OMNI-WEATHER: UNIFIED MULTIMODAL FOUNDATION MODEL FOR WEATHER GENERATION AND UNDERSTANDING

Anonymous authors
Paper under double-blind review

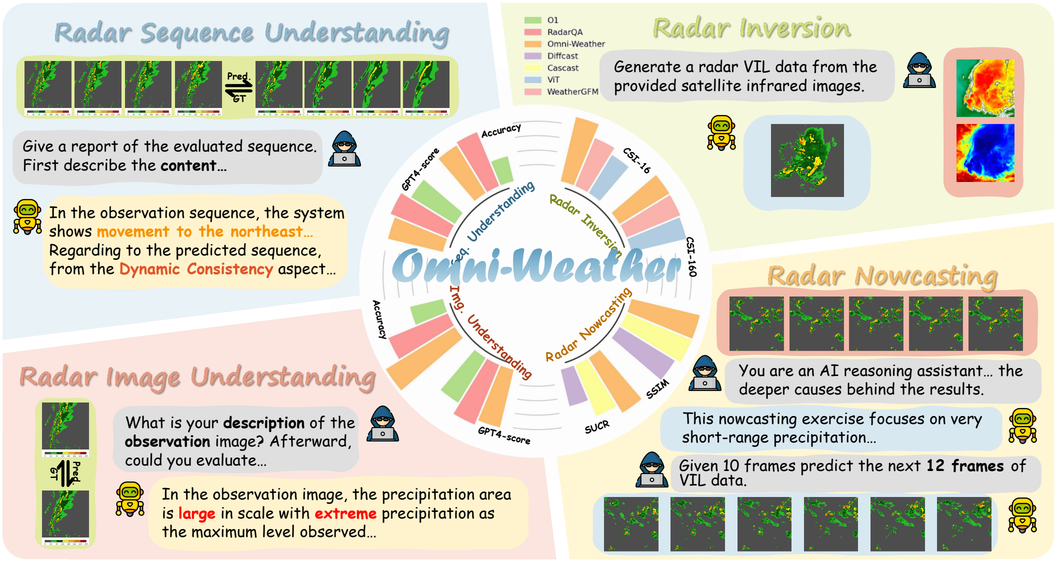

Radar Sequence Understanding
Pred.
GT
Give a report of the evaluated sequence. First describe the **content**...
In the observation sequence, the system shows movement to the northeast... Regarding to the predicted sequence, from the Dynamic Consistency aspect...
Radar Inversion
🟩 O1
🟥 RadarQA
🟧 Omni-Weather
🟪 Diffcast
🟨 Cascast
🟦 ViT
🟥 WeatherGFM
Generate a radar VIL data from the provided satellite infrared images.
Accuracy
GPT4-score
CSI-16
CSI-160
Seq. Understanding
Radar Inversion
Omni-Weather
Img. Understanding
Radar Nowcasting
Accuracy
GPT4-score
SUCR
SSIM
Radar Nowcasting
You are an AI reasoning assistant... the deeper causes behind the results.
This nowcasting exercise focuses on very short-range precipitation...
Given 10 frames predict the next **12 frames** of VIL data.
Radar Image Understanding
GT
Pred.
What is your **description** of the **observation** image? Afterward, could you evaluate...
In the observation image, the precipitation area is large in scale with extreme precipitation as the maximum level observed...

Figure 1: **Illustration of Omni-Weather's unified capabilities.**

ABSTRACT

Weather modeling requires both accurate prediction and mechanistic interpretation, yet existing methods treat these goals in isolation, separating generation from understanding. To address this gap, we present **Omni-Weather**, the first multimodal foundation model that unifies weather generation and understanding within a single architecture. Omni-Weather integrates a radar encoder for weather generation tasks, followed by unified processing using a shared self-attention mechanism. Moreover, we construct a Chain-of-Thought dataset for causal reasoning in weather generation, enabling interpretable outputs and improved perceptual quality. Extensive experiments show Omni-Weather achieves state-of-the-art performance in both weather generation and understanding. Our findings further indicate that generative and understanding tasks in the weather domain can mutually enhance each other. Omni-Weather also demonstrates the feasibility and value of unifying weather generation and understanding.

1 INTRODUCTION

A significant trend in AI research is the rise of foundation models that unify *generation* and *understanding* within a single architecture. Multimodal LLMs such as InternVL Chen et al. (2024c), UniGen Tian et al. (2025), and Lumina-omnilv Pu et al. (2025) demonstrate that perception and synthesis can be integrated seamlessly, achieving strong generalization across visual and textual domains. These advances highlight the opportunity to extend unified generation–understanding paradigms to weather domain, where both predictive accuracy and interpretability are essential.

