation for task unification. By controlling generation conditions, generative methods can naturally support forecasting, imputation, and generation simultaneously under arbitrary masks, providing a flexible formulation for incomplete time series modeling.

Despite this natural unification capability, most existing generative time series methods focus on fitting the global data manifold, which becomes problematic for Global In-situ Observation (GIO), where the global weather manifold is often dominated by smooth evolution, whereas local dynamics are frequently affected by strong extreme noise. Simply landing on this manifold may lead to over-smoothed predictions that miss local extremes; worse, under sparse and noisy conditions, models often fit spurious noise instead of real dynamics, resulting in uncontrolled fluctuations.

We empirically apply advanced generative time series methods, such as diffusion models, to GIO and find that their conditional information is often weakly exploited or even largely discarded, resulting in degraded performance. This indicates that available observations serve as critical anchors for preserving realistic local states and constraining physically plausible generation. Motivated by this, UniGIO adopts an observation-conditioned CVAE whose prior mean represents observation-supported states and whose variance controls uncertain latent perturbations, thereby balancing smooth temporal evolution and localized abrupt mutations in a decoupled manner.

B TASK-SPECIFIC MASK FORMULATIONS

This section specifies how task-specific masks are constructed. We use an observation mask $\mathbf{M}$, where $M_{n,t,c} = 1$ indicates that the corresponding value is available to the model and $M_{n,t,c} = 0$ denotes a masked position requiring reconstruction. In real deployment scenarios, missing observations generally lack corresponding ground truth. To simulate this setting while enabling controlled evaluation, we train models with curated masks constructed from available observations and evaluate them on both curated and native Weather-5K missing patterns. The curated masks provide supervised targets during training, while native missing patterns are used as a zero-shot evaluation of robustness to realistic observation failures.

B.1 FIXED- AND VARIABLE-WINDOW FORECASTING

For look-back length i within a window of length T,

$$\begin{aligned} M_{n,t,c} &= \mathbf{1}\{t \le i\} \\ n &= 1, \dots, N,\ c = 1, \dots, C. \end{aligned} \tag{B.1}$$

Fixed-window forecasting holds i fixed; variable-window forecasting samples it. The forecast cutoff is shared across conditioning stations, so observations after the cutoff are unavailable even at neighboring stations. Native gaps may remain within the look-back window. This distinguishes direct forecasting with incomplete history from first imputing that history and then applying a forecasting model.

B.2 RANDOM AND PERIOD IMPUTATION

Let $\mathcal{P}_n$ contain randomly selected time-variable entries at station n, and let $\mathcal{B}_n = \bigcup_{l=1}^{L_n} [a_{n,l}, b_{n,l}]$ contain missing temporal blocks. Then

$$M_{n,t,c} = \mathbf{1}\{(t, c) \notin \mathcal{P}_n \text{ and } t \notin \mathcal{B}_n\}. \tag{B.2}$$

Point entries are sampled with probability p_{point}; block locations and lengths are sampled to represent sustained outages. Selected variables can be unmasked within sampled blocks to reflect independent sensor availability.

B.3 RESOLUTION IMPUTATION

For a sampled temporal sampling interval S,

$$M_{n,t,c} = \mathbf{1}\{t = 0 \pmod S\}. \tag{B.3}$$

The retained observations simulate a lower-frequency station record, and the intervening hours are reconstructed. Unlike random masking, the missing positions follow a regular temporal pattern.

B.4 VARIABLE GENERATION

For a selected variable c^*,

$$M_{n,t,c} = \mathbf{1}\{c \neq c^*\}. \tag{B.4}$$

The selected variable is hidden throughout the relevant station sequence, requiring the model to use other variables and available station context. This simulates a sensor type that is absent rather than temporarily malfunctioning.

B.5 IN-REGION AND CROSS-REGION STATION GENERATION

For a set of target stations $\mathcal{S}_{\text{gen}}$,

$$M_{n,t,c} = \mathbf{1}\{n \notin \mathcal{S}_{\text{gen}}\}. \tag{B.5}$$

All variables and timestamps at target stations are unavailable to the model, while the remaining stations provide spatial context. A singleton target set corresponds to generating one missing station, while larger sets enable multi-station generation. In-region generation targets stations within the geographic training coverage, whereas cross-region generation holds out an unseen region. These tasks share the same mask formulation but differ in the spatial training and evaluation split. The South America and Africa experiments further vary the fraction of regional stations used for fine-tuning, with the remaining stations reserved for generation (Section E.7).

C UniGIO Pipeline

The prior supplies the conditioning features in both phases. The posterior supplies reconstruction samples only during training; inference instead samples from the conditional prior. Detailed module definitions appear in Appendix D.

Algorithm C.1 UniGIO CVAE training and inference

Require: Incomplete observations $\mathbf{X}_M$, mask $\mathbf{M}$, metadata $\mathbf{I}$; complete target $\mathbf{X}$ during training.
Ensure: Completed sequence $\hat{\mathbf{X}}$.

1. **Conditional prior branch:**
2. $\mathbf{f}_{\mathrm{C}} \leftarrow \mathrm{OM}(\mathbf{X}_M, \mathbf{M}, \mathbf{I})$.
3. $\mathbf{f}_{\mathrm{T}} \leftarrow \mathrm{ATM}(\mathbf{f}_{\mathrm{C}}; \mathbf{M}, \mathbf{I})$.
4. $(\boldsymbol{\mu}_p, \boldsymbol{\sigma}_p) \leftarrow \mathrm{PriorMLP}(\mathbf{f}_{\mathrm{T}})$.
5. $p_\psi(\mathbf{z} \mid \mathbf{X}_M, \mathbf{M}, \mathbf{I}) = \mathcal{N}(\boldsymbol{\mu}_p, \mathrm{diag}(\boldsymbol{\sigma}_p^2))$ (Eq. D.3).
6. **Training posterior branch:**
7. $\mathbf{f}_{\mathrm{S}} \leftarrow \mathrm{EA}(\mathbf{X}, \mathbf{I})$ (Eq. D.10).
8. $\mathbf{f}_{\mathrm{V}} \leftarrow \mathrm{MHSA}_{\mathrm{var}}(\mathbf{f}_{\mathrm{S}})$.
9. $\mathbf{f}_{\mathrm{Q}} \leftarrow \mathrm{BiSSM}(\mathbf{f}_{\mathrm{V}})$.
10. $(\boldsymbol{\mu}_q, \boldsymbol{\sigma}_q) \leftarrow \mathrm{PostMLP}(\mathbf{f}_{\mathrm{Q}})$.
11. $q_\phi(\mathbf{z} \mid \mathbf{X}, \mathbf{I}) = \mathcal{N}(\boldsymbol{\mu}_q, \mathrm{diag}(\boldsymbol{\sigma}_q^2))$ (Eq. D.4).
12. **CVAE training:**
13. $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d), \quad \mathbf{z}_q \leftarrow \boldsymbol{\mu}_q + \boldsymbol{\sigma}_q \odot \boldsymbol{\epsilon}$.
14. $\hat{\mathbf{X}} \leftarrow p_\theta\big(\mathrm{LR}(\mathbf{z}_q), \mathbf{f}_{\mathrm{T}}\big)$ (Eq. D.16).
15. Optimize reconstruction, prior–posterior KL, and orthonormal losses (Eq. D.17).
16. **CVAE inference / generation:**
17. Choose pointwise sampling strength $\boldsymbol{\alpha} \in \mathbb{R}^{N \times T \times C}$ and draw $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$.
18. $\mathbf{z}_p \leftarrow \boldsymbol{\mu}_p + \boldsymbol{\alpha} \odot \boldsymbol{\sigma}_p \odot \boldsymbol{\epsilon}$ (Eq. D.15).
19. $\hat{\mathbf{X}} \leftarrow p_\theta\big(\mathrm{LR}(\mathbf{z}_p), \mathbf{f}_{\mathrm{T}}\big)$ (Eq. D.16).

C.1 Observation Mixer Details

The Observation Mixer complements discrete station observations with continuous field representations and aligns asynchronous inter-station events. The Station-Centric Field Complement diffuses station-level features into local fields through field and memory tokens, while Event Aligner corrects temporal offsets between stations using frequency-domain correlation. Sections D.3–D.4 give their definitions.

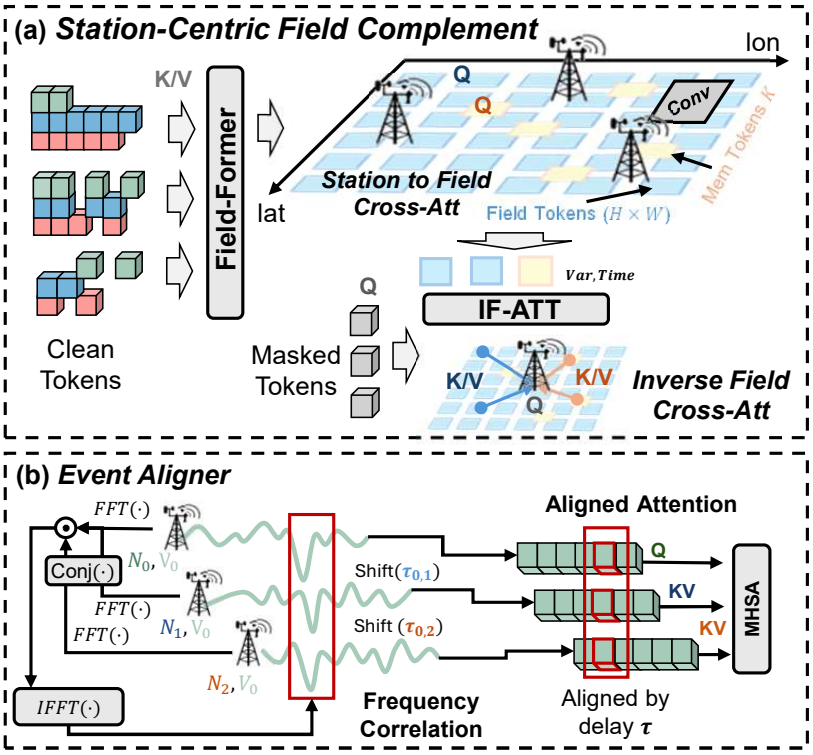

Figure C.1: Detailed designs in Observation Mixer: (a) Station-Centric Field Complement to diffuse discrete station level observations into continuous local space with field and memory tokens; (b) Event Aligner corrects inter-station temporal delays via frequency-correlation alignment.

