WEATHER TEXT AS OBSERVATION FOR METEOROLOGICAL FIELD GENERATION VIA A UNIFIED MULTIMODAL MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Natural-language weather descriptions have a long history as meteorological records but have rarely been exploited as an information source for atmospheric state inference, particularly for reconstructing meteorological fields. Directly conditioning field generators on weather text, however, provides weak constraints because descriptions are ambiguous and are sparsely aligned with gridded meteorological observation. We address this problem by converting weather descriptions into quantitative meteorological pseudo-observations before field generation. To support this formulation while providing dense supervision, we construct ***MeteoText***, a large-scale dataset that aligns daily city-level weather descriptions with meteorological variables and fields, and define five tasks spanning weather-text understanding and field generation. We further propose ***MeteoOmni***, a unified multimodal model with a ***Value Probe*** for text-to-value mapping and ***Text-Derived Observation Guidance*** for injecting the resulting psedo-observations as local constrains during generation inference. Experiments show that MeteoOmni consistently outperforms strong baselines across all five tasks and generalizes zero-shot to real-world meteorological warnings. More importantly, we demonstrate that weather texts can serve as pseudo-observations to improve local generation, with their benefits extending to the full meteorological field.

1 INTRODUCTION

Meteorological observations are essential for atmospheric state estimation and downstream applications such as forecasting and disaster warning (Kalnay, 2003; Gong et al., 2026b; Chen et al., 2023). However, current structured observations have limited historical coverage, as modern observing networks were only well established from the past century to the past few decades (Brázdil et al., 2005). Textual records offer a complementary source: as one of the earliest ways humans recorded weather, they can date back thousands of years and preserve rich weather information from times and regions with sparse instrumental measurements (Burgdorf, 2022; Burgdorf et al., 2023).

Although textual records offer this potential, prior work mainly focused on generating weather text from meteorological data, including weather reports (Ma et al., 2024; He et al., 2026; Thulke et al., 2024) and warning messages (Tang et al., 2025; Li et al., 2025; Jiang et al., 2025). By contrast, extracting observational information from weather descriptions remains largely unexplored, despite it could provide local constraints for meteorological field generation and improve consistency with local weather conditions. A straightforward way is to incorporate weather text into the generation process as a conditioning signal. However, the influence of such textual conditioning is often weak, which can be attributed to two aspects: (1) *Semantic Ambiguity*. Weather descriptions often express meteorological states in vague and qualitative language, providing only coarse information for constraining meteorological fields (Nash et al., 2020); and (2) *Sparse Text–Field Alignment*. Existing datasets are mostly built around extreme weather events, which limits text to provide dense supervision for meteorological fields. Therefore, ***effectively extracting observational information from weather text to improve meteorological field generation*** remains a key challenge.

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.

Table 1: Comparison between MeteoText and representative weather-text datasets.

Dataset	Dataset Characteristics				Supported Tasks	
	Spatial	Temporal	Construction Unit	Variables Source	Understanding	Generation
WeatherQA (Ma et al., 2024)	CONUS	2014–2020	Storm Event	Reanalysis, Observation	✓	✗
CLLMate (Li et al., 2025)	Global	2015–2023	News Event	Reanalysis	✓	✗
RadarQA (He et al., 2026)	CONUS	2017–2019	Storm Event	Remote Sensing	✓	✗
EWE (Jiang et al., 2025)	Global	2014–2024	Extreme Event	Reanalysis	✓	✗
MP-Bench (Tang et al., 2025)	China, CONUS	2023–2024	Warning Event	Reanalysis	✓	✗
MeteoText	**Global**	**1911–2018**	**Daily City Weather Record**	**Reanalysis, Simulation, Observation**	✓	✓

We address this challenge by using meteorological variables as a bridge between weather text and meteorological fields, where textual information can first be grounded into quantitative observations of local weather states and then used to guide field generation. To support this formulation while providing dense supervision, we construct ***MeteoText***, a spatiotemporally indexed dataset that jointly aligns daily city-level weather descriptions, corresponding meteorological values, and meteorological fields, using the European Centre for Medium-Range Weather Forecasts Reanalysis version 5 (ERA5) (Hersbach et al., 2020) and Phase 6 of the Coupled Model Intercomparison Project (CMIP6) (Eyring et al., 2016). This text-value-field alignment supports five tasks across two stages: three *understanding* tasks establish and evaluate the bidirectional correspondence between weather descriptions and meteorological variables, while two *generation* tasks assess whether observational information derived from weather text can improve meteorological field generation. We also evaluate Out-Of-Distribution (OOD) generalization on real-world weather text issued by the China Meteorological Administration (CMA) from 2021 to 2023.

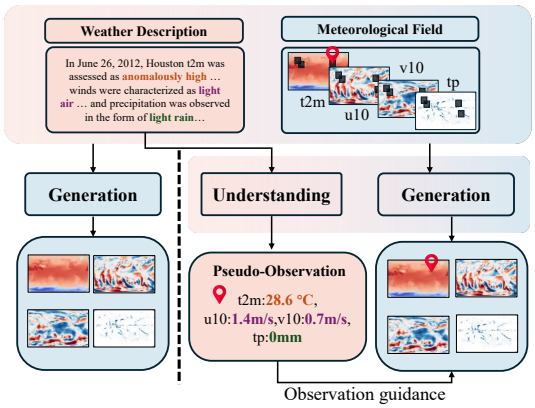

Figure 1: **Comparison** between direct text-conditioned field generation (left) and our approach that bridges weather text and meteorological fields through quantitative pseudo-observations (right).

Building on this task formulation, we propose ***MeteoOmni***, a unified multimodal model that integrates city weather question answering, text quantification, and text-conditioned field generation within a shared backbone. In MeteoOmni, the understanding branch first quantifies weather descriptions into meteorological variables and treats these as local pseudo-observations. We introduce ***Value Probe*** to improve the accuracy of these text-derived pseudo-observations through auxiliary regression supervision. ***Text-Derived Observation Guidance*** then injects the resulting pseudo-observations as explicit numerical constraints during field generation inference.

Our main contributions can be summarized as:

- We introduce the first paradigm for using weather text to improve meteorological field generation by grounding weather descriptions into quantitative local observations and using them as explicit constraints during generation.
- We construct MeteoText, a large-scale and spatiotemporally indexed dataset that aligns text, meteorological values, and fields, with approximately 100K samples spanning more than a century. We further define five benchmark tasks and evaluate zero-shot generalization on real-world CMA warnings in two downstream tasks.
- We propose MeteoOmni, a unified multimodal model for weather-text understanding and meteorological field generation. Value Probe improves text-to-value grounding, while Text-Derived Observation Guidance uses the resulting pseudo-observations as explicit constraints for field generation.

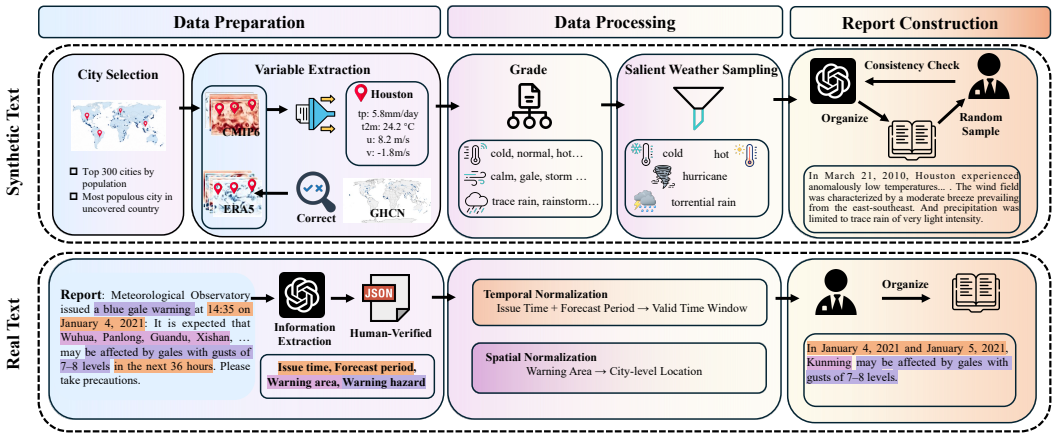

Figure 2: Overview of the MeteoText construction pipeline.

2 METEOTEXT DATASET

MeteoText is a large-scale weather-text dataset that aligns city-level weather descriptions with corresponding meteorological values and meteorological fields. It is designed to study how weather descriptions can serve as observational constraints for meteorological field generation. As shown in Table 1, MeteoText has three main advantages over existing weather-text datasets: (1) ***Daily city weather construction***, which organizes weather records at a daily resolution for each city, allowing each meteorological field to be associated with multiple city-level weather descriptions and providing denser textual supervision for field generation; (2) ***Longer temporal range***, extending the coverage of weather descriptions to historical periods not covered by existing datasets; and (3) ***Broader spatial coverage***, spanning more regions and climate types.

2.1 DATA SOURCE

MeteoText aligns city-level weather descriptions with meteorological values and fields from ERA5 and CMIP6. We use daily 2-m temperature (t2m), 10-m zonal and meridional winds (u10 and v10), and 24-h accumulated precipitation (tp) from ERA5 and CNRM-CM6-1 (Hersbach et al., 2020; Eyring et al., 2016). ERA5 precipitation is further corrected using GHCN station observations (Peterson et al., 1998) and details are provided in Appendix A.1.2. Weather descriptions come from synthetic and real-world sources. Synthetic descriptions are generated from city-level meteorological states derived from ERA5 and CMIP6, with GeoNames providing city metadata for selection and spatial alignment. Real-world descriptions are collected from CMA disaster warnings, which provide issue time, warning type, and affected region.