Recently, weather generation and understanding tasks have made notable progress. On the generation side, nowcasting models such as PreDiff Gao et al. (2023), DiffCast Yu et al. (2024), and CasCast Gong et al. (2024) forecast convective evolution from radar sequences, supporting early warnings of hazards like flooding. Radar inversion methods He et al. (2025b) further reconstruct radar observables from satellite channels, enabling precipitation monitoring in regions without radar coverage. On the understanding side, models such as RadarQA He et al. (2025a) and WeatherQA Ma et al. (2024) generate diagnostic reports or identify severe weather impacts from atmospheric fields.

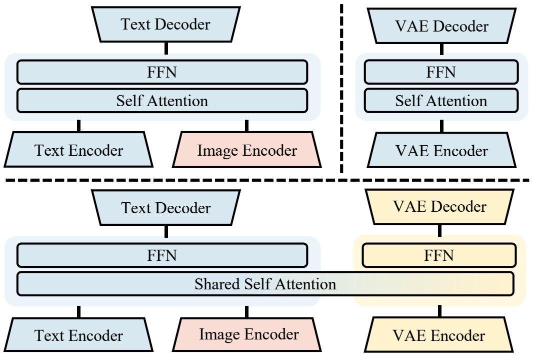


Figure 2: **Comparison** between separated architectures for weather understanding / generation (top) and unified framework with shared self-attention (bottom).

Despite these advances, unified architectures remain absent in the weather domain. As shown in Figure 2, existing approaches are divided into two disjoint paradigms: model such as ClimaX Nguyen et al. (2023) and WeatherGFM Zhao et al. (2024) excel at forecasting and downscaling but lack interpretation, while understanding models such as RadarQA He et al. (2025a) and WeatherQA Ma et al. (2024) provide diagnostic reasoning yet cannot synthesize physical fields. However, atmospheric systems are inherently multiscale, shaped by storm genesis, intensification and decay, where accurate prediction is often accompanied by the need for mechanistic interpretation. Moreover, extreme events such as rapid intensification of cyclones demand models that can not only predict hazardous outcomes but also explain the underlying drivers for actionable decision-making. Current studies isolate these links—*generative nowcasting models do not understand radar observations, yet MLLMs do not predict radar variables*. Bridging this gap with a foundation model that unifies generation and understanding is therefore an urgent requirement for weather domain.

To this end, we propose **Omni-Weather**, a unified multimodal foundation model for both weather generation and understanding. By consolidating these tasks within a shared backbone (Figure 2, bottom), we further propose a Chain-of-Thought dataset tailored for causal reasoning in generation tasks, which enables Omni-Weather to be finetuned with explicit reasoning supervision and to perform thinking inference. Through this integration, Omni-Weather bridges predictive accuracy with interpretability, marking a step toward reasoning unified foundation models for weather.

The main contributions of this work can be summarized as follows:

- We introduce the first multimodal foundation model in weather that jointly addresses both generation tasks (e.g., nowcasting, inversion) and understanding tasks (e.g., diagnostic reasoning, question answering) within a single model.
- We demonstrate that training both generation and understanding tasks together provides complementary supervision signals, enabling Omni-Weather to learn more transferable representations of storm evolution and improving performance on both sides.
- We propose a Chain-of-Thought (CoT) dataset and explore its integration into weather generation, enhancing perceptual quality and interpretability as a first step toward explainable generative modeling in weather domain.

2 Related Work

Weather generation models.

Weather generation models aim to synthesize physically consistent weather fields from historical or multi-modal observations Han et al. (2024); Chen et al. (2023); Gao et al. (2023); Yu et al. (2024); Gao et al. (2022); Lam et al. (2023). Examples include DiffSR He et al. (2025b), which reconstructs composite radar reflectivity from satellite infrared and lightning inputs, and CasCast Gong et al. (2024), which predicts precipitation evolution from past VIL sequences. Beyond convective-scale nowcasting, GenCast Price et al. (2023) and Stormer Nguyen et al. (2024) scale diffusion and transformer architectures to medium-range ensemble forecasting, while Pangu-Weather Bi et al. (2022) enables fast and accurate 3D high-resolution global forecasts. More recently, foundation-scale Earth-system models such as ClimaX Nguyen et al. (2023), FengWu Chen et al. (2023); Han et al. (2024), Aurora Bodnar et al. (2025), and Prithvi WxC Schmude et al. (2024) extend the foundation-model paradigm Wiggins & Tejani (2022) to climate and weather, and WeatherGFM Zhao et al. (2024) introduces in-context learning for generalist nowcasting and inversion. Parallel to these modeling efforts, CLLMate Li et al. (2025) proposes a multimodal benchmark for weather and climate event forecasting that unifies several datasets, including SEVIR, which is also the primary radar dataset used in this work; while we follow prior radar nowcasting literature in focusing on SEVIR, a systematic evaluation on the broader CLLMate suite is an interesting direction for future work. However, existing models and benchmarks are primarily designed for medium-range or global forecasting and downscaling, and they do not explicitly address understanding or reasoning at the event level, leaving interpretability and evaluation of their predictions largely unexplored.