C.2 Temporal Adaptation and Controllable Sampling

The upper diagram separates distance-based reliability weighting from pattern/geographic expert routing. Input weighting acts before recurrent state updates, whereas expert routing adapts the resulting temporal features. The lower diagram illustrates pointwise perturbation of the conditional prior. Reliability weights and prior variance are different quantities: the former controls feature entry into ATM, while the latter controls the spread of latent samples. Sections D.5–D.7 give their definitions.

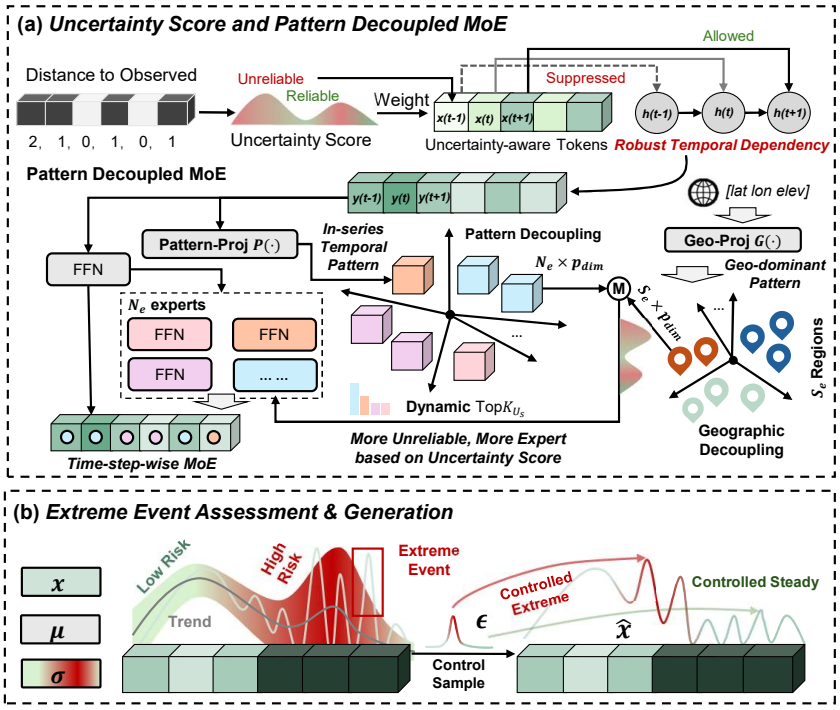

Figure C.2: Details for extreme pattern capture, assessment, and generation. (a) Uncertainty Score and Pattern Decoupled MoE suppress missing-induced state noise and decouple global similarity from local heterogeneity. (b) Extreme event Assessment and Generation use CVAE variance and sampling-noise control to estimate and generate extremes.

D Detailed Method Formulation

This section expands the compact formulation in the main text. Sections D.1–D.2 define the notation and conditional branches; Sections D.3–D.7 describe the modules; Section D.8 gives the learning objective. Algorithm C.1 separates training from inference.

D.1 Task Formulation and Notation

Let $\mathbf{X} \in \mathbb{R}^{N\times T\times C}$ denote a local collection of weather sequences, where N counts stations in the current neighborhood, T is the window length, and C counts variables. Thus N need not equal the total number of stations in the benchmark. The observation mask has the same shape:

$$M_{n,t,c} = \mathbf{1}\{X_{n,t,c} \text{ is available to the model}\}, \quad n = 1,\dots,N,\ t = 1,\dots,T,\ c = 1,\dots,C. \tag{D.1}$$

Here $\mathbf{1}\{\cdot\}$ is an indicator. The incomplete input $\mathbf{X}_M$ contains the values available to the model, while $\mathbf{M}$ denotes their availability, with $M_{n,t,c} = 1$ for an observed position and $M_{n,t,c} = 0$ for a missing or intentionally hidden position. Accordingly, positions requiring reconstruction are indicated by $1 - \mathbf{M}$. A numerical placeholder at a missing position is not treated as an observed weather measurement. Metadata $\mathbf{I}$ provides station and temporal information.

The conditional modeling problem is

$$\begin{aligned} \hat{\mathbf{X}} &= f_\theta(\mathbf{X}_M, \mathbf{M}, \mathbf{I}) \\ \min_\theta \mathcal{L}\big(\mathbf{X}, f_\theta(\mathbf{X}_M, \mathbf{M}, \mathbf{I})\big). \end{aligned} \tag{D.2}$$

The target notation $\mathbf{X}$ refers to values available for supervision; it does not imply that genuinely missing observations have known ground truth. Appendix E distinguishes input availability from evaluation eligibility. Forecasting, imputation, and generation differ in their masks, as specified in Appendix B.

Table D.1: Major notation used in the detailed method. Station and variable indices are suppressed in token-wise temporal equations.

Symbol	Meaning
N, T, C, d	Local station count, sequence length, variable count, and feature dimension.
$\mathbf{X}_M, \mathbf{M}, \mathbf{I}$	Incomplete observations, availability mask, and metadata.
$\mathbf{f}_{\mathrm{C}}$	Observation-complemented features before temporal modeling.
$\mathbf{h}_t, \mathbf{f}_{\mathrm{T}}^t$	Temporal features before and after expert aggregation.
$\mathbf{a}_t^{\rightarrow}, \mathbf{a}_t^{\leftarrow}$	Forward and backward internal state-space representations.
$\mathbf{z}_{\mathrm{field}}, \mathbf{z}_{\mathrm{mem}}$	Regional field queries and shared learnable memory tokens.
H, W, K	Field-grid height, width, and memory-token count.
D_{ij}, d_t	Inter-station event delay and distance to an observed time index.
N_e, S_e, d_p	Expert count, geographic-subspace count, and routing-subspace dimension.
$E_t, \mathcal{E}_t, w_{t,e}$	Active expert count, selected index set, and aggregation weight.
$\boldsymbol{\mu}_p, \boldsymbol{\sigma}_p$	Conditional-prior mean and standard deviation.
$\boldsymbol{\mu}_q, \boldsymbol{\sigma}_q$	Training-posterior mean and standard deviation.

Throughout, $[\cdot;\cdot]$ denotes concatenation along the indicated token or feature dimension, $\odot$ is elementwise multiplication, and MLP denotes a learned multilayer perceptron. Operators with different subscripts have different roles and need not share parameters. The identity matrix with k dimensions is written $\mathbf{I}_k$ to distinguish it from metadata $\mathbf{I}$.

D.2 Framework and Conditional Branches

UniGIO first collects complementary observation evidence and then models temporal evolution (Fig. 2). OM exchanges information across variables and stations and summarizes their shared regional field. ATM applies reliability weighting before bidirectional state-space modeling, followed by PD-MoE adaptation. LR combines sampled latent information with the adapted temporal features for local reconstruction.

Conditional prior. Write $\mathcal{O} = (\mathbf{X}_M, \mathbf{M}, \mathbf{I})$ for the available conditioning information. The inference branch is

$$\begin{aligned} \mathbf{f}_{\mathrm{C}} &= \mathrm{OM}(\mathcal{O}) \\ \mathbf{f}_{\mathrm{T}} &= \mathrm{ATM}(\mathbf{f}_{\mathrm{C}}; \mathbf{M}, \mathbf{I}), \\ (\boldsymbol{\mu}_p, \boldsymbol{\sigma}_p) &= \mathrm{PriorMLP}(\mathbf{f}_{\mathrm{T}}) \\ p_\psi(\mathbf{z} \mid \mathcal{O}) &= \mathcal{N}(\boldsymbol{\mu}_p, \mathrm{diag}(\boldsymbol{\sigma}_p^2)). \end{aligned} \tag{D.3}$$

The semicolon makes explicit the mask and metadata used by ATM's reliability and routing mechanisms. The Gaussian notation flattens latent tensor coordinates only for expressing the diagonal covariance.

Training posterior. The posterior additionally accesses target sequences for event alignment. EA aligns neighboring evidence, variable-wise attention combines aligned variable features, and a bidirectional state-space encoder produces posterior features $\mathbf{f}_{\mathrm{Q}}$:

$$\begin{aligned} \mathbf{f}_{\mathrm{S}} &= \mathrm{EA}(\mathbf{X}, \mathbf{I}), \\ \mathbf{f}_{\mathrm{Q}} &= \mathrm{BiSSM}\left(\mathrm{MHSA}_{\mathrm{var}}(\mathbf{f}_{\mathrm{S}})\right), \\ (\boldsymbol{\mu}_q, \boldsymbol{\sigma}_q) &= \mathrm{PostMLP}(\mathbf{f}_{\mathrm{Q}}) \\ q_\phi(\mathbf{z} \mid \mathbf{X}, \mathbf{I}) &= \mathcal{N}(\boldsymbol{\mu}_q, \mathrm{diag}(\boldsymbol{\sigma}_q^2)). \end{aligned} \tag{D.4}$$

Here $\mathbf{f}_{\mathrm{S}}$ is EA's aligned output from Section D.4. As in the main text, $q_\phi(\mathbf{z} \mid \mathbf{X}, \mathbf{I})$ suppresses ancillary conditioning arguments. The posterior is used to learn reconstruction and to supervise the conditional prior through KL divergence; it is not evaluated at inference.