2.2 DATASET CONSTRUCTION PIPELINE

As illustrated in Fig. 2, MeteoText contains separate pipelines for synthetic and real weather descriptions. ***For synthetic text***, we first select the 300 most populous cities worldwide and additionally include the most populous city from each uncovered country. For each city-date record, ERA5 or CMIP6 variables are converted into discrete meteorological attributes: wind speed and direction are derived from zonal and meridional wind components, with wind speed graded by the Beaufort scale (Wheeler & Wilkinson, 2004), while precipitation and temperature follow national standards[1][2]. We then sample records with salient weather conditions and use GPT-5.1 to verbalize the city, date, meteorological grades, and wind direction into weather descriptions, without access to the continuous target values. Manual inspection of 80% of the generated reports in the test set confirms that all

[1] https://openstd.samr.gov.cn/bzgk/std/newGbInfo?hcno=B4A00E4ABCF80F8C6A048C1D0121A97D

[2] https://openstd.samr.gov.cn/bzgk/std/newGbInfo?hcno=9CEE599D2B75DC1F3F35D056B131B07D

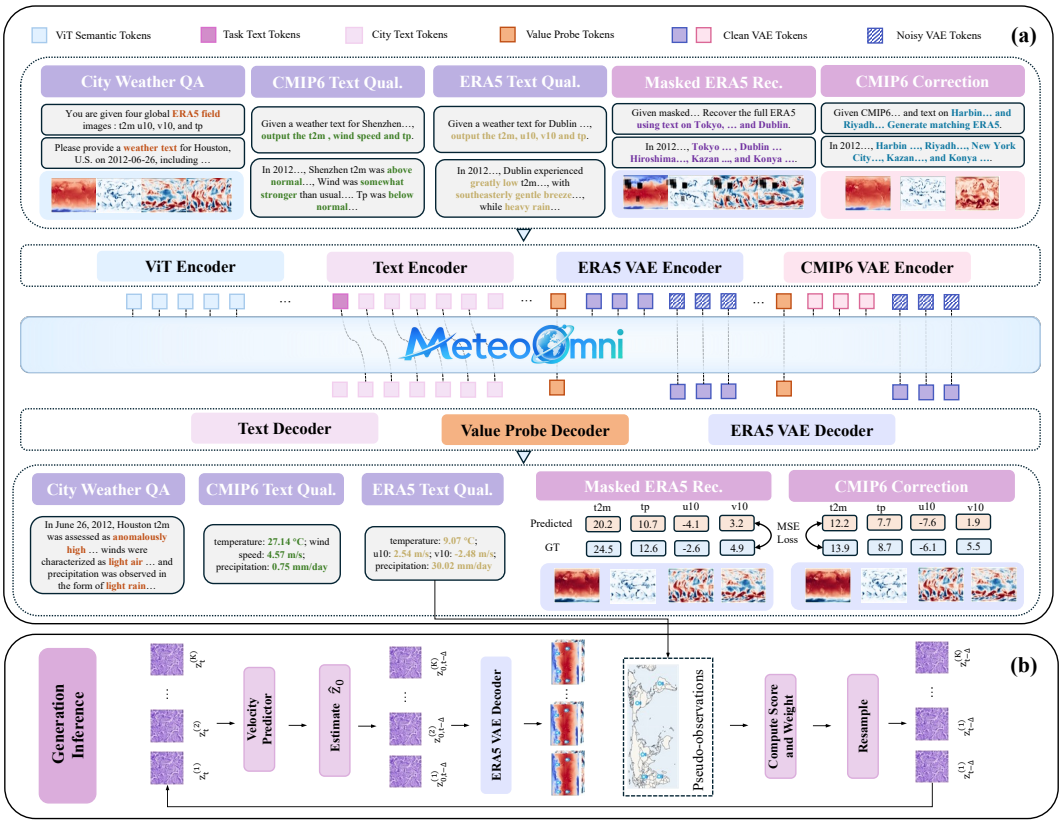

Figure 3: Overview of MeteoOmni. (a) Architecture, task paradigm, and training framework. (b) Generation inference with Text-Derived Observation Guidance using pseudo-observations.

are consistent with their conditioning attributes. ***For real text***, we use GPT-5.1 to extract the issue time, forecast period, warning area, and weather hazard from CMA warnings. We manually validate approximately 80% of the records, achieving an extraction accuracy of 96.28% across these four fields, and correct any identified errors, including incomplete warning areas and incorrect forecast dates. Further details are provided in Appendix A.1.1-A.1.3.

2.3 Dataset Statistics

The synthetic portion of MeteoText contains approximately 100K city-date samples, covering 314 cities across 131 countries. Temporally, ERA5 spans the period from 1979 to 2018, whereas CMIP6 covers a broader range from 1911 to 2014. Geographically, the samples are distributed across diverse regions worldwide. Salient weather conditions, including precipitation, high temperature, low temperature, and wind, are approximately balanced, with 12K to 13K cases per category in both ERA5 and CMIP6. For real-world OOD evaluation, MeteoText further includes 739 CMA warning samples collected from 22 cities in China between 2021 and 2023. Notably, these real-world samples are temporally disjoint from the synthetic data. Additional statistics are provided in Appendix A.1.4.

3 Methods

3.1 Task Formulation

Using MeteoText, we formulate five tasks over aligned triplets (s, v, X), where s, v, and X denote the city-level weather description, corresponding meteorological values and meteorological field, respectively. The five tasks cover three understanding settings, i.e., *(i) City Weather Ques-*

tion Answering (QA), *(ii) CMIP6 Text Quantification*, and *(iii) ERA5 Text Quantification*, and two generation settings, i.e., *(iv) Masked ERA5 Reconstruction* and *(v) CMIP6 Correction*.

Specifically, the understanding tasks evaluate the bidirectional correspondence between text and observation. *City Weather QA* takes an ERA5 field $X^E \in \mathbb{R}^{4\times H\times W}$ and city metadata as input and generates a weather description s^E for the queried city. *CMIP6 and ERA5 Text Quantification* convert weather descriptions s^C and s^E into structured textual representations of numerical meteorological variables. The generation tasks further assess whether weather descriptions can improve meteorological field generation. *Masked ERA5 Reconstruction* recovers the complete X^E from a masked field $\tilde{X}^E$ and weather descriptions of the masked cities, while *CMIP6 Correction* generates the X^E from a CMIP6 field $X^C \in \mathbb{R}^{3\times H\times W}$ and weather descriptions of partial cities s^E.

Formally, all five tasks can be unified as prompt-conditioned mappings over the triplets (s, v, X):

$$\hat{y}_t = F_\theta(p_t, x_t), \quad t \in \{1, \dots, 5\}, \tag{1}$$

where p_t is the task prompt, x_t is the task-specific input constructed from (s, v, X), and y_t is the corresponding target. For convenience, we denote these five tasks as T1–T5.

Architecture. As illustrated in Fig. 3(a), MeteoOmni unifies weather understanding and meteorological field generation within a shared multimodal backbone. Task prompts, city metadata, and weather descriptions are encoded by a shared *text encoder*, while meteorological fields are processed by task-specific vision modules. For understanding, the ERA5 field in *City Weather QA* is encoded by the *vit encoder* and fused with the textual tokens through shared self-attention layers, after which the *text decoder* generates the city-level weather description. The same shared backbone also supports text quantification, which maps weather descriptions to meteorological values. For generation, the masked ERA5 or CMIP6 background is encoded into latent tokens by the corresponding *VAE encoder* following DAWP (Gong et al., 2026a). These latent tokens are fused with city-level weather description tokens through the shared self-attention layers, and the *ERA5 VAE decoder* reconstructs the target field. Weather descriptions therefore provide implicit textual conditions for generation, while their quantified meteorological values further serve as pseudo-observations for explicit numerical guidance in generation inference, as shown in Fig. 3(b).

Value Probe. Although MeteoOmni can generate structured numerical text from weather descriptions in understanding tasks, token-level supervision provides weak constraints on continuous physical variables. This limits the reliability of *ERA5 Text Quantification* (T3) and the pseudo-observation guidance used in *Masked ERA5 Reconstruction* (T4) and *CMIP6 Correction* (T5). We therefore introduce *Value Probe*, which applies auxiliary numerical loss to the city-level weather descriptions on the generation side during training. Specifically, we append a learnable probe token e_{probe} to each city weather description. After this token is jointly processed with the weather-description tokens and latent field tokens by the shared self-attention layers, we take its hidden state h^i_{probe} and use a two-layer MLP head to predict the ERA5 observations of the corresponding city as follows:

$$\tilde{\mathbf{v}}_i = \text{MLP}(h^i_{\text{probe}}), \tag{2}$$

where $\tilde{\mathbf{v}}_i = [\tilde{t}^{2m}_i, \tilde{u}^{10}_i, \tilde{v}^{10}_i, \tilde{tp}_i]$ denotes the ERA5 values predicted by the probe for the i-th city.

Text-Derived Observation Guidance. Direct text conditioning provides weak and implicit constraints on field generation. We therefore use T3 to quantify city-level descriptions into pseudo-observations and inject them into T4 and T5 as explicit numerical constraints during inference, as shown in Fig. 3(b). Specifically, for the i-th city in prompt, we firstly obtain a pseudo-observation $\hat{\mathbf{v}}_i = [\hat{t}^{2m}_i, \hat{u}^{10}_i, \hat{v}^{10}_i, \hat{tp}_i]$. Combined with its grid location (r_i, c_i), they form a sparse pseudo-observation set $\mathcal{O} = \{(r_i, c_i, \hat{\mathbf{v}}_i)\}_{i=1}^N$. To inject $\mathcal{O}$ into generation, we maintain K parallel latent particles $\{z^{(k)}_t\}_{k=1}^K$ during inference. Every Δ sampling steps, we estimate the clean latent of each particle as $\hat{z}^{(k)}_0 = z^{(k)}_t - t\cdot v_\theta(z^{(k)}_t, t, p_t, x_t)$. Then, we decode each $\hat{z}^{(k)}_0$ with the ERA5 VAE decoder to reconstruct field $\hat{\mathbf{y}}^{(k)}$. Each particle is scored by its consistency with the pseudo-observations:

$$\log w_k = -\sum_{(r_i, c_i, \hat{\mathbf{v}}_i)\in\mathcal{O}} \frac{\left\|\hat{\mathbf{y}}^{(k)}(:, r_i, c_i) - \hat{\mathbf{v}}_i\right\|_2^2}{2\sigma^2}, \tag{3}$$

where σ is the standard deviation of pseudo-observation errors in the normalized ERA5 space. Particles are then resampled according to $\{w_k\}_{k=1}^K$, favoring trajectories that better satisfy the local

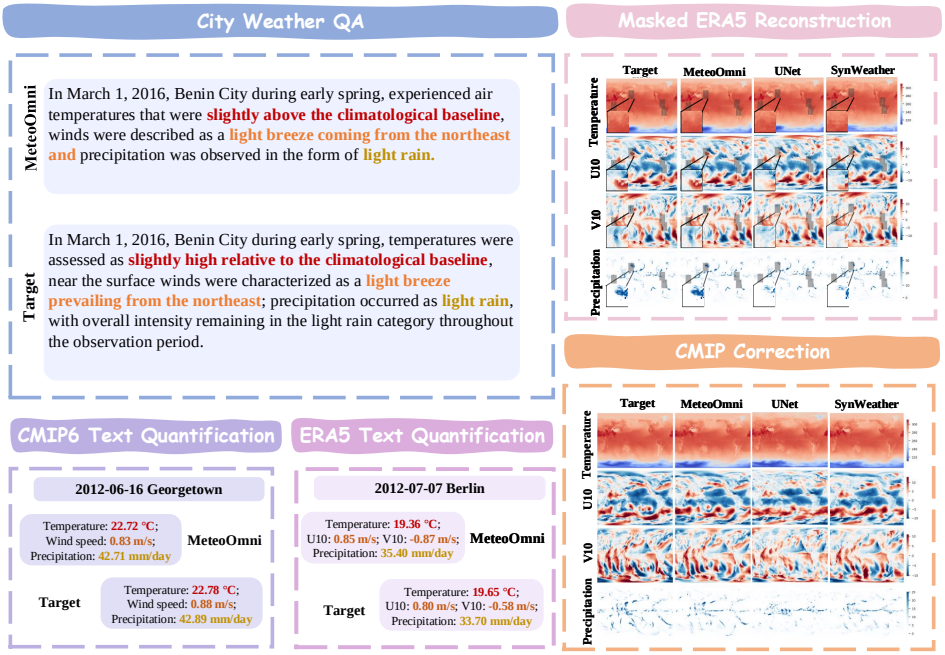

Figure 4: Qualitative results across the five tasks in MeteoText.

constraints derived from city-level descriptions. Thus, without updating model parameters, MeteoOmni steers generation using text-derived pseudo-observations.

Training Objectives. We initialize MeteoOmni from the pretrained BAGEL-7B-MoT (Deng et al., 2025) and jointly fine-tune it on T1-T5 under the unified task formulation. For each task t, MeteoOmni learns the target y_t conditioned on (p_t, x_t). We denote T1-T3 as set $\mathcal{T}_{\text{und}}$ and T4-T5 as set $\mathcal{T}_{\text{gen}}$, with λ_t balancing the contribution of each task. For understanding tasks, $y_t = (y_{t,1}, \ldots, y_{t,n_t})$ is the target text sequence, and we optimize the token-level cross-entropy loss, which is defined as:

$$\mathcal{L}_{\text{und}} = \sum_{t\in\mathcal{T}_{\text{und}}} \lambda_t \left(-\sum_{k=1}^{n_t} \log p_\psi \left(y_{t,k} \mid y_{t,<k}, p_t, x_t\right)\right). \tag{4}$$

For generation tasks, we optimize the flow matching together with the value probe objective:

$$\mathcal{L}_{\text{gen}} = \sum_{t\in\mathcal{T}_{\text{gen}}} \lambda_t \left(\frac{1}{|\Omega_t|}\sum_{j\in\Omega_t} \|\hat{y}_{t,j} - y_{t,j}\|_2^2 + \beta \frac{1}{N_t}\sum_{i=1}^{N_t} \|\tilde{\mathbf{v}}_i - \mathbf{v}_i^*\|_2^2\right), \tag{5}$$

where Ω_t indexes latent positions, and N_t counts cities in generation task t. y_t denotes the flow matching target velocity, while $\hat{y}_t$ denotes the velocity predicted from (p_t, x_t). $\tilde{\mathbf{v}}_i$ is the value probe prediction and $\mathbf{v}_i^*$ is the corresponding ERA5 observation. β balances the two generation objectives. The overall training objective combines the understanding and generation losses as follows:

$$\mathcal{L} = \mathcal{L}_{\text{und}} + \mathcal{L}_{\text{gen}}. \tag{6}$$

4 Experiments

4.1 Implementation and Evaluation

We train MeteoOmni for 200K steps on four NVIDIA H200 GPUs using AdamW with a learning rate of 2×10^{-5}. The learning rate is linearly warmed up for the first 2K steps and remains constant thereafter. For multi-task training, we set the task weights λ_t to 1 for T1-T3 and 4 for T4-T5, with the Value Probe loss weight β set to 0.1. Further implementation details are provided in Appendix A.2.

We evaluate models on both weather understanding and field generation tasks. For *City Weather QA*, ROUGE-L (Lin, 2004), METEOR (Banerjee & Lavie, 2005), and BERTScore (Zhang et al., 2019) measure the quality of generated descriptions, while Meteorological Grade Accuracy (MGA) evaluates their consistency with the discrete grades of meteorological variables. For *Text Quantification*,

Table 2: Quantitative results on weather understanding and generation tasks. Best results are in bold, and second-best results are underlined. ↑ indicates higher is better, and ↓ indicates lower is better.

Understanding

Method	City Weather QA							CMIP6 Text Quantification				ERA5 Text Quantification				
	R-L ↑	MET ↑	B.S. ↑	M-T ↑	M-W ↑	M-P ↑	M-O ↑	A-T ↑	A-W ↑	A-P ↑	A-O ↑	A-T ↑	A-U ↑	A-V ↑	A-P ↑	A-O ↑
GPT-5.5	0.297	0.273	0.879	0.536	0.443	0.063	0.364	0.741	0.554	0.567	0.621	0.809	0.020	0.105	0.328	0.315
GPT-5.6-Sol	0.164	0.147	0.862	0.599	0.496	0.086	0.399	0.787	0.630	0.570	0.662	0.828	0.040	0.020	0.327	0.303
GPT-6-Astra	0.159	0.142	0.854	**0.612**	0.503	0.097	0.404	0.733	0.630	0.577	0.647	0.858	0.020	0.000	0.300	0.295
Claude-Sonnet-4.6	0.209	0.250	0.853	0.441	0.485	0.097	0.341	0.714	0.535	0.523	0.590	0.831	0.251	0.056	0.317	0.364
Kimi-K3	0.133	0.136	0.841	0.482	0.413	0.046	0.314	0.758	0.612	0.526	0.632	0.822	0.025	0.041	**0.563**	0.363
Qwen-3.8	0.135	0.186	0.836	0.396	0.270	0.074	0.247	0.749	0.581	0.503	0.611	0.815	0.004	0.007	0.509	0.334
GLM-5.3	0.161	0.233	0.863	0.511	**0.586**	0.000	0.366	0.417	0.331	0.502	0.416	0.741	0.066	0.054	0.540	0.350
MeteoOmni	**0.531**	**0.532**	**0.926**	0.604	0.536	**0.480**	**0.540**	**0.877**	**0.892**	**0.734**	**0.834**	**0.872**	**0.701**	**0.702**	0.379	**0.664**

Understanding notes. R-L = ROUGE-L F1; MET = METEOR; B.S. = BERTScore F1; MGA = Meteorological Grade Accuracy; APE = clamped APE score. For City Weather QA, M-T, M-W, and M-P denote MGA for temperature, wind, and precipitation, respectively, while M-O denotes their mean. For text quantification, A-T, A-W, A-P, A-U, and A-V denote clamped APE scores for temperature, wind, precipitation, U10, and V10, respectively, while A-O denotes their mean.

Generation

Method	Masked ERA5 Reconstruction				CMIP6 Correction							
	Ms-t2m ↓	Ms-u10 ↓	Ms-v10 ↓	Ms-tp ↓	Glo-t2m ↓	Glo-u10 ↓	Glo-v10 ↓	Glo-tp ↓	City-t2m ↓	City-u10 ↓	City-v10 ↓	City-tp ↓
SynWeather (Xu et al., 2026)	3.527	2.908	2.790	12.920	4.224	4.837	4.810	6.230	4.112	4.138	3.558	13.954
UNet (Ronneberger et al., 2015)	2.639	1.559	1.534	10.811	4.443	4.694	4.633	6.306	4.103	3.563	3.275	13.119
MeteoOmni-no text	1.304	1.182	1.171	8.827	3.958	4.707	4.695	6.040	3.324	3.560	3.398	13.074
MeteoOmni	**1.273**	**1.142**	**1.135**	**8.641**	**3.830**	**4.645**	**4.626**	**6.001**	**2.675**	**2.640**	**2.534**	**10.691**

Generation notes. Ms = Masked; Glo = Global. t2m, u10, v10, and tp denote 2-meter temperature, 10-meter zonal wind, 10-meter meridional wind, and total precipitation, respectively.

Table 3: Zero-shot evaluation on real text for ERA5 Text Quantification.

Model	A-T ↑	A-P ↑	A-U ↑	A-V ↑
GPT-6-Astra	0.7446	0.1392	0.0440	0.0573
GPT-5.6-sol	0.5954	0.1224	0.0333	0.0334
MeteoOmni	**0.8284**	**0.3276**	**0.1093**	**0.1040**

Notes: A-T, A-P, A-U, and A-V denote clamped APE scores for temperature, precipitation, zonal wind, and meridional wind.

Table 4: Zero-shot evaluation on real text on Masked ERA5 Reconstruction.

Setting	Ms-t2m ↓	Ms-tp ↓
w/o real text	1.759	7.737
with real text	**1.747**	**7.693**

Notes: Ms-t2m and Ms-tp denote masked-region RMSE of 2-meter temperature and total precipitation, respectively.

we use the clamped APE score to evaluate the recovery of numerical meteorological variables from text Bell et al. (2025). Field generation is evaluated using RMSE between the generated and target fields, computed over masked regions for *Masked ERA5 Reconstruction* and over both global and city-centered regions for *CMIP6 Correction*. Full metric definitions are provided in Appendix A.3.

4.2 Experiment Results

MeteoOmni achieves strong capabilities on weather understanding tasks. Table 2 shows that MeteoOmni achieves the best text similarity scores on *City Weather QA* and improve overall MGA from 0.404 to 0.540, demonstrating stronger city-level meteorological understanding. More importantly, MeteoOmni shows consistent gains across meteorological variables in text quantification, achieving the best overall scores of 0.834 on CMIP6 and 0.664 on ERA5. Together, these results establish a reliable correspondence between weather language and meteorological states, enabling weather descriptions to be converted into quantitative pseudo-observations that provide useful constraints for meteorological field generation.

MeteoOmni also shows superior performance on field generation tasks. As shown in Table 2 and Fig. 4, MeteoOmni achieves the lowest RMSE across all variables in both *Masked ERA5 Reconstruction* and *CMIP6 Correction*. The improvement is especially pronounced at the city level in CMIP6 correction, where t2m, u10, and v10 RMSE are reduced by 34.8%, 25.9%, and 22.6% over the strongest baselines, respectively.

Weather descriptions provide effective observational constraints for meteorological field generation. Compared with MeteoOmni-no text, the MeteoOmni further reduces RMSE on nearly all

Table 5: Training with only understanding tasks, only generation tasks, and joint understanding-generation tasks. T1 is evaluated by R-L, B.S., and Overall MGA; T2/T3 by APE; and T4/T5 by RMSE. T4 is evaluated over masked regions, while T5 is evaluated over 5×5 regions centered at cities. R-L = ROUGE-L F1; B.S. = BERTScore F1.

Setting	Understanding							Generation			
	T1			T2/T3				T4/T5			
	R-L ↑	B.S ↑	Overall-MGA ↑	APE-t2m ↑	APE-wind ↑	APE-tp ↑	APE-all ↑	RMSE-t2m ↓	RMSE-u10 ↓	RMSE-v10 ↓	RMSE-tp ↓
Und-Only	0.529	0.925	0.497	0.846 / 0.861	0.788 / 0.643	0.693 / 0.355	0.776 / 0.620	-	-	-	-
Gen-Only	-	-	-	-	-	-	-	1.449 / 3.888	1.268 / 3.282	1.244 / 3.036	9.913 / 13.390
Joint (U+G)	**0.531**	**0.926**	**0.540**	**0.869 / 0.872**	**0.849 / 0.787**	**0.713 / 0.379**	**0.810 / 0.679**	**1.273 / 2.675**	**1.142 / 2.640**	**1.135 / 2.534**	**8.641 / 10.691**

Table 6: Ablation of Value Probe on T3. Entries report variable-wise RMSE.