Why separate the stages? Sparse observations create two coupled difficulties: missing states must be initialized, and temporal dependencies must be learned despite uncertain initialization. OM supplies cross-variable and regional evidence before ATM processes the sequence. Mamba then provides an ordered temporal backbone, while experts adapt that backbone to heterogeneous patterns. This is a modeling design intended to reduce interference between smooth evolution and abrupt events, rather than a claim that either module guarantees perfect denoising or a unique decomposition.

D.3 Observation Mixer

Variable and station complementarity. A variable may be absent while other sensors at the same station remain available; an entire station segment may be absent while neighboring stations observe the same weather system. OM therefore first attends across variables and then across neighboring stations. With latent width d, the operations are

$$\begin{aligned} \mathbf{f} &= \mathrm{MLP}(\mathbf{X}_M) + \mathrm{STE}(\mathbf{X}_M) \in \mathbb{R}^{N \times T \times C \times d}, \\ \mathbf{f}^t_{\mathrm{vars}} &= \mathbf{f}_{:,t,:,:} \in \mathbb{R}^{N \times C \times d}, \\ \rho_{ij} &= \mathrm{dist}(\mathbf{s}_i, \mathbf{s}_j) \\ A_{ij} &= \begin{cases} \exp(-\rho_{ij}/\tau), & \rho_{ij} \le \tau, \\ 0, & \rho_{ij} > \tau, \end{cases} \\ \mathbf{f}^t_{\mathrm{self}} &= \mathrm{MHSA}_{\mathrm{var}}(\mathbf{f}^t_{\mathrm{vars}}) \\ \mathbf{f}^t_{\mathrm{adj}} &= \mathrm{GAT}_{\mathrm{sta}}(\mathbf{f}^t_{\mathrm{self}}, \mathbf{A}). \end{aligned} \tag{D.5}$$

STE denotes spatiotemporal encoding attached to the input positions; its input notation includes the associated metadata, not just numerical weather values. ρ_{ij} is spatial distance, and τ sets both the neighborhood cutoff and distance-decay scale. We use ρ_{ij} to avoid confusion with temporal distance d_t. Variable attention acts within a station and time step; graph attention exchanges station information using $\mathbf{A}$. The graph weights express proximity and should not themselves be interpreted as normalized attention coefficients.

Why a field representation? Pairwise station aggregation remains tied to discrete sensor locations. Weather processes extend through the regions between those locations, and uneven station coverage

leaves each station with only a partial view. Regional field tokens provide a common latent representation in which evidence from several stations can jointly constrain missing positions. They represent the spatially extended process without first interpolating weather observations onto a dense global measurement grid.

Let lat_i and lon_i denote station coordinates. A padded local bounding region and its query tokens are constructed as

$$\begin{aligned}
\phi_{\min} &= \min_i \mathrm{lat}_i - \delta, \qquad \phi_{\max} = \max_i \mathrm{lat}_i + \delta, \\
\lambda_{\min} &= \min_i \mathrm{lon}_i - \delta, \qquad \lambda_{\max} = \max_i \mathrm{lon}_i + \delta, \\
\mathbf{G}_s &= \mathrm{Grid}([\phi_{\min}, \phi_{\max}], [\lambda_{\min}, \lambda_{\max}], H, W), \\
\mathbf{z}_{\mathrm{field}} &= \mathrm{STE}(\mathbf{G}_s) \in \mathbb{R}^{H\times W\times d}, \qquad \mathbf{z}_{\mathrm{mem}} \in \mathbb{R}^{K\times d}, \\
\mathbf{F}_{\mathrm{adj}} &= \mathrm{Stack}_{t=1}^{T}\left(\mathbf{f}_{\mathrm{adj}}^{t}\right), \\
[\mathbf{z}'_{\mathrm{field}}; \mathbf{z}'_{\mathrm{mem}}] &= \mathrm{Ff}\left(q = [\mathbf{z}_{\mathrm{field}}; \mathbf{z}_{\mathrm{mem}}],\ k/v = \mathbf{F}_{\mathrm{adj}}\right).
\end{aligned} \tag{D.6}$$

Here δ denotes spatial padding, while H and W specify the resolution of the local field grid. The field representation is flattened into HW query tokens for attention. The K memory tokens are learnable parameters shared across regional inputs and provide global contextual capacity rather than measurements at additional stations. $\mathbf{F}_{\mathrm{adj}}$ stacks station features over the complete temporal window and serves as the key/value memory of the Field-former, allowing field and memory queries to aggregate regional evidence across time. Field and station tokens follow the same spatiotemporal encoding convention.

Field-former and inverse cross-attention. Ff denotes the Field-former: field and memory tokens query station features as keys and values. To make the operator roles explicit, the basic scaled dot-product attention operation is

$$\mathrm{Att}(Q, K, V) = \mathrm{Softmax}_{\mathrm{keys}}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V, \tag{D.7}$$

where Q, K, V are projected query, key, and value arrays and d_k is the key dimension. This equation explains the attention primitive; Ff denotes the complete learned block rather than an additional unexplained function.

The aggregated field representation is refined and retrieved at missing station positions:

$$\begin{aligned}
\mathbf{z}_{\mathrm{ref}} &= \mathrm{CNNs}(\mathbf{z}'_{\mathrm{field}}), \\
\mathbf{f}_{\mathrm{C}} &= \mathrm{ICAtt}(q = \mathbf{f}_{\mathrm{mask}},\ k/v = \mathbf{z}_{\mathrm{ref}}).
\end{aligned} \tag{D.8}$$

CNNs is the convolutional refinement block that operates on the spatial field representation to capture local spatial continuity. The convolutional input remains a two-dimensional field feature map, while memory tokens are not involved in this refinement operation. $\mathbf{f}_{\mathrm{mask}}$ contains feature queries associated with the positions to reconstruct. ICAtt reverses the information flow of the Field-former: station queries retrieve regional evidence from the refined field representation. "Inverse" refers to this retrieval direction rather than an algebraic inverse of attention. The complemented sequence $\mathbf{f}_{\mathrm{C}}$ is then passed to ATM. Figure C.1(a) illustrates this station-to-field-to-station path.

D.4 Event Aligner

Weather events can reach neighboring stations at different times. Aggregating the same timestamp may therefore combine the rising phase at one station with the peak or recovery phase at another. EA estimates temporal shifts before neighboring features are aggregated (Fig. C.1(b)).

For the station signals supplied to the delay estimator, define the FFT-based cross-correlation

$$\mathrm{Corr}(\mathbf{X}_i, \mathbf{X}_j)[\ell] = \left[\mathcal{F}^{-1}\left(\mathcal{F}(\mathbf{X}_i) \odot \overline{\mathcal{F}(\mathbf{X}_j)}\right)\right]_{\ell}. \tag{D.9}$$

The transforms act along time; ℓ is a temporal lag and the overbar denotes complex conjugation. This is the cross-correlation used for delay estimation, not a Pearson correlation coefficient. The alignment

and aggregation operations are

$$\begin{aligned} D_{ij} &= \arg\max_{\ell} \mathrm{Corr}(\mathbf{X}_i, \mathbf{X}_j)[\ell], \\ \mathbf{f}_a &= \mathrm{Shift}(\mathbf{f}_{\mathrm{C}}, \mathbf{D}) \in \mathbb{R}^{N\times T\times N\times C\times d}, \\ \mathbf{f}_{\mathrm{ego}} &= \mathrm{Diag}(\mathbf{f}_a) \in \mathbb{R}^{N\times T\times C\times d} \\ \mathbf{f}_{\mathrm{nbr}} &= \mathrm{OffDiag}(\mathbf{f}_a), \\ \mathbf{f}_{\mathrm{S}} &= \mathrm{MHA}(Q = \mathbf{f}_{\mathrm{ego}},\ K/V = \mathbf{f}_{\mathrm{nbr}};\ \mathrm{mask} = \mathbf{M}_s). \end{aligned} \tag{D.10}$$

D_{ij} is an integer lag in time-step units. In $\mathbf{f}_a$, the first station axis identifies the target station and the second identifies the neighbor whose features are aligned to it. Diag selects matching target/source station indices, while OffDiag collects other stations. $\mathbf{M}_s$ is the validity mask for station-attention entries. Shift includes temporal reindexing and the handling of valid shifted positions. We denote neighboring aligned features by $\mathbf{f}_{\mathrm{nbr}}$.

Complete temporal targets are used to estimate posterior delays during training. EA operates only in the posterior branch to obtain inter-station event-aligned representations, which guide the conditional prior through posterior-prior alignment. At inference, EA and the target-conditioned posterior are absent.

D.5 Reliability-weighted Temporal Modeling

ATM couples a Mamba temporal backbone with PD-MoE adaptation. The state-space stage compresses sequential evidence into a recurrent representation, aiming to retain persistent dependencies while filtering perturbations. The expert stage then handles patterns that a common sequential representation may not fully distinguish. This avoids asking one global attention operation to account simultaneously for large extreme values and smooth ordered evolution.

Observation-distance weighting. Even after OM, features in long gaps are inferred evidence rather than direct measurements. Unweighted state updates could propagate their errors through the temporal representation. For a token sequence with nonempty observed-index set $\mathcal{T}_{\mathrm{obs}}$, ATM computes

$$\begin{aligned} d_t &= \begin{cases} \min\limits_{t'\in\mathcal{T}_{\mathrm{obs}}} |t-t'|, & \mathcal{T}_{\mathrm{obs}} \neq \varnothing, \\ T, & \mathcal{T}_{\mathrm{obs}} = \varnothing. \end{cases}, \\ \mathbf{U}_s^t &= e^{[-\,\mathrm{ReLU}(\mathrm{MLP}(d_t))]}, \\ \tilde{\mathbf{f}}_{\mathrm{C}}^t &= \mathbf{U}_s^t \odot \mathbf{f}_{\mathrm{C}}^t. \end{aligned} \tag{D.11}$$

Station and variable indices are suppressed. d_t measures distance in time steps. For fully masked sequences, we set the temporal distance to the finite fallback value T, consistent with the implementation. and the output of the learned map is broadcast or matched to the input-feature dimension. Each component of $\mathbf{U}_s^t$ lies in $(0, 1]$ for finite network outputs. Smaller weights attenuate input evidence more strongly. Although the module was termed an uncertainty score, $\mathbf{U}_s^t$ functions as a reliability weight; it is distinct from the CVAE variance used for sampling.

Distance supplies a cue about observation support. The stated unconstrained MLP does not by itself guarantee that weights decrease monotonically with distance, nor that an observed position receives exactly unit weight. Weighting is intended to protect the recurrent representation from unreliable initialization; its benefit is empirical rather than a hard noise-removal guarantee.

Bidirectional state-space encoding. The weighted sequence is processed in both temporal directions:

$$\begin{aligned} \mathbf{a}_t^{\rightarrow} &= \mathrm{SSM}_f(\tilde{\mathbf{f}}_{\mathrm{C}}^1, \ldots, \tilde{\mathbf{f}}_{\mathrm{C}}^t), \\ \mathbf{a}_t^{\leftarrow} &= \mathrm{SSM}_b(\tilde{\mathbf{f}}_{\mathrm{C}}^T, \ldots, \tilde{\mathbf{f}}_{\mathrm{C}}^t), \\ \mathbf{h}_t &= \mathrm{MLP}([\mathbf{a}_t^{\rightarrow}; \mathbf{a}_t^{\leftarrow}]) \in \mathbb{R}^d \\ \mathbf{h}_{1:T} &\in \mathbb{R}^{T\times d}. \end{aligned} \tag{D.12}$$

SSM_f and SSM_b are the forward and backward Mamba state-space passes. We distinguish their internal representations $\mathbf{a}$ from the projected pre-expert feature $\mathbf{h}_t$ used in the main text. The

projection combines both directions into width d for each station-variable token sequence. Missing future entries remain masked inputs, so a backward pass does not imply access to future ground truth.

D.6 Pattern-decoupled Expert Adaptation

Global pattern association after sequential modeling. PD-MoE applies N_e separately parameterized experts to the temporal backbone. Expert parameters are shared across stations and time positions, so similar states can reuse transformations even when they occur far apart. This provides association through shared pattern transformations rather than explicit all-to-all exchange of token values. Separate experts reduce the pressure to fit stable evolution and abrupt extremes with one transformation; their specialization is learned rather than prescribed by extreme-event labels.

Joint pattern and geographic modulation. Pattern information alone is insufficient because similar temporal fluctuations may have different implications across regions, while geography alone cannot characterize the evolving dynamical states of a station. We therefore jointly condition expert routing on temporal patterns and geographic context.

$$
\begin{aligned}
\mathbf{p}_t &= P(\mathbf{h}_t) \in \mathbb{R}^{N_e \times d_p}, \qquad \mathbf{g} = G(\mathbf{I}) \in \mathbb{R}^{S_e \times d_p}, \\
\alpha_s &= \frac{\|\mathbf{g}_{s,:}\|_2 + \epsilon}{\sum_{k=1}^{S_e} (\|\mathbf{g}_{k,:}\|_2 + \epsilon)}, \qquad \bar{\mathbf{g}} = \sum_{s=1}^{S_e} \alpha_s \mathbf{g}_{s,:}, \\
[\mathbf{w}_g, \mathbf{b}_g] &= \mathrm{MLP}_g(\bar{\mathbf{g}}), \\
r_{t,e} &= \frac{\exp\left(\frac{\langle \mathbf{p}_{t,e,:}, \mathbf{w}_g \rangle}{\sqrt{d_p}} + [\mathbf{b}_g]_e\right)}{\sum_{j=1}^{N_e} \exp\left(\frac{\langle \mathbf{p}_{t,j,:}, \mathbf{w}_g \rangle}{\sqrt{d_p}} + [\mathbf{b}_g]_j\right)}.
\end{aligned}
\tag{D.13}
$$

Here N_e is the number of experts, S_e is the number of geographic subspaces, and d_p is the shared routing dimension. The ℓ_2-norm of each geographic subspace is used as its activation strength and normalized into α_s to aggregate multiple geographic representations. The geographic routing module $\mathcal{R}_g$ then maps the aggregated representation $\bar{\mathbf{g}}$ through MLP_g to generate the location-conditioned projection vector $\mathbf{w}_g$ and expert-specific bias $\mathbf{b}_g$. The routing distribution $r_{t,e}$ is obtained by evaluating each expert-specific pattern representation $\mathbf{p}_{t,e,:}$ under the geographic context, thereby jointly modeling temporal patterns and spatial information.

Reliability-dependent expert selection. With weaker observation support, multiple weather patterns may remain plausible. Activating more experts allows diverse learned transformations to contribute, while well-supported features can use fewer experts for focused adaptation. The reliability-dependent expert selection and aggregation are defined as

$$
\begin{aligned}
E_t = \mathrm{Relative}(U_s^t) &= \max\left(1, \mathrm{int}((1 - U_s^t) N_e)\right) \\
\mathcal{E}_t &= \mathrm{TopK}(\mathbf{r}_t, E_t), \\
\mathbf{f}_{\mathrm{T}}^t &= \sum_{e \in \mathcal{E}_t} r_{t,e}\, \mathrm{MLP}_e(\mathbf{h}_t), \\
\mathbf{f}_{\mathrm{T}} &= \mathrm{Stack}_{t=1}^{T}(\mathbf{f}_{\mathrm{T}}^t).
\end{aligned}
\tag{D.14}
$$

Here $\mathrm{Relative}$ maps the reliability score $\mathbf{U}_s^t$ to an adaptive integer expert count. Less reliable inputs activate more experts to preserve diverse candidate weather patterns, whereas more reliable inputs use fewer experts for focused adaptation. TopK selects the activated expert set according to the routing distribution $\mathbf{r}_t$, and the selected routing weights directly aggregate the outputs of the corresponding experts. Each MLP_e represents an expert-specific transformation.

Thus, temporal patterns and geographic context determine the expert allocation, while reliability controls the sparsity level of expert activation. Reliability weighting before Mamba and reliability-dependent expert selection after Mamba serve different purposes: the former attenuates uncertain

input evidence, whereas the latter adjusts representational flexibility by controlling the number of active experts. The adaptive expert count should therefore not be interpreted as a calibrated probability of extreme events.

D.7 Local Refiner and Latent Sampling

Expert transformations operate token-wise, and independent latent perturbations may weaken local continuity. LR recovers neighborhood structure through temporal convolutions with several receptive fields, then uses temporal features to select their contributions.

Sampling source. Training reconstruction uses posterior samples; inference uses the conditional prior. With independent standard-normal noise $\boldsymbol{\epsilon}$, reparameterized samples are

$$\begin{aligned} \mathbf{z}_{\text{train}} &= \boldsymbol{\mu}_q + \boldsymbol{\sigma}_q \odot \boldsymbol{\epsilon}, \\ \mathbf{z}_{\text{infer}} &= \boldsymbol{\mu}_p + \boldsymbol{\alpha} \odot \boldsymbol{\sigma}_p \odot \boldsymbol{\epsilon}, \\ \boldsymbol{\epsilon} &\sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d). \end{aligned} \tag{D.15}$$

$\boldsymbol{\alpha} \in \mathbb{R}^{N\times T\times C}$ is broadcast over latent channels and controls pointwise perturbation strength. Unit strength samples from the learned prior; zero strength decodes its latent mean. The nonlinear decoder means that decoding the latent mean need not equal the predictive mean. Changing $\boldsymbol{\alpha}$ changes latent variability, rather than directly specifying a weather value or guaranteeing an extreme event.

Multi-scale local reconstruction. For $\mathbf{z} \in \mathbb{R}^{N\times T\times C\times d}$ and receptive fields $\mathcal{R} = \{r_1, \ldots, r_{N_r}\}$,

$$\begin{aligned} \mathbf{z}_c &= \text{Concat}_{j=1}^{N_r} \left[\text{Conv}(\mathbf{z}, r_j) \right], \\ \mathbf{S} &= \text{Softmax}(\text{SG}(\mathbf{f}_{\text{T}})), \\ \hat{\mathbf{X}} &= \text{MLP}_{\text{dec}}([\mathbf{f}_{\text{T}}; \mathbf{S} \odot \mathbf{z}_c]) \in \mathbb{R}^{N\times T\times C}. \end{aligned} \tag{D.16}$$

$\text{Conv}(\cdot, r_j)$ extracts temporal neighborhoods at scale r_j. In the equal-width construction, each branch outputs d/N_r channels, and their concatenation preserves the feature width d. SG denotes the selective MLP gate conditioned on temporal features, whose output weights follow the same concatenated layout as $\mathbf{z}_c$. The decoder combines deterministic temporal context from $\mathbf{f}_{\text{T}}$ with gated stochastic local features $\mathbf{z}_c$, enabling reconstruction by jointly modeling stable temporal evolution and local uncertain variations.