Value Probe	t2m	u10	v10	tp
✗	2.354	0.648	0.676	10.991
✓	**2.096**	**0.639**	**0.613**	**10.771**

Table 7: Ablation of Value Probe and TDOG on T4 and T5. Entries report variable-wise RMSE as T4 / T5.

Value Probe	TDOG	t2m	u10	v10	tp
✗	✗	1.482 / 3.394	2.022 / 3.341	1.996 / 3.057	9.121 / 12.706
✗	✓	1.469 / 3.117	1.936 / 3.133	1.984 / 3.054	9.092 / 12.005
✓	✗	1.301 / 3.020	1.165 / 3.146	1.156 / 2.931	8.746 / 12.016
✓	✓	**1.273 / 2.675**	**1.142 / 2.640**	**1.135 / 2.534**	**8.641 / 10.691**

metrics in Table 2. In *Masked ERA5 Reconstruction*, the gains from text conditioning are relatively moderate, because the unmasked ERA5 regions already provide reliable spatial background information. As a contrast, in *CMIP6 Correction*, where the CMIP6 background field contains systematic biases, text information brings larger gains at the city level, reducing the RMSE of t2m, u10, v10, and tp by 19.5%, 25.8%, 25.4%, and 18.2%, respectively. By converting weather descriptions into pseudo observations and injecting them into the generation process, MeteoOmni improves the local consistency and propagates to full-field.

Real-world weather text contain observational information, which MeteoOmni effectively extracts and utilizes. To evaluate whether weather descriptions beyond synthetic data can be treated as meteorological observations, we conduct zero-shot experiments on real-world weather data. As shown in Table 3, MeteoOmni extracts quantitative meteorological information from real-world texts and consistently outperforms general-purpose language models. Furthermore, the extracted variables are used as pseudo-observations for *Masked ERA5 Reconstruction*. As shown in Table 4 and Fig 15, real-world weather text improves field reconstruction performance. These results demonstrate that real-world weather texts contain valuable observational information, and MeteoOmni can effectively convert such information into constraints for meteorological field generation.

4.3 Ablation Studies and Analysis

Impact of joint training on understanding and generation. To examine whether understanding and generation tasks benefit each other, we compare fine-tuning with only understanding data, only generation data, and both jointly. As shown in Table 5, joint training improves T1 captioning and T2/T3 numerical consistency over understanding-only training, and reduces T4/T5 RMSE over generation-only training. This complementarity can be attributed to the fact that generation tasks encourage physically consistent field representations that help numerical grounding, while understanding tasks improve the alignment between weather descriptions and meteorological variables, making text conditions and pseudo-observation guidance more reliable for generation.

Effect of value probe on text quantification and generation fields. We conduct ablation studies on the understanding task T3 and the generation tasks T4 and T5 to evaluate the effect of the value probe. As shown in Table 6, adding the value probe consistently reduces RMSE across all meteorological variables in T3. It indicates that explicit numerical constraints on continuous physical variables help the model ground meteorological quantities from text more accurately. This improvement in T3 directly benefits Text-Derived Observation Guidance (TDOG), which uses values predicted from text as guidance for field generation. More accurate text quantification can therefore provide more reliable guidance for T4 and T5. The value probe also improves generation independently of TDOG.

As shown in Table 7, when TDOG is not used, the value probe improves generation performance on T4 and T5 across all variables, suggesting that auxiliary value learning encourages the model to form more physically consistent representations.

Effect of text-derived observation guidance on generation fields. As shown in Table 7, TDOG improves generation in both tasks, especially on T5. The larger gains arise because the CMIP6 field is biased relative to ERA5, whereas the unmasked ERA5 regions in T4 provide a reliable spatial context. By incorporating text-derived local observations, TDOG supplies correction signals and produces more physically consistent fields. The extreme-event case of the Late Season Mid-Atlantic and Northeast Winter Storm in Fig 5 further illustrates this effect. Compared with the raw CMIP6 input and the baseline without TDOG, TDOG recovers a clearer precipitation band, better preserves the local cold anomaly, and more closely matches ERA5 wind gradients and directions.

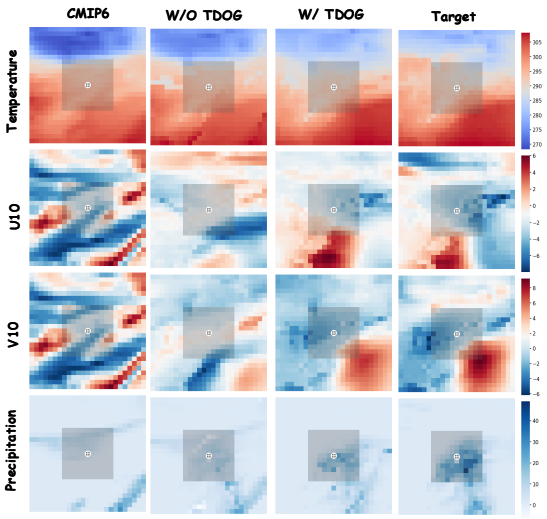

Figure 5: Case study of the Late Season Mid-Atlantic and Northeast Winter Storm, comparing CMIP6 input, baseline without TDOG, TDOG-refined result, and ERA5 target.

5 Related Work

Weather-Text Datasets. Weather-text datasets have recently emerged to connect meteorological data with natural language, but most focus on data-to-text generation. WeatherSyn (Zheng et al., 2026) generates forecast reports from multi-source weather information, while ClimateGPT (Thulke et al., 2024) studies climate-related text synthesis and question answering from scientific resources. Besides, most dataset mainly focus on severe weather warnings and event understanding. For example, WeatherQA (Ma et al., 2024) evaluates multimodal reasoning for severe convective weather. CLLMate (Li et al., 2025) aligns ERA5 with environmental news for forecasting textual weather and climate events, and EWE (Jiang et al., 2025) provides a benchmark for diagnosing high-impact extreme weather. However, their event-driven construction offers sparse textual supervision for ordinary city-date weather states, limiting dense alignment with continuous meteorological fields.

Earth VLMs. Recent Earth VLMs extend vision-language modeling to weather and Earth science by aligning observational data with natural language. RadarQA (He et al., 2026) studies whether MLLMs can assess radar forecast quality through numerical and descriptive evaluations. OmniWeather (Zhou et al., 2026) further unifies radar understanding and generation, using language to describe and evaluate radar observations and generated results. And MP-Bench (Tang et al., 2025) studies severe-weather prediction from meteorological inputs and text. In these works, language primarily serves to understand, describe, or evaluate meteorological data and model outputs. In contrast, we treat weather text as an observational source, grounding it into quantitative pseudo-observations that explicitly constrain meteorological field generation.

6 Conclusion and Discussion

We construct the first dataset aligning weather descriptions, meteorological variables, and meteorological fields (i.e. MeteoText), and propose a unified multimodal model that grounds weather text into quantitative pseudo-observations for field generation (i.e. MeteoOmni). Experiments demonstrate that weather text can serve as a source of meteorological observations, providing local constraints whose benefits extend to the full meteorological field. This formulation is particularly promising for historical climate reconstruction, where documentary records could provide local constraints for reconstructing large-scale fields or correcting biased climate simulations.

AI USE DISCLOSURE

In this work, we used generative AI tools as part of dataset construction and manuscript polishing. Specifically, GPT-5.1 was used to verbalize structured meteorological attributes into synthetic weather descriptions. The model was provided with city, date, discrete meteorological grades, and wind direction, but did not have access to the continuous target values. For real-world CMA warnings, GPT-5.1 was used to extract the issue time, forecast period, warning area, and weather hazard. The extracted information was manually verified and subsequently normalized to valid dates and city-level locations. We additionally used generative AI tools to improve the readability and language quality of the manuscript. All AI-assisted data processing and manuscript editing were reviewed by the authors. The authors take full responsibility for the final content of the paper, including all text, data, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses publicly available meteorological, climate, geographic, and operational weather-warning data for research purposes and does not involve human-subject experiments or personal sensitive information. All data processing, experimental results, and conclusions were reviewed by the authors, who take responsibility for the content of this work.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure the reproducibility of this work by providing sufficient methodological, experimental, and implementation details in the main paper and supplementary materials. Relevant configurations, evaluation procedures, and data processing information are documented to support independent reproduction of the reported results. Code and releasable data resources will be made available upon publication.

REFERENCES

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Aaron Bell, Amit Aides, Amr Helmy, Arbaaz Muslim, Aviad Barzilai, Aviv Slobodkin, Bolous Jaber, David Schottlander, George Leifman, Joydeep Paul, et al. Earth ai: unlocking geospatial insights with foundation models and cross-modal reasoning. *arXiv preprint arXiv:2510.18318*, 2025.

Rudolf Brázdil, Christian Pfister, Heinz Wanner, Hans Von Storch, and Jürg Luterbacher. Historical climatology in europe–the state of the art. *Climatic change*, 70(3):363–430, 2005.

Angela-Maria Burgdorf. A global inventory of quantitative documentary evidence related to climate since the 15th century. *Climate of the Past*, 18(6):1407–1428, 2022.

Angela-Maria Burgdorf, Stefan Brönnimann, George Adamson, Tatsuya Amano, Yasuyuki Aono, David Barriopedro, Teresa Bullón, Chantal Camenisch, Dario Camuffo, Valérie Daux, et al. Docu-clim: A global documentary climate dataset for climate reconstructions. *Scientific data*, 10(1):402, 2023.

Kang Chen, Tao Han, Junchao Gong, Lei Bai, Fenghua Ling, Jing-Jia Luo, Xi Chen, Leiming Ma, Tianning Zhang, Rui Su, et al. Fengwu: Pushing the skillful global medium-range weather forecast beyond 10 days lead. *arXiv preprint arXiv:2304.02948*, 2023.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Veronika Eyring, Sandrine Bony, Gerald A Meehl, Catherine A Senior, Bjorn Stevens, Ronald J Stouffer, and Karl E Taylor. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016.

Junchao Gong, Jingyi Xu, Ben Fei, Fenghua Ling, Wenlong Zhang, Kun Chen, Wanghan Xu, Weidong Yang, Xiaokang Yang, and Lei Bai. Dawp: A framework for global observation forecasting via data assimilation and weather prediction in satellite observation space. *Advances in Neural Information Processing Systems*, 38:148646–148682, 2026a.

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*, 2026b.

Xuming He, Zhiyuan You, Junchao Gong, Couhua Liu, Xiaoyu Yue, Peiqin Zhuang, Wenlong Zhang, and Lei Bai. Radarqa: Multi-modal quality analysis of weather radar forecasts. *Advances in Neural Information Processing Systems*, 38:131737–131767, 2026.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The era5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730):1999–2049, 2020.