Interpreting variance. The prior standard deviation controls the magnitude of latent perturbations before decoding, thereby adjusting generation diversity under uncertain conditions. Different variables and time periods may exhibit different responses to sampling strength, reflecting heterogeneous uncertainty in the learned dynamics. By controlling latent perturbations, UniGIO can sample diverse plausible states and generate potential extreme events at arbitrary time-variable locations. However, latent variance itself is not a calibrated extreme-event probability, since extreme outcomes are jointly determined by the decoder mapping and the event threshold.

D.8 Training Objective

UniGIO combines posterior reconstruction, conditional-prior alignment, and orthonormal regularization. $\mathcal{O}$ denote the incomplete conditioning information:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\text{rec}} + \beta \mathcal{L}_{\text{KL}} + \gamma \mathcal{L}_{\text{orth}}, \\ \mathcal{L}_{\text{rec}} &= -\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{X},\mathbf{I})} \left[\log p_\theta(\mathbf{X} \mid \mathbf{z}, \mathcal{O}) \right], \\ \mathcal{L}_{\text{KL}} &= D_{\text{KL}} \left(q_\phi(\mathbf{z} \mid \mathbf{X}, \mathbf{I}) \parallel p_\psi(\mathbf{z} \mid \mathcal{O}) \right). \end{aligned} \tag{D.17}$$

β and γ weight conditional-prior alignment and orthonormal regularization, respectively. The decoder is conditioned on $\mathcal{O}$ through the deterministic temporal representation $\mathbf{f}_{\text{T}}$ rather than through an additional input branch. Reconstruction is supervised only where target values are available.

For diagonal Gaussian prior and posterior distributions, the KL term before any implementation-specific batch reduction is

$$\mathcal{L}_{\text{KL}} = \frac{1}{2} \sum_k \left[\log \frac{\sigma_{p,k}^2}{\sigma_{q,k}^2} + \frac{\sigma_{q,k}^2 + (\mu_{q,k} - \mu_{p,k})^2}{\sigma_{p,k}^2} - 1 \right], \tag{D.18}$$

where k indexes latent coordinates and $\sigma_{p,k}, \sigma_{q,k} > 0$ are standard deviations. This closed-form term encourages the incomplete-input conditional prior to approximate the posterior inferred from complete target sequences.

The orthonormal penalty is applied to the designated pattern, geographic, and memory parameter bases. For a basis matrix $\mathbf{B} \in \mathbb{R}^{m \times k}$ whose columns are basis vectors, its contribution is

$$\mathcal{L}_{\text{orth}}(\mathbf{B}) = \left\| \mathbf{B}^\top \mathbf{B} - \mathbf{I}_k \right\|_F^2, \tag{D.19}$$

where $\mathbf{I}_k$ denotes the $k \times k$ identity matrix. $\mathcal{L}_{\text{orth}}$ aggregates these contributions over the constrained parameter groups. This notation distinguishes the parameter basis $\mathbf{B}$ from the input-dependent pattern representation $P(\mathbf{h}_t)$. The penalty encourages approximately orthogonal and normalized columns within each basis matrix; it acts as a soft regularizer and does not imply statistical independence between expert outputs.

E BENCHMARK AND SETUP

E.1 DATASET AND EVALUATION SCOPE

GIO-U is constructed from Weather-5K (Han et al., 2026), a large-scale global station weather benchmark derived from NOAA's Integrated Surface Database (ISD). It contains hourly in-situ observations of temperature, dew point, wind rate, wind direction, and sea-level pressure from 5,672 stations worldwide during 2014–2024. The dataset provides broad spatial coverage, long temporal span, and diverse meteorological variables with 8 years for training and 2 years for testing. ISD integrates observations from a large and heterogeneous collection of global meteorological sources, while Weather-5K offers substantially broader station coverage, temporal span, and variable diversity than existing station forecasting benchmarks. Therefore, although our main experiments are conducted on a single benchmark, Weather-5K provides a comprehensive and representative testbed for GSWF to cover the major station forecasting settings represented by existing datasets, including missing observation reconstruction, temporal forecasting, spatial generation, and extreme-event evaluation. The main experiments use 120h windows, while 72h and 240h windows further examine sensitivity to temporal context. All settings evaluate pointwise accuracy, distributional fidelity, and extreme-event capture; task-specific comparisons emphasize accuracy and F1-SEDI.

E.2 CURATED MASKS, NATIVE MASKS, AND SCORING SUPPORT

Native masks record missingness already present in station observations. Curated masks deliberately hide observed values across time, variables, or stations, making their retained values available as training targets. Training uses curated masks because genuinely missing values do not supply supervised targets in the real world. Evaluation under native masks uses the trained model without additional training to assess transfer to real-world incompleteness. Curated masks also enable the controlled task comparisons in Appendix B.

Let $\mathcal{E}$ denote the set of masked entries to be predicted:

$$\mathcal{E} = \{(n, t, c) : M_{n,t,c} = 0\}. \tag{E.1}$$

E.3 BASELINE FAMILIES AND ADAPTATIONS

General time-series generators. TimeGAN (Yoon et al., 2019) and TimeVAE (Desai et al., 2021) represent adversarial and variational generation. DiffWave (Kong et al., 2021), SSSD (Alcaraz & Strodthoff, 2022), and Diffusion-TS (Yuan & Qiao, 2024) represent diffusion-based approaches. Together, these baselines test whether general temporal generative modeling is sufficient for incomplete GIO. Their station-variable treatment is contrasted with UniGIO's explicit complementary observation modeling.

Interaction-aware generators. UniMTD (Yang et al., 2025b) and SportsTraj (Xu & Fu, 2025) originally model interacting trajectories. Their reimplementations adapt the interaction representation to station and variable dependencies. These comparisons evaluate whether interaction-aware generation transfers to weather observations. The † marker in result tables identifies reimplemented methods; it should not be interpreted as an unchanged original benchmark configuration.

Task-specific and pretrained comparisons. Forecasting comparisons include WSSM (Yang et al., 2025a), iTransformer (Liu et al., 2024), Corrformer (Wu et al., 2023), and the time-series foundation model Chronos2 (Ansari et al., 2024). Imputation comparisons include specialized methods (Liu et al., 2019; Xu et al., 2023). The 24h experiment additionally compares the fine-tuned observation foundation model ESFM (Ozdemir et al., 2026) and operational HRES (European Centre for Medium-Range Weather Forecasts, 2024) under the ESFM evaluation setting. HRES is an operational numerical weather prediction product rather than a generative time-series baseline. These systems differ in their original training data and input scope, so the comparison measures predictive performance in the reported setting rather than equal-compute training efficiency.

Deterministic control. UniGIO without CVAE removes its generative structure to assess the contribution of latent-variable modeling. This differs from setting sampling strength to zero: the latter retains a trained CVAE and decodes the prior latent mean.

E.4 Accuracy Metrics

With i indexing eligible entries in $\mathcal{E}$, accuracy is measured by

$$\begin{aligned} \mathrm{MAE} &= \frac{1}{|\mathcal{E}|}\sum_{i\in\mathcal{E}}|\hat{X}_i - X_i| \\ \mathrm{MSE} &= \frac{1}{|\mathcal{E}|}\sum_{i\in\mathcal{E}}(\hat{X}_i - X_i)^2. \end{aligned} \tag{E.2}$$

For a complete reference window, this reduces to the original masked-region formula with weights $1 - M_i$. MAE emphasizes absolute pointwise deviations, whereas MSE gives larger errors greater weight. Both are lower-is-better, and results are reported separately by weather variable.

E.5 Extreme-event Capture: F1-SEDI

F1-SEDI combines the recall-style quantity used in this benchmark with precision. It is an F1-form evaluation of tail-event detection, accounting for both missed events and false alarms. Here the recall term is written R to make the implemented count-based definition explicit; the name F1-SEDI follows the main text.

For each variable and threshold setting p, let Q^p_{lower} and Q^p_{upper} denote its lower and upper tail thresholds. Define matching-tail hits, predicted events, and reference events on the scoring support:

$$\begin{aligned} H_p &= \sum_{i\in\mathcal{E}}\left[\mathbf{1}\{\hat{X}_i < Q^p_{\mathrm{lower}}, X_i < Q^p_{\mathrm{lower}}\} + \mathbf{1}\{\hat{X}_i > Q^p_{\mathrm{upper}}, X_i > Q^p_{\mathrm{upper}}\}\right], \\ P_p &= \sum_{i\in\mathcal{E}}\left[\mathbf{1}\{\hat{X}_i < Q^p_{\mathrm{lower}}\} + \mathbf{1}\{\hat{X}_i > Q^p_{\mathrm{upper}}\}\right], \\ O_p &= \sum_{i\in\mathcal{E}}\left[\mathbf{1}\{X_i < Q^p_{\mathrm{lower}}\} + \mathbf{1}\{X_i > Q^p_{\mathrm{upper}}\}\right]. \end{aligned} \tag{E.3}$$

For nonzero denominators,

$$\begin{aligned} \mathrm{Precision}_p &= \frac{H_p}{P_p} \\ R_p &= \frac{H_p}{O_p} \\ \text{F1-SEDI}_p &= \frac{2\,\mathrm{Precision}_p\, R_p}{\mathrm{Precision}_p + R_p} = \frac{2H_p}{P_p + O_p}. \end{aligned} \tag{E.4}$$

The count form is defined whenever $P_p + O_p > 0$, including zero-hit cases. A predicted upper-tail event does not count as a hit for a lower-tail reference event, or vice versa. This preserves the same-tail matching in the original formulation. The reported threshold settings are the 99.5th and 90.0th percentiles. Higher F1-SEDI indicates better event capture. Wind direction is excluded because its circular scale does not admit the same ordered lower/upper-tail interpretation.

E.6 Distributional Fidelity

FID. Fréchet Inception Distance (Heusel et al., 2017; Yuan & Qiao, 2024) compares the means and covariances of reference and generated sequence features:

$$\mathrm{FID} = \|\boldsymbol{\mu}_r - \boldsymbol{\mu}_f\|_2^2 + \mathrm{Tr}\left(\boldsymbol{\Sigma}_r + \boldsymbol{\Sigma}_f - 2(\boldsymbol{\Sigma}_r\boldsymbol{\Sigma}_f)^{1/2}\right). \tag{E.5}$$

The subscripts r and f denote reference and generated features. Tr is the matrix trace. The Gaussian feature approximation measures distributional separation and complements pointwise errors; the metric name alone does not specify the time-series feature extractor.

STE. Statistical error compares local temporal variation:

$$\mathrm{STE} = \frac{1}{|\mathcal{E}|}\sum_{i\in\mathcal{E}}\left|\sigma^w_r(X_i) - \sigma^w_f(\hat{X}_i)\right|. \tag{E.6}$$

σ^w denotes standard deviation in temporal window w associated with an eligible entry. STE penalizes discrepancies in fluctuation magnitude, including both excessive smoothing and excessive variation. It does not require the fluctuations to occur at the same individual timestamps, which is why accuracy is also reported.

CroS. Correlation Score (Yoon et al., 2019; Yuan & Qiao, 2024) compares cross-variable correlations. Let ρ^r and ρ^f denote the reference and generated variable-pair correlations:

$$\begin{aligned} \rho^r &= \mathbb{E}_{N,T}[\tilde{\mathbf{X}} \odot \tilde{\mathbf{X}}'], \\ \rho^f &= \mathbb{E}_{N,T}[\tilde{\hat{\mathbf{X}}} \odot \tilde{\hat{\mathbf{X}}}'], \\ \mathrm{CroS} &= \sum_i |\rho_i^f - \rho_i^r|. \end{aligned} \tag{E.7}$$

Tildes denote normalized data, and the prime denotes the corresponding lower-triangular variable-pair arrangement rather than a temporal derivative. Expectations are taken over station and time axes. The resulting correlation representation is defined on variable pairs, and CroS measures the discrepancy between generated and reference cross-variable dependencies. Lower CroS indicates better preservation of inter-variable consistency.

E.7 Spatial Transfer Protocol

The cross-region experiment uses 120h windows and targets South America and Africa, focusing on temperature, wind rate, and wind direction. The worldwide comparison uses a random station split. Cross-region pretraining excludes the target region; zero-shot evaluation generates all target-region stations. Additional settings reserve 90%, 60%, or 30% of target-region stations for generation and use the remaining 10%, 40%, or 70% for fine-tuning. The percentages in Table 9 therefore refer to generation stations, not fine-tuning stations. This regional fine-tuning experiment is distinct from zero-shot transfer between curated and native masks.

F Additional Time-window Results

Tables F.1–F.3 report 72h and 240h results. Each cell lists curated/native performance, while the window column identifies the temporal setting. Red and blue mark the best and second-best results within the corresponding comparison. The tables retain the original numerical values and reporting precision.

Longer windows generally increase reconstruction difficulty, but UniGIO remains competitive across all three evaluation aspects. At 72h, its accuracy/F1-SEDI/fidelity gains are 8.8%/13.5%, 3.6%/3.6%, and 7.1%/5.3%, increasing to 11.1%/10.0%, 5.1%/6.0%, and 21.4%/16.4% at 240h under curated/native masks. Improved CroS at longer windows is consistent with richer context for estimating inter-variable relationships, even when pointwise accuracy deteriorates.

Table F.1: Quantitative Accuracy Comparison on Curated / Native Masks (72h/240h) † means re-implemented

Baselines	Window	Temperature		Dewpoint		Wind Rate		Wind Direc.		Sea-Level Pressure	
		MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓	MAE↓	MSE↓ ($\times 10^3$)	MAE↓	MSE↓
TimeGAN	72h	5.24 / 9.42	33.2 / 63.4	3.33 / 5.71	16.3 / 28.7	2.49 / 4.20	8.05 / 13.8	97.0 / 136	11.4 / 15.9	4.55 / 6.49	31.1 / 45.0
	240h	5.88 / 11.2	43.5 / 88.8	4.51 / 8.18	28.9 / 54.4	2.82 / 5.16	10.4 / 19.8	111 / 165	13.9 / 20.8	6.66 / 10.2	63.9 / 101
TimeVAE	72h	4.41 / 3.36	23.2 / 17.7	3.00 / 2.28	12.9 / 9.87	2.03 / 1.54	5.52 / 4.20	93.0 / 70.9	10.1 / 7.66	4.24 / 3.23	25.8 / 19.6
	240h	4.88 / 11.6	29.7 / 70.6	3.94 / 9.34	21.8 / 51.6	2.20 / 5.22	6.75 / 16.0	103 / 244	11.4 / 27.2	6.00 / 14.2	50.4 / 119
UniMTD †	72h	2.00 / 2.03	8.96 / 10.4	2.04 / 1.99	9.94 / 10.7	1.36 / 1.28	4.05 / 4.21	66.2 / 61.1	8.21 / 7.08	2.65 / 2.35	19.8 / 17.2
	240h	2.79 / 2.66	16.4 / 17.1	2.97 / 2.71	19.3 / 18.5	1.60 / 1.47	5.46 / 5.48	77.0 / 69.7	9.94 / 8.53	4.54 / 3.64	47.1 / 36.3
SportsTraj †	72h	1.65 / 1.98	5.18 / 6.66	1.80 / 2.08	6.69 / 8.19	1.54 / 1.69	4.34 / 4.62	75.8 / 77.0	9.96 / 8.45	1.03 / 1.14	2.48 / 2.50
	240h	2.37 / 2.82	10.3 / 12.0	2.28 / 2.68	10.4 / 12.2	1.77 / 1.99	5.71 / 6.16	86.5 / 87.2	11.9 / 9.64	1.92 / 2.15	7.26 / 6.68
UniGIO	72h	1.26 / 1.37	4.15 / 5.56	1.32 / 1.33	4.80 / 5.93	1.22 / 1.14	3.34 / 3.47	63.1 / 56.1	8.47 / 6.60	0.75 / 0.81	1.91 / 2.02
	240h	1.72 / 1.71	7.34 / 7.63	1.76 / 1.65	8.16 / 7.80	1.36 / 1.28	4.12 / 3.32	68.9 / 62.3	9.15 / 8.52	1.47 / 1.27	8.21 / 5.87

Table F.2: Quantitative F1-SEDI Comparison on Curated / Native Masks (72h/240h) † means re-implemented

Baselines	Window	Temperature		Dewpoint		Wind Rate		Wind Direc.		Sea-Level Pressure	
		99.5th↑	90.0th↑	99.5th↑	90.0th↑	99.5th↑	90.0th↑	99.5th↑	90.0th↑	99.5th↑	90.0th↑
TimeGAN	72h	0.15 / 0.23	0.49 / 0.51	0.28 / 0.34	0.65 / 0.66	0.04 / 0.08	0.24 / 0.29	/	/	0.35 / 0.35	0.62 / 0.63
	240h	0.04 / 0.06	0.39 / 0.40	0.10 / 0.14	0.48 / 0.50	0.02 / 0.05	0.15 / 0.20	/	/	0.03 / 0.04	0.36 / 0.36
TimeVAE	72h	0.15 / 0.16	0.50 / 0.51	0.27 / 0.26	0.66 / 0.66	0.03 / 0.02	0.19 / 0.19	/	/	0.33 / 0.33	0.63 / 0.63
	240h	0.03 / 0.03	0.40 / 0.39	0.05 / 0.06	0.50 / 0.50	0.02 / 0.01	0.09 / 0.10	/	/	0.01 / 0.01	0.34 / 0.34
UniMTD †	72h	0.48 / 0.49	0.75 / 0.75	0.37 / 0.45	0.68 / 0.71	0.17 / 0.24	0.37 / 0.46	/	/	0.51 / 0.55	0.70 / 0.72
	240h	0.29 / 0.36	0.65 / 0.67	0.21 / 0.32	0.54 / 0.60	0.11 / 0.20	0.27 / 0.41	/	/	0.32 / 0.38	0.47 / 0.57
SportsTraj †	72h	0.54 / 0.56	0.80 / 0.82	0.43 / 0.48	0.74 / 0.77	0.23 / 0.31	0.42 / 0.52	/	/	0.80 / 0.80	0.89 / 0.89
	240h	0.40 / 0.42	0.72 / 0.75	0.30 / 0.39	0.68 / 0.72	0.17 / 0.26	0.31 / 0.48	/	/	0.63 / 0.64	0.80 / 0.81
UniGIO	72h	0.60 / 0.63	0.83 / 0.85	0.51 / 0.58	0.78 / 0.81	0.22 / 0.28	0.43 / 0.55	/	/	0.85 / 0.83	0.92 / 0.91
	240h	0.50 / 0.53	0.76 / 0.80	0.42 / 0.48	0.72 / 0.77	0.15 / 0.24	0.35 / 0.52	/	/	0.70 / 0.74	0.82 / 0.87