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

Zhe Jiang, Jiong Wang, Xiaoyu Yue, Zijie Guo, Wenlong Zhang, Fenghua Ling, Wanli Ouyang, and Lei Bai. Ewe: An agentic framework for extreme weather analysis. *arXiv preprint arXiv:2511.21444*, 2025.

Eugenia Kalnay. *Atmospheric modeling, data assimilation and predictability*. Cambridge university press, 2003.

Haobo Li, Zhaowei Wang, Jiachen Wang, YueYa Wang, Alexis Kai Hon Lau, and Huamin Qu. Cllmate: A multimodal benchmark for weather and climate events forecasting. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 17547–17573, 2025.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Chengqian Ma, Zhanxiang Hua, Alexandra Anderson-Frey, Vikram Iyer, Xin Liu, and Lianhui Qin. Weatherqa: Can multimodal language models reason about severe weather? *arXiv preprint arXiv:2406.11217*, 2024.

David J Nash, George CD Adamson, Linden Ashcroft, Martin Bauch, Chantal Camenisch, Dagomar Degroot, Joelle Gergis, Adrian Jusopović, Thomas Labbé, Kuan-Hui Elaine Lin, et al. Climate indices in historical climate reconstructions: A global state-of-the-art. *Climate of the Past Discussions*, 2020:1–48, 2020.

Thomas C Peterson, Russell Vose, Richard Schmoyer, and Vyachevslav Razuvaëv. Global historical climatology network (ghcn) quality control of monthly temperature data. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 18(11):1169–1179, 1998.

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

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Shuo Tang, Jian Xu, Jiadong Zhang, Yi Chen, Qizhao Jin, Lingdong Shen, Chenglin Liu, and Shiming Xiang. Meteorpred: A meteorological multimodal large model and dataset for severe weather event prediction. *arXiv preprint arXiv:2508.06859*, 2025.

David Thulke, Yingbo Gao, Petrus Pelser, Rein Brune, Richa Jalota, Floris Fok, Michael Ramos, Ian Van Wyk, Abdallah Nasir, Hayden Goldstein, et al. Climategpt: Towards ai synthesizing interdisciplinary research on climate change. *arXiv preprint arXiv:2401.09646*, 2024.

Dennis Wheeler and Clive Wilkinson. From calm to storm: the origins of the beaufort wind scale. *The Mariner's Mirror*, 90(2):187–201, 2004.

Kaiyi Xu, Junchao Gong, Zhiwang Zhou, Zhangrui Li, Yuandong Pu, Yihao Liu, Ben Fei, Fenghua Ling, Wenlong Zhang, and Lei Bai. Synweather: Weather observation data synthesis across multiple regions and variables via a general diffusion transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 1346–1354, 2026.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Zinan Zheng, Yang Liu, Nuo Chen, Juepeng Zheng, Hong Cheng, and Jia Li. Weathersyn: An instruction tuning mllm for weather forecasting report generation. *arXiv preprint arXiv:2605.07522*, 2026.

Zhiwang Zhou, Yuandong Pu, Xuming He, Yidi Liu, Yixin Chen, Junchao Gong, Xiang Zhuang, Wanghan Xu, Qinglong Cao, Shixiang Tang, et al. Omni-weather: A unified multimodal model for weather radar understanding and generation. In *International Conference on Learning Representations*, volume 2026, pp. 7471–7496, 2026.

A APPENDIX

APPENDIX TABLE OF CONTENTS

A.1 Details of MeteoText dataset 14
A.1.1 City-to-Grid Alignment and Variable Preprocessing 14
A.1.2 Precipitation Correction and Meteorological Grading 14
A.1.3 Salient City-Date Sampling 16
A.1.4 Dataset Statistics 16
A.1.5 Dataset Splits 16
A.2 Details of implementation 17
A.2.1 MeteoOmni 17
A.2.2 Baseline 18
A.3 Details of evaluation metrics 18
A.3.1 Text Quality Metrics 19
A.3.2 Meteorological Understanding Metrics 19
A.3.3 Field Generation Metrics 19
A.4 Additional experimental results for MeteoOmni 20
A.5 More qualitative results of MeteoOmni 21

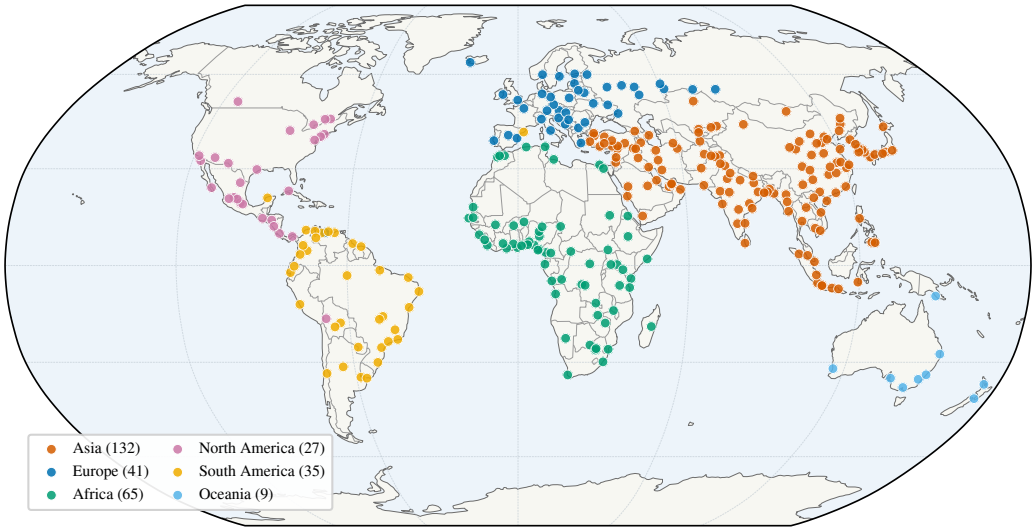

Figure 6: Spatial distribution of the selected cities.

A.1 Details of MeteoText dataset

This section provides additional construction details of MeteoText that are not fully covered in the preceding dataset description.

A.1.1 City-to-Grid Alignment and Variable Preprocessing

We construct a compact core set of cities rather than using all candidate cities. Candidate city metadata are obtained from GeoNames, including city names, countries, coordinates, and population. We first select the global top 300 cities by population, and then add the most populous city from each country that is not covered by these 300 cities. This procedure yields 314 cities across 131 countries. Fig. 6 shows the global distribution of the selected cities.

Each city is aligned to the same 128×256 global grid as the meteorological fields. Given latitude ϕ and longitude λ, the city grid location is computed as:

$$r = \left\lfloor \frac{90 - \phi}{180} H \right\rfloor, \quad c = \left\lfloor \frac{\lambda \bmod 360}{360} W \right\rfloor, \tag{7}$$

where $H = 128$ and $W = 256$. Row 0 corresponds to $90°$N, and column 0 corresponds to $0°$E.

For ERA5, hourly t2m, u10, and v10 are averaged into daily city-level values, while hourly tp is accumulated into 24-hour precipitation. For CMIP6, daily tas, pr, and sfcWind fields are used, with pr converted from precipitation flux to mm/day before city-level values are extracted.

A.1.2 Precipitation Correction and Meteorological Grading

ERA5 precipitation is corrected at city grid points using GHCN-Daily station observations. For each city, we search for stations within 150 km and use at most the three nearest valid stations. Daily precipitation is computed by inverse distance weighting after station records with quality control flags are removed. Specifically, for city i on day t, the corrected precipitation is computed as:

$$p_{i,t}^{\mathrm{GHCN}} = \frac{\sum_{s \in \mathcal{S}_i(t)} w_{i,s} p_{s,t}}{\sum_{s \in \mathcal{S}_i(t)} w_{i,s}}, \quad w_{i,s} = \frac{1}{\max(d_{i,s}, \epsilon)}, \tag{8}$$

where $\mathcal{S}_i(t)$ denotes the valid GHCN stations matched to city i on day t, $p_{s,t}$ is the precipitation observation at station s, $d_{i,s}$ is the distance between city i and station s, and ϵ is a small constant used to avoid division by zero. The corrected precipitation replaces ERA5 tp only at the corresponding

city grid point, while the remaining gridded precipitation field remains unchanged. This correction reduces the effect of spatial smoothing in gridded reanalysis precipitation, which can weaken local heavy-rainfall signals at the city scale.

After variable extraction, each city-date record is converted into meteorological grades using variable-specific criteria. For ERA5, wind speed is computed as $\sqrt{u10^2 + v10^2}$ and graded by the Beaufort scale. Precipitation is graded using fixed 24-hour accumulation thresholds, ranging from trace rain to extreme torrential rain. Temperature is graded by standardized anomaly relative to a day-of-year climatology computed from 1993-2016. For CMIP6, we use anomaly-based grading for all variables to account for model-specific climatology. For each year Y, the city-level climatology is computed from a 30-year rolling window $[Y-15, Y+14]$, truncated at the available data boundaries. The tas and sfcWind variables are graded directly by standardized anomaly. The pr variable is first converted to mm/day and then graded in $\log(1+x)$ space to reduce the effect of the long-tailed precipitation distribution. The standardized anomaly grades are divided into seven bins: extremely low, significantly low, slightly low, normal, slightly high, significantly high, and extremely high, with thresholds at -2.0, -1.5, -0.5, 0.5, 1.5, and 2.0. Tables 8-10 summarize the grading criteria for all variables.

Table 8: ERA5 wind grading based on the Beaufort scale.

Level	Wind speed range (m/s)	Label
0	$[0, 0.2]$	Calm
1	$(0.2, 1.5]$	Light Air
2	$(1.5, 3.3]$	Light Breeze
3	$(3.3, 5.4]$	Gentle Breeze
4	$(5.4, 7.9]$	Moderate Breeze
5	$(7.9, 10.7]$	Fresh Breeze
6	$(10.7, 13.8]$	Strong Breeze
7	$(13.8, 17.1]$	Near Gale
8	$(17.1, 20.7]$	Gale
9	$(20.7, 24.4]$	Strong Gale
10	$(24.4, 28.4]$	Storm
11	$(28.4, 32.6]$	Violent Storm
12	> 32.6	Hurricane

Table 9: ERA5 precipitation grading based on 24-hour accumulated precipitation.

Level	Precipitation range (mm/day)	Label
0	< 0.1	Trace Rain
1	$[0.1, 10.0)$	Light Rain
2	$[10.0, 25.0)$	Moderate Rain
3	$[25.0, 50.0)$	Heavy Rain
4	$[50.0, 100.0)$	Torrential Rain
5	$[100.0, 250.0)$	Severe Torrential Rain
6	≥ 250.0	Extreme Torrential Rain

Table 10: Standardized-anomaly grading used for ERA5 temperature and CMIP6 variables.