Table F.3: Quantitative Fidelity Comparison on Curated / Native Masks (72h/240h) † means re-implemented

Baselines	Window	Temperature		Dewpoint		Wind Rate		Wind Direc.		Sea-Level Pressure		CroS↓
		FID↓	STE↓	FID↓	STE↓	FID↓	STE↓	FID↓	STE↓ ($\times 10^3$)	FID↓	STE↓	
TimeGAN	72h	0.47 / 0.39	5.67 / 5.54	0.42 / 0.34	2.16 / 2.15	0.42 / 0.32	1.47 / 1.49	0.90 / 0.91	3.78 / 3.78	1.18 / 1.06	2.03 / 2.00	3.42 / 3.50
	240h	3.24 / 2.61	5.76 / 5.65	3.01 / 2.38	2.57 / 2.51	3.01 / 2.28	1.53 / 1.54	3.42 / 2.57	3.91 / 3.89	5.65 / 4.77	3.26 / 3.23	2.74 / 2.82
TimeVAE	72h	0.28 / 0.26	5.34 / 5.11	0.75 / 0.67	2.52 / 2.27	0.39 / 0.37	1.24 / 1.19	0.89 / 0.86	3.06 / 3.11	1.56 / 1.69	1.31 / 1.44	4.53 / 4.77
	240h	2.74 / 2.43	6.72 / 6.74	3.04 / 3.37	2.48 / 2.42	4.11 / 3.80	1.85 / 1.80	4.06 / 3.93	4.26 / 4.13	6.15 / 4.90	1.68 / 1.61	4.24 / 4.25
UniMTD †	72h	0.19 / 0.22	3.07 / 2.03	0.31 / 0.17	1.53 / 1.16	0.26 / 0.30	1.02 / 0.66	0.56 / 0.67	2.61 / 1.59	0.15 / 0.12	1.26 / 0.87	2.56 / 1.52
	240h	1.82 / 0.97	3.85 / 2.46	1.07 / 0.64	1.74 / 1.33	1.66 / 0.99	1.13 / 0.72	2.03 / 1.46	2.86 / 1.73	1.23 / 0.92	1.46 / 0.99	2.18 / 1.21
SportsTraj †	72h	0.12 / 0.18	2.59 / 1.97	0.32 / 0.17	2.00 / 1.55	0.35 / 0.31	1.19 / 0.90	0.61 / 0.74	2.78 / 1.95	0.12 / 0.10	0.84 / 0.68	1.70 / 1.05
	240h	1.90 / 0.99	3.90 / 2.83	2.28 / 1.27	2.83 / 2.05	2.87 / 1.67	1.55 / 1.15	4.30 / 2.48	3.48 / 2.26	1.52 / 1.02	2.16 / 1.52	1.24 / 0.81
UniGIO	72h	0.14 / 0.16	2.20 / 1.55	0.19 / 0.15	1.38 / 1.02	0.22 / 0.28	0.96 / 0.64	0.50 / 0.54	2.41 / 1.52	0.07 / 0.05	0.61 / 0.51	1.63 / 0.94
	240h	0.30 / 0.32	2.86 / 1.78	0.43 / 0.45	1.53 / 1.17	0.50 / 0.44	1.10 / 0.69	1.02 / 0.97	2.67 / 1.62	0.17 / 0.12	1.05 / 0.65	1.26 / 0.69

G FULL QUALITATIVE COMPARISONS

The following figures expand the compact main-text visualization using the full comparison panels supplied with the manuscript. Columns show temperature, dew point, wind direction, wind rate, and sea-level pressure; rows compare ground truth (GT), UniGIO (Ours), SportsTraj, UniMTD, Diffusion-TS, and TimeGAN. These examples illustrate temporal behavior and complement the aggregate metrics; they do not introduce additional quantitative experiments.

G.1 FORECASTING

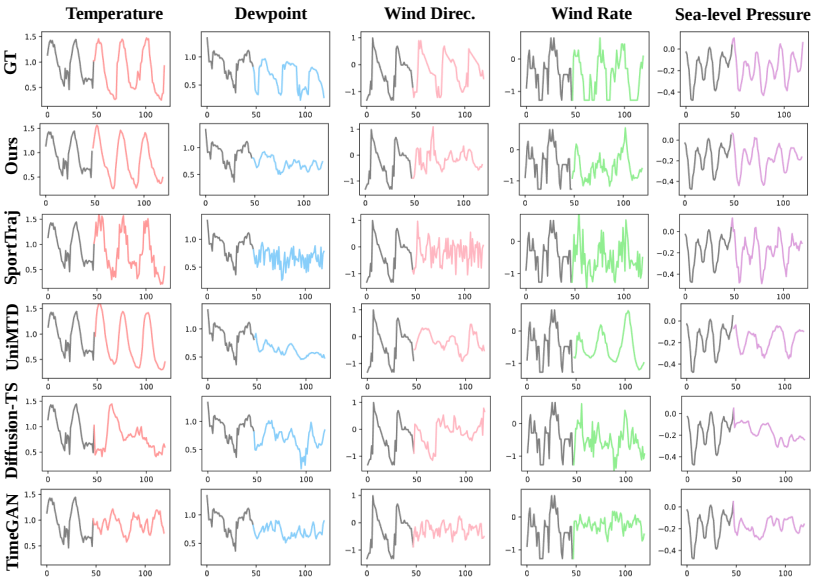

Figure G.1: Full forecasting comparison across all five weather variables. The panel labeled Ours denotes UniGIO; model and ground-truth labels follow the original visualization.

G.2 IMPUTATION

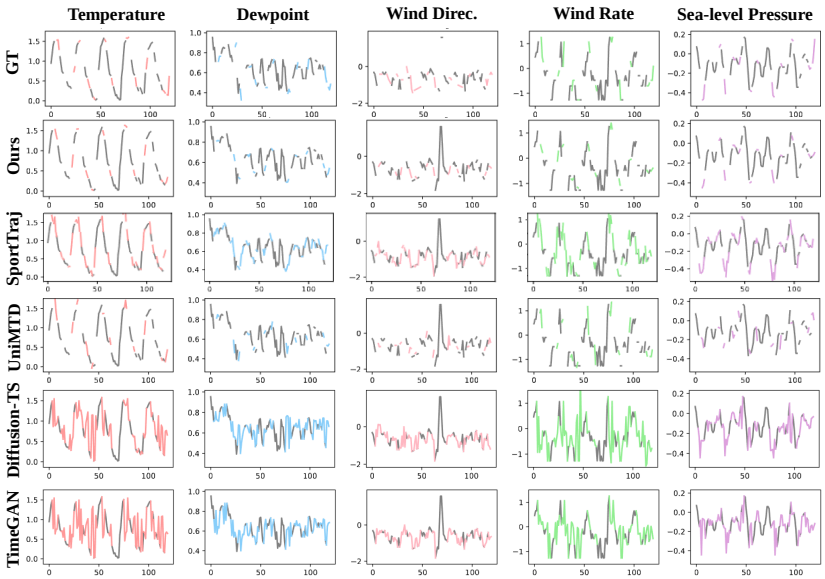

Figure G.2: Full imputation comparison across all five weather variables. The panel labeled Ours denotes UniGIO; model and ground-truth labels follow the original visualization.

G.3 GENERATION

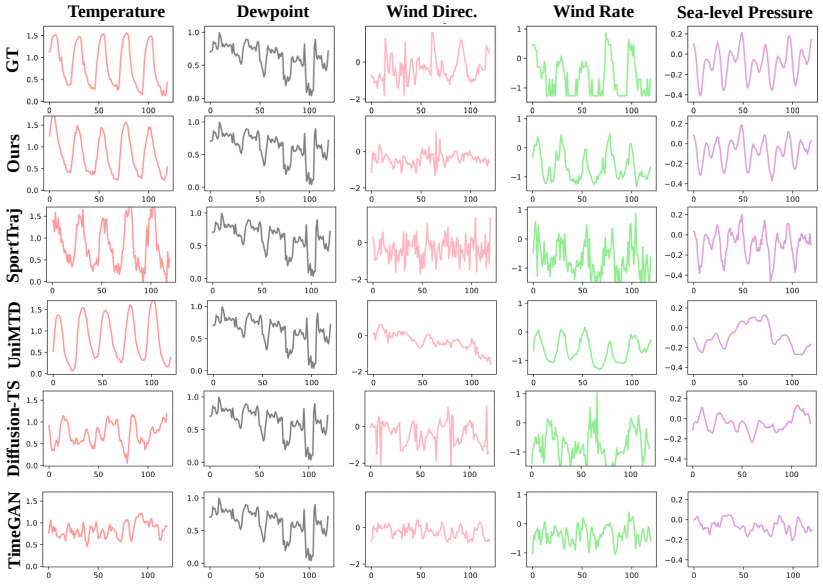

Figure G.3: Full generation comparison across all five weather variables. The panel labeled Ours denotes UniGIO; model and ground-truth labels follow the original visualization.