Level	Standardized anomaly range	Label
0	$z < -2.0$	Extremely / Anomalously Low
1	$[-2.0, -1.5)$	Significantly Low
2	$[-1.5, -0.5)$	Slightly Low
3	$[-0.5, 0.5]$	Normal
4	$(0.5, 1.5]$	Slightly High
5	$(1.5, 2.0]$	Significantly High
6	$z > 2.0$	Extremely / Anomalously High

A.1.3 SALIENT CITY-DATE SAMPLING

We sample salient city-date records rather than uniformly sampling all days, since normal weather cases would otherwise dominate the generated descriptions. Filtering is performed independently for each city and year, allowing the thresholds to adapt to local climate conditions and interannual variability. If a city-date sample satisfies multiple salient weather criteria, we keep only the highest-priority category. After filtering, daily files are randomly sampled until the target number of city-date samples is reached. The category distribution of the sampled records is shown in Fig. 7.

Table 11: Salient weather sampling rules. Thresholds are computed for each city-year.

Category	Selection Rule
Extreme precipitation	Top 10% of rainy days
Extreme wind	Top 10% of wind-speed days
Extreme hot temperature	Top 5% of temperature days
Extreme cold temperature	Bottom 5% of temperature days
Multi-category priority	Precipitation, hot, cold, wind

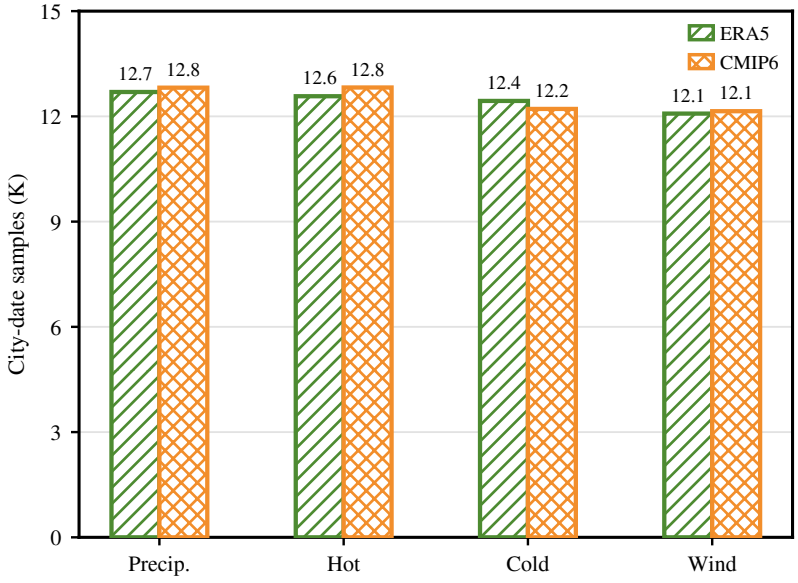

Figure 7: Distribution of sampled city-date records by salient-weather category.

A.1.4 DATASET STATISTICS

The final dataset contains 99,788 city-date text samples, including 49,788 ERA5 samples and 50,000 CMIP6 samples. Fig. 9 and Fig. 8 summarize the temporal and spatial coverage of the final samples. Temporally, MeteoText combines two complementary sources. The ERA5 subset covers 1979–2018 and provides modern reanalysis-based records with a relatively uniform yearly distribution, while the CMIP6 subset covers 1911–2014 and extends the dataset to earlier decades where direct reanalysis coverage is unavailable. This design gives MeteoText both reliable modern weather-text alignment and long term historical coverage. Spatially, the samples are distributed across all major populated continents. Asia contributes the largest share, followed by Africa, Europe, South America, North America, and Oceania. This pattern is consistent with the population-based city selection strategy, but the country-fallback rule ensures that the dataset is not limited to a few densely populated regions and maintains broad geographic coverage across diverse climate zones.

A.1.5 DATASET SPLITS

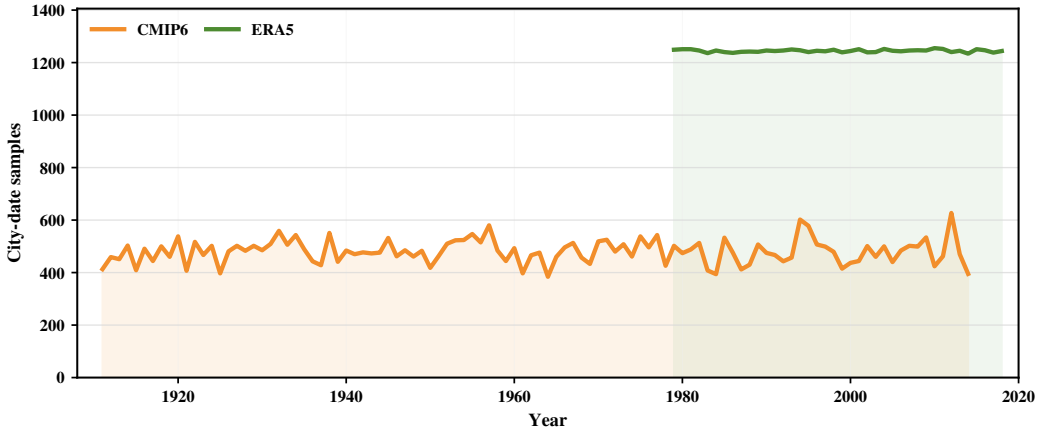

Figure 9: Temporal distribution of final sampled city-date records.

Table 12: Dataset splits for the five MeteoText tasks and the real-world OOD evaluation. The in-distribution splits are chronologically separated with no temporal overlap between training, validation, and test sets. The CMA warning set is used exclusively for zero-shot OOD evaluation.

Setting	Task	Data Source	Train	Validation	Test
ID	City Weather QA	ERA5	1979–2014	2015	2016
	CMIP6 Text Quantification	CMIP6	1911–2007	2008–2011	2012
	ERA5 Text Quantification	ERA5	1979–2007	2008–2011	2012
	Masked ERA5 Reconstruction	ERA5	1979–2007	2008–2011	2012
	CMIP6 Correction	ERA5 + CMIP6	1979–2007	2008–2011	2012
OOD	ERA5 Text Quantification	CMA text + ERA5	–	–	2021–2023
	Masked ERA5 Reconstruction	CMA text + ERA5	–	–	2021–2023

We use chronological splits with no temporal overlap between training, validation, and test sets. For T2–T5, data up to 2007 are used for training, 2008–2011 for validation, and 2012 for testing. For T1, we use 1979–2014 for training, 2015 for validation, and 2016 for testing. For T5, only the overlapping period between ERA5 and CMIP6 is used.

We additionally evaluate on a real-world OOD set of 739 CMA warning samples from 22 Chinese cities during 2021–2023. This set is used only for zero-shot evaluation on T3 and T4, without further training or adaptation. The complete splits are summarized in Table 12.

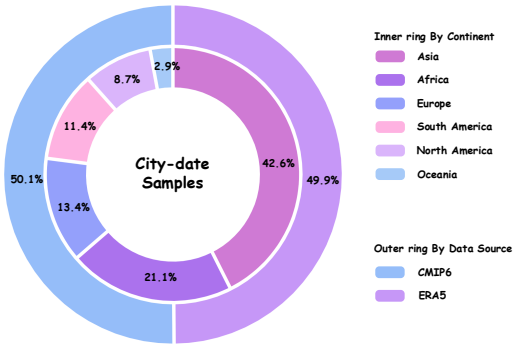

Figure 8: Distribution of samples by continent and data source.

A.2 Details of Implementation

This section provides the implementation details of MeteoOmni and the baseline models used in the field generation experiments.

A.2.1 MeteoOmni

Training. We train MeteoOmni by fine-tuning the pretrained BAGEL-7B-MoT. Training is performed with PyTorch distributed training on 4 NVIDIA H200 GPUs. The model is wrapped with FSDP using `HYBRID_SHARD`, bfloat16 mixed precision, activation checkpointing, and EMA weights. The VAE is frozen during training, while the language model and visual understanding

modules are fine-tuned. We use AdamW with a learning rate of 2×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 10^{-15}$, and zero weight decay. The learning rate is linearly warmed up for the first 2,000 steps and remains constant thereafter. Training runs for 200K optimizer steps. Each packed batch uses a soft target of 5,120 tokens, with a maximum of 5,120 tokens per sample and 5,120 tokens per packed batch. We use a gradient accumulation step of 1 and clip the gradient norm to 1.0. For multi-task training, the task weights λ_t are set to 1 for T1–T3 and 4 for T4–T5, and the Value Probe loss weight β is set to 0.1.

Inference. TDOG is applied only during inference and introduces no additional parameter updates. For each generation sample, we maintain $K = 4$ parallel latent particles over 50 sampling steps. Every $\Delta = 10$ steps, we estimate and decode the clean latent state of each particle, and evaluate its consistency with the text-derived pseudo-observations. The particle likelihood in Eq. 3 is computed jointly over t2m, u10, v10, and tp, using an observation noise of $\sigma = 1.0$ in the normalized decoded space. Pseudo-observations are obtained from MeteoOmni's own T3 text-quantification predictions and normalized using the same ERA5 statistics as the generation model; no ground-truth meteorological values are used during inference.

A.2.2 Baseline

For the two field generation tasks, we compare MeteoOmni with two task-specific baselines trained under the same data splits and evaluation protocol. SynWeatherDiff Xu et al. (2026) follows the diffusion training objective Ho et al. (2020) and uses DDIM sampling Song et al. (2020) at inference. It operates in the ERA5 latent space and uses a transformer denoiser to generate complete ERA5 latent fields. The other baseline is deterministic pixel space regressors based on UNet Ronneberger et al. (2015). For Masked ERA5 Reconstruction, UNet maps masked ERA5 inputs to complete ERA5 fields. For CMIP6 Correction, it map CMIP6 inputs to ERA5 fields. Table 13 reports the main model and optimization hyperparameters.

Table 13: Hyperparameters of the baseline models.

Module	Hyperparameter	Value
SynWeatherDiff	Input space	ERA5 latent space
	Latent size	$16 \times 32 \times 64$
	Spatial downsampling ratio	$4\times$
	Denoiser	Transformer denoiser
	Patch size	2
	Hidden size	512
	Transformer depth	12
	Attention heads	8
	MLP ratio	4.0
	Diffusion timesteps	1,000
	Beta schedule	Linear
	Beta range	10^{-4} to 0.02
	Sampling method	DDIM, 20 steps
UNet	Input space	Pixel space
	Image size	128×256
	Input channels for T4	4
	Input channels for T5	3
	Output channels	4
	Base channels	64
	Channel multipliers	$[1, 2, 4, 8]$
	Residual blocks per stage	2
	Attention resolutions	$[16, 8]$
	GroupNorm groups	32

A.3 Details of evaluation metrics

This section provides formal definitions of the evaluation metrics used in our experiments.