G.4 SPATIAL GENERALIZATION ANALYSIS

To evaluate the generalization capability of UniGIO under unseen spatial distributions, we conduct cross-region generation experiments in Africa and South America. Unlike conventional missing-value

Table G.1: Spatial generalization results in Africa.

Setting	Temperature		Dewpoint		Wind Rate		Wind Direc.		Sea-Level Pressure	
	MAE↓/99.5↑	MSE↓/90↑	MAE↓/99.5↑	MSE↓/90↑	MAE↓/99.5↑	MSE↓/90↑	MAE↓/99.5↑	MSE(10^3)↓/90↑	MAE↓/99.5↑	MSE↓/90↑
Global training	1.88/0.44	8.03/0.75	1.87/0.34	8.53/0.70	1.52/0.16	5.13/0.34	75.3/–	10.4/–	1.43/0.72	5.68/0.84
All stations	2.22/0.29	11.4/0.64	2.53/0.16	14.3/0.51	1.62/0.12	5.77/0.21	84.9/–	12.2/–	1.72/0.34	7.86/0.65
90% stations	2.07/0.27	9.68/0.68	2.43/0.17	13.06/0.52	1.53/0.08	4.78/0.24	82.1/–	11.43/–	1.51/0.40	5.37/0.69
60% stations	1.84/0.30	8.00/0.71	2.10/0.17	10.20/0.57	1.36/0.10	3.78/0.25	81.0/–	11.1/–	1.17/0.48	3.67/0.76
30% stations	1.70/0.33	7.17/0.71	2.09/0.19	10.0/0.57	1.35/0.16	3.05/0.27	79.9/–	10.7/–	1.16/0.53	3.38/0.77

Table G.2: Spatial generalization results in South America.

Setting	Temperature		Dewpoint		Wind Rate		Wind Direc.		Sea-Level Pressure	
	MAE↓/99.5↑	MSE↓/90↑	MAE↓/99.5↑	MSE↓/90↑	MAE↓/99.5↑	MSE↓/90↑	MAE↓/99.5↑	MSE(10^3)↓/90↑	MAE↓/99.5↑	MSE↓/90↑
Global training	1.88/0.44	8.03/0.75	1.87/0.34	8.53/0.70	1.52/0.16	5.13/0.34	75.3/–	10.4/–	1.43/0.72	5.68/0.84
All stations	3.14/0.14	20.5/0.46	2.22/0.09	11.9/0.46	1.81/0.08	6.91/0.20	80.9/–	11.7/–	3.26/0.15	25.3/0.46
90% stations	2.28/0.22	15.9/0.60	1.92/0.11	11.9/0.53	1.83/0.08	7.49/0.25	74.0/–	10.2/–	2.46/0.32	17.7/0.58
60% stations	2.06/0.32	8.77/0.64	1.81/0.11	8.20/0.54	1.68/0.14	6.11/0.27	73.4/–	10.1/–	1.72/0.39	7.63/0.67
30% stations	2.04/0.36	8.20/0.70	2.05/0.29	9.67/0.61	1.65/0.13	5.76/0.28	72.8/–	10.5/–	2.11/0.51	10.6/0.74

reconstruction, these settings require the model to infer complete station sequences in regions with different spatial characteristics from the training distribution. We further vary the fraction of available regional stations (100%, 90%, 60%, and 30%) to examine robustness under limited spatial support.

Table G.1 and Table G.2 report the results in Africa and South America, respectively. The results evaluate both conventional reconstruction errors (MAE/MSE) and extreme-event preservation measured by high-quantile statistics (99.5th/90.0th percentiles).

H Detailed Ablation Results

H.1 Model Capacity and Sampling Strength

The default model uses 16 experts and 3.4M parameters. UniGIO-M has 0.9M parameters, reducing capacity by 73.5%, while the 32-expert variant has 4.5M parameters. Tables H.1 and H.2 separate the effects on accuracy, extreme-event capture, and fidelity. UniGIO-M retains comparable fidelity with approximately 5% average performance loss. Increasing sampling strength shifts the accuracy/extreme-event trade-off; zero sampling retains the trained CVAE and therefore is not the deterministic no-CVAE ablation.

H.2 Component Contributions

The labels w/o OM-Mem, OM-Field, LR, ATM-Geo, ATM-MoE, and EA denote removal of memory tokens, field complementarity, local refinement, geographic routing, expert adaptation, and event alignment, respectively. The tables preserve the original settings and values. These component comparisons support the main-text interpretation of complementary observation modeling, temporal adaptation, and local reconstruction; they should not be read as proof that learned representations are fully independent.

Table H.1: Ablation Study on Accuracy and F1-SEDI

Baselines	Temperature		Dewpoint		Wind Rate		Wind Direc.		Sea-Level Pressure	
	MAE↓/99.5th↑	MSE↓/90.0th↑	MAE↓/99.5th↑	MSE↓/90.0th↑	MAE↓/99.5th↑	MSE↓/90.0th↑	MAE↓/99.5th↑	MSE↓ ($\times 10^3$)/90.0th↑	MAE↓/99.5th↑	MSE↓/90.0th↑
UniGIO 3.4M	1.45 / 0.55	5.35 / 0.81	1.50 / 0.46	5.99 / 0.75	1.27 / 0.22	3.63 / 0.39	64.6 /	8.45 /	1.02 / 0.79	3.67 / 0.88
UniGIO-M 0.9M	1.47 / 0.52	5.89 / 0.76	1.55 / 0.45	6.16 / 0.75	1.33 / 0.19	3.74 / 0.37	66.0 /	8.59 /	1.09 / 0.77	3.82 / 0.86
expert 32 4.5M	1.41 / 0.59	4.71 / 0.81	1.45 / 0.49	5.46 / 0.76	1.27 / 0.20	3.49 / 0.41	65.1 /	8.32 /	0.84 / 0.83	2.94 / 0.91
sample 0	1.40 / 0.45	4.14 / 0.66	1.41 / 0.37	5.24 / 0.58	1.15 / 0.10	3.02 / 0.27	56.4 /	6.05 /	0.73 / 0.76	1.97 / 0.77
sample 1.5	1.84 / 0.55	7.54 / 0.81	1.83 / 0.49	8.20 / 0.76	1.77 / 0.27	4.91 / 0.49	86.4 /	13.40 /	1.29 / 0.81	4.59 / 0.89
w/o OM-Mem	1.46 / 0.56	5.39 / 0.81	1.54 / 0.47	6.18 / 0.75	1.29 / 0.22	3.74 / 0.39	65.1 /	8.49 /	1.05 / 0.79	3.69 / 0.87
w/o OM-Field	1.64 / 0.53	5.9 / 0.79	1.62 / 0.44	6.63 / 0.73	1.33 / 0.20	3.82 / 0.39	66.9 /	8.75 /	1.17 / 0.77	3.97 / 0.87
w/o LR	1.69 / 0.54	6.11 / 0.80	1.67 / 0.45	6.75 / 0.74	1.39 / 0.20	3.95 / 0.39	67.4 /	8.96 /	1.21 / 0.78	4.21 / 0.88
w/o ATM-Geo	1.71 / 0.53	6.53 / 0.78	1.73 / 0.44	6.97 / 0.74	1.48 / 0.19	4.07 / 0.38	67.2 /	8.95 /	1.28 / 0.77	4.26 / 0.87
w/o ATM-MoE	1.92 / 0.52	7.27 / 0.77	1.90 / 0.41	7.46 / 0.74	1.56 / 0.17	4.38 / 0.35	75.6 /	9.85 /	1.37 / 0.75	4.97 / 0.87
w/o EA	2.02 / 0.50	7.86 / 0.75	2.03 / 0.41	7.88 / 0.72	1.60 / 0.16	4.49 / 0.34	78.7 /	10.3 /	1.52 / 0.73	5.36 / 0.86

Table H.2: Ablation Study on Fidelity

Baselines	Temperature		Dewpoint		Wind Rate		Wind Direc.		Sea-Level Pressure		CroS↓
	FID↓	STE↓	FID↓	STE↓	FID↓	STE↓	FID↓	STE↓ ($\times 10^3$)	FID↓	STE↓	
UniGIO 3.4M	0.17	2.48	0.25	1.45	0.28	1.07	0.60	2.53	0.11	0.79	1.46
UniGIO-M 0.9M	0.19	2.47	0.29	1.46	0.30	1.09	0.64	2.63	0.05	0.71	1.49
expert 32 4.5M	0.14	2.45	0.21	1.43	0.23	0.97	0.55	2.46	0.02	0.68	1.43
sample 0	0.29	2.77	0.41	1.48	0.54	1.10	0.94	2.83	0.19	0.72	1.54
sample 1.5	0.27	3.09	0.45	1.87	0.66	1.57	1.25	4.11	0.26	1.15	1.56
w/o OM-Mem	0.16	2.53	0.28	1.45	0.25	1.11	0.69	2.54	0.11	0.79	1.47
w/o OM-Field	0.24	2.58	0.45	1.51	0.39	1.24	0.97	2.63	0.19	0.95	1.49
w/o LR	0.30	2.99	0.54	1.68	0.54	1.33	1.21	2.81	0.27	1.14	1.52
w/o ATM-Geo	0.31	3.13	0.57	1.74	0.65	1.35	1.26	2.80	0.29	1.17	1.52
w/o ATM-MoE	0.39	3.21	0.68	1.91	0.79	1.34	1.45	3.02	0.36	1.24	1.54
w/o EA	0.43	3.21	0.75	2.17	0.85	1.32	1.42	3.08	0.39	1.28	1.54