A.3.1 TEXT QUALITY METRICS

For *City Weather QA*, text quality metrics are used to evaluate whether the generated weather description is close to the reference description in wording and semantic content. We report ROUGE-L Lin (2004), METEOR Banerjee & Lavie (2005), and BERTScore Zhang et al. (2019). ROUGE-L is based on the longest common subsequence between the generated and reference descriptions, and is used to measure sequence-level lexical overlap. METEOR aligns generated and reference tokens with exact matching, stemming, and synonym matching, and is used to evaluate lexical similarity with better tolerance to word variation. BERTScore compares contextual token embeddings between the generated and reference descriptions, and is used to measure semantic similarity beyond exact lexical overlap. These metrics evaluate text similarity, but they do not directly measure whether the generated description preserves the correct meteorological grades. We therefore also report meteorological understanding metrics.

A.3.2 METEOROLOGICAL UNDERSTANDING METRICS

For City Weather QA, Meteorological Grade Accuracy (MGA) is used in addition to text similarity metrics. MGA evaluates whether the generated description accurately reflects the discrete meteorological grades of each variable, such as temperature, wind, and precipitation. It therefore measures meteorological correctness rather than textual overlap with the reference description.

Let $\mathcal{A} = \{\text{temperature}, \text{wind}, \text{precipitation}\}$ denote the evaluated meteorological attributes. MGA is computed only on samples where the corresponding attribute is expressed in the generated description and the reference grade is available. Let $\mathcal{I}_a$ denote the subset of valid samples for attribute a, and let $g_{i,a}$ and $\hat{g}_{i,a}$ denote the reference and predicted grades. The variable-wise MGA is computed as:

$$\mathrm{MGA}_a = \frac{1}{|\mathcal{I}_a|} \sum_{i \in \mathcal{I}_a} \mathbf{1}\left[\hat{g}_{i,a} = g_{i,a}\right]. \tag{9}$$

The overall MGA is the average over attributes with non-empty evaluation subsets. If a variable is not expressed in the generated description or the reference grade is unavailable, that sample is excluded from the MGA denominator for the corresponding attribute.

For *CMIP6 Text Quantification* and *ERA5 Text Quantification*, clamped APE is used to evaluate the numerical consistency between the predicted meteorological values and the reference values Bell et al. (2025). It measures relative numerical error and clamps large errors so that the score remains bounded. For one scalar value, the score is computed as:

$$\mathrm{APE}(y, \hat{y}) = 1 - \min\left(\frac{|y - \hat{y}|}{|y| + \epsilon}, 1\right), \tag{10}$$

where y is the reference value, $\hat{y}$ is the predicted value, and ϵ is a small constant for numerical stability. The score is bounded between 0 and 1, with a higher value indicating more accurate text-to-value quantification. For a variable set $\mathcal{V}$, the overall score is computed as:

$$\mathrm{APE}_{\mathrm{all}} = \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \mathrm{APE}(y_v, \hat{y}_v). \tag{11}$$

For CMIP6 quantification, $\mathcal{V}$ includes temperature, wind speed, and precipitation. For ERA5 quantification, $\mathcal{V}$ includes t2m, u10, v10, and tp, and we additionally report wind magnitude computed from u10 and v10.

A.3.3 FIELD GENERATION METRICS

RMSE is used to evaluate the pointwise deviation between the generated field and the target ERA5 field. Given an evaluation region Ω, RMSE is computed as

$$\mathrm{RMSE} = \sqrt{\frac{1}{|\Omega|} \sum_{j \in \Omega} \left(\hat{X}_j - X_j\right)^2}, \tag{12}$$

where X_j and $\hat{X}_j$ denote the target and generated values at grid location or variable-location index j. Lower RMSE indicates a generated field that is closer to the ERA5 target. For Masked ERA5

Table 14: Effect of weather text information on T3 ERA5 Text Quantification.

Input	A-T ↑	A-U ↑	A-V ↑	A-P ↑	A-O ↑
City + Date only	0.638	0.024	0.023	0.209	0.224
Shuffled text	0.563	0.019	0.014	0.163	0.190
Correct text	**0.872**	**0.701**	**0.702**	**0.379**	**0.664**

Table 15: Comparison between MeteoOmni and the ground truth meteorological grade reference on T3 ERA5 Text Quantification using the clamped APE score (↑). The grade reference samples continuous values from the ground truth grade intervals without access to the weather descriptions.

Variable	MeteoOmni	Grade Reference
t2m	**0.872**	0.852
u10	**0.701**	0.527
v10	**0.702**	0.519
tp	**0.379**	0.338

Reconstruction, Ω is the masked region, so the metric measures reconstruction quality only where the input field is removed. For CMIP6 Correction, we report both global RMSE over all grid cells and city-level RMSE over 5×5 regions centered at the selected city grid points. The city-level metric measures local consistency with city weather descriptions, while the global metric evaluates the quality of the full generated field.

A.4 Additional experimental results for MeteoOmni.

Effect of Weather Text Information. To test whether T3 relies on weather semantics rather than city/date priors, we compare three settings: *City + Date only*, *Shuffled text*, and *Correct text*. For *Shuffled text*, the weather description is replaced by one from another date of the same city and month, while keeping the original city, date, and target unchanged. As shown in Table 14, *Correct text* consistently outperforms both control settings. The gap over *City + Date only* indicates that city-specific climatology and seasonal information alone are insufficient to explain the quantification performance. Moreover, replacing the correct description with a mismatched report substantially degrades performance despite preserving the same city and seasonal context, showing that the model is sensitive to the meteorological state expressed in the text. The degradation is particularly pronounced for wind components, which are difficult to infer from climatological priors alone, while temperature retains relatively stronger performance under the metadata-only setting due to its stronger seasonal dependence. Overall, these results show that T3 extracts event-specific meteorological information from weather descriptions beyond city- and date-dependent priors.

Information Beyond Discrete Meteorological Grades. A natural question is whether T3 simply recovers the discrete meteorological grades used to construct the synthetic descriptions, rather than exploiting additional information expressed in language. We therefore compare MeteoOmni with a grade reference that samples continuous values from the ground truth grade intervals. For temperature, the sampled anomaly value is converted to absolute temperature using the corresponding climatology. As shown in Table 15, the grade reference already achieves 0.852 for temperature, close to MeteoOmni at 0.872, indicating that the anomaly grade and climatological background account for much of the temperature information. For wind, the grade reference samples wind speed from the corresponding Beaufort interval and combines it with the wind direction to recover the zonal and meridional components. It achieves 0.527 and 0.519 for u10 and v10, while MeteoOmni reaches 0.701 and 0.702, respectively. This improvement suggests that MeteoOmni learns the conditional distribution of continuous wind values given the discrete wind attributes and contextual information, rather than simply inverting the grade intervals. For precipitation, MeteoOmni similarly improves over the grade reference (0.379 vs. 0.338), indicating that it can better capture the non-uniform distribution of continuous values within each precipitation grade.

A.5 MORE QUALITATIVE RESULTS OF METEOOMNI

This section provides additional qualitative examples across all tasks in MeteoText. Fig 10-12 cover City Weather QA, CMIP6 Text Quantification, ERA5 Text Quantification, Masked ERA5 Reconstruction, and CMIP6 Correction. Fig. 13 shows an extreme-event ablation case in Yekaterinburg on 2012-06-05. Compared with the raw CMIP6 input and the variant without TDOG, TDOG better recovers the local precipitation pattern, preserves the temperature anomaly around the selected city, and produces wind gradients and directions that are more consistent with the ERA5 target.

We further provide a real-world example based on a CMA high-temperature warning issued by the Xiamen Meteorological Observatory on July 24, 2021. As shown in Fig. 14, the original warning is normalized by resolving the valid date and retaining meteorologically relevant information, including the affected locations and reported temperature ranges. We then use the normalized weather description in Masked ERA5 Reconstruction. As shown in Fig. 15, incorporating the real-world text provides additional local observational constraints and produces a reconstructed temperature pattern around Xiamen that is more consistent with the ERA5 target than the reconstruction without text.

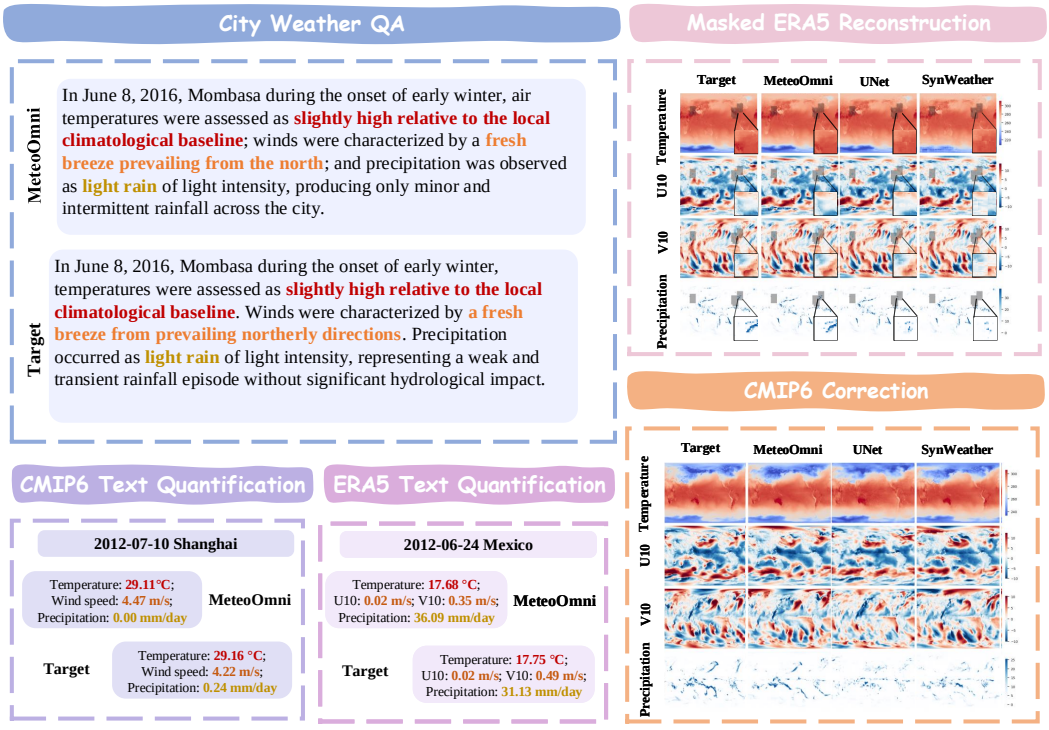

Figure 10: Qualitative results across all tasks.

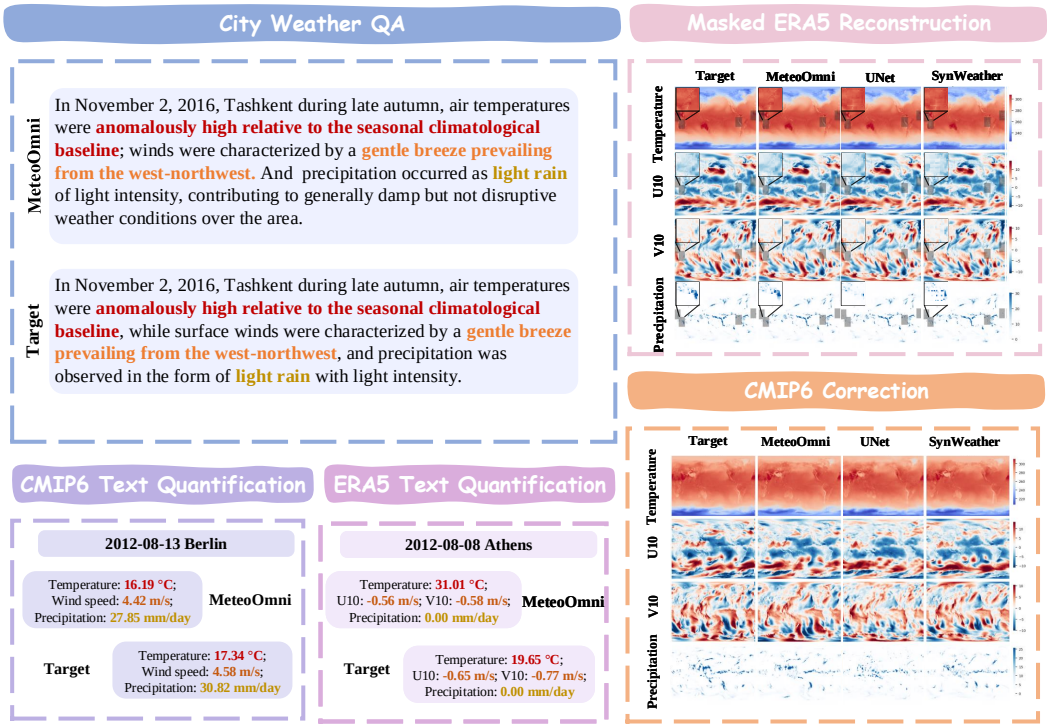

Figure 11: Qualitative results across all tasks.

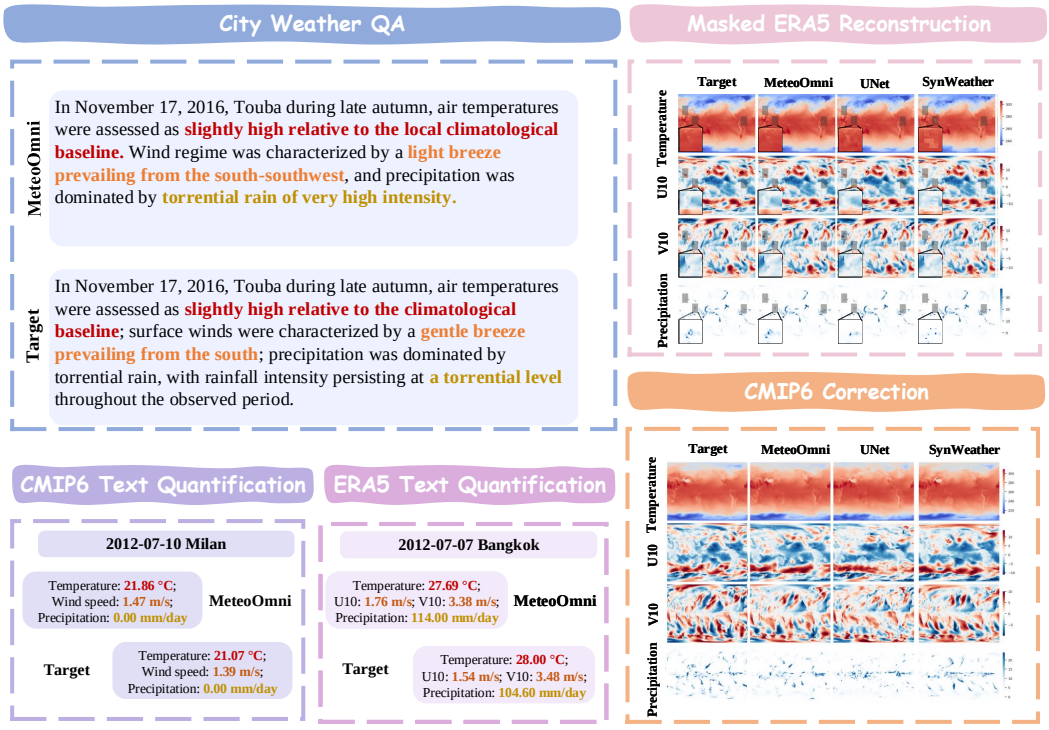

Figure 12: Qualitative results across all tasks.

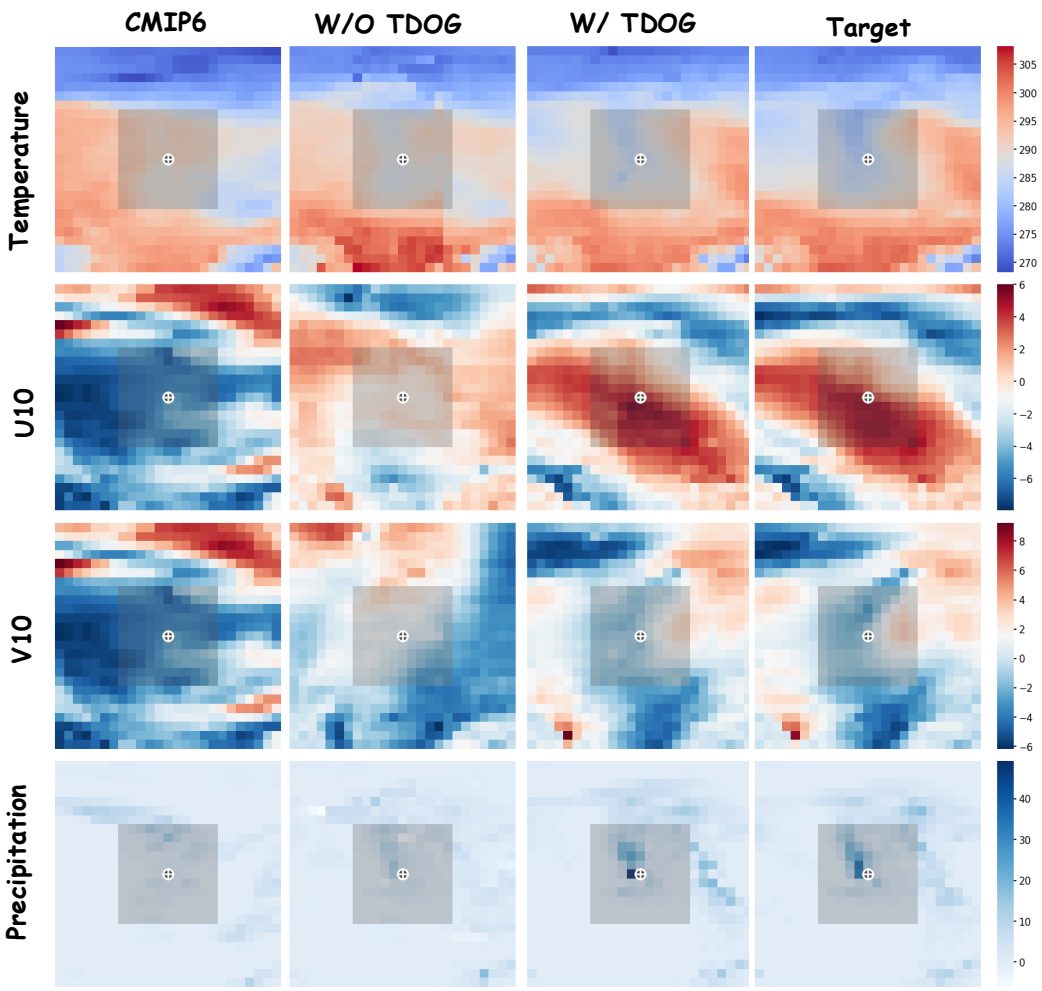

Figure 13: Ablation case for CMIP6 Correction on an extreme event in Yekaterinburg on 2012-06-05.

Raw Weather Warning:
Xiamen Meteorological Observatory continued to issue an Orange Warning Signal for High Temperatures at 08:49 on July 24, 2021: Xiamen is expected to continue experiencing hot weather today. In some towns and subdistricts of Huli District and Siming District, maximum temperatures are expected to reach 35-37°C, with localized areas exceeding 37°C. In some towns and subdistricts of Jimei District and Haicang District, maximum temperatures are expected to reach 36-38°C, with localized areas exceeding 39°C. Please take precautions against the heat and implement appropriate heatstroke-prevention and cooling measures. Tong'an District and Xiang'an District should follow locally issued warning signals.

Normalized Weather Description:
On July 24, 2021, Xiamen was expected to continue experiencing high temperatures. In some parts of the city, maximum temperatures were expected to reach 35-37°C, with localized areas exceeding 37°C.In other parts, maximum temperatures were expected to reach 36-38°C, with localized areas exceeding 39°C.

Figure 14: Example of real-world weather-text processing. The raw CMA high-temperature warning is normalized into a weather description by resolving its valid date and retaining meteorologically relevant information.

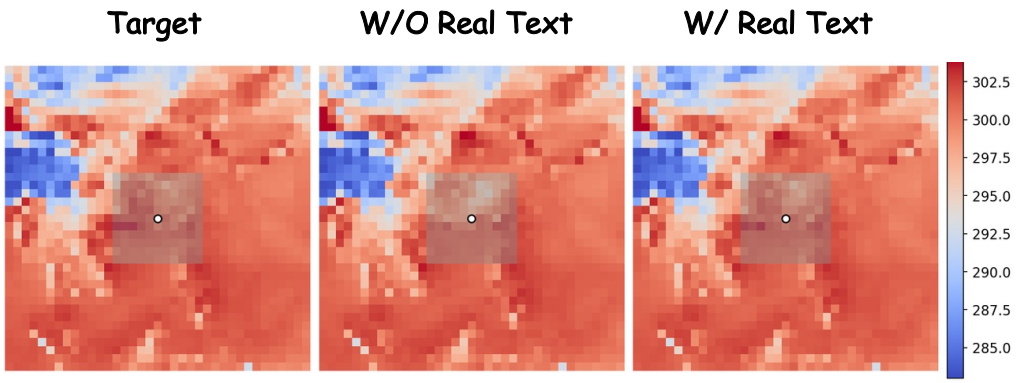

Figure 15: Qualitative comparison of Masked ERA5 Reconstruction with and without real-world weather text. The white circle marks Xiamen. Incorporating the normalized weather description provides additional local observational constraints and guides the reconstructed temperature field toward the ERA5 target.