Weather understanding models aim to interpret weather signals and provide human-readable insights, often through natural language or diagnostic reasoning. Early studies adapt pretrained language models such as ClimateBERT Webersinke et al. (2021); Schimanski et al. (2023) and ClimateNLP Krishnan & Anoop (2023) to analyze textual weather reports, focusing on tasks such as climate risk assessment, report classification, and domain adaptation. More recent work emphasizes multimodal inputs, combining imagery with text. For example, WeatherQA Ma et al. (2024) takes 20 images of atmospheric parameters to predict regions impacted by severe convection, while RadarQA He et al. (2025a) leverages both radar observations and numerical forecasts to generate expert-like quality assessment reports. In parallel, Aquilon Varambally et al. takes an initial step toward multimodal weather LLMs that jointly process gridded fields and language, showing that our model is among the first multimodal architectures specifically targeting the weather/radar domain rather than standing entirely on its own. These approaches demonstrate the feasibility of applying large language models to meteorology, but they remain largely limited to understanding tasks alone. In particular, existing models specialize in either textual analysis or visual reasoning without integrating predictive generation, leaving the connection between physical simulation and diagnostic interpretation underexplored.

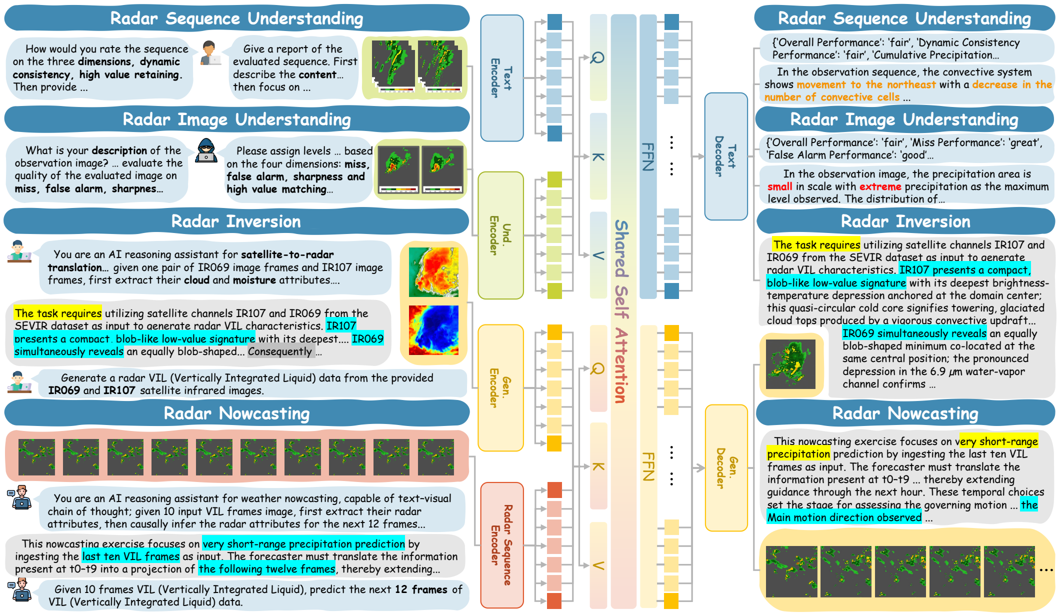
Unified multimodal models integrate visual understanding and generation within a single architecture, leveraging advances in LLMs and diffusion models Chen et al. (2025a; 2024b;a; 2025b); Pu et al. (2025); Zhao et al. (2024). Transfusion Zhou et al. (2024) unifies text prediction and image diffusion within a single transformer trained end-to-end on both modalities. LMFusion Shi et al. (2024) adapts pretrained text-only LLMs by freezing language modules and introducing parallel image-specific branches for efficient multimodal generation. MetaMorph Tong et al. (2024) employs Visual-Predictive Instruction Tuning (VPiT) to enable LLMs to jointly predict text and continuous visual tokens from multimodal instructions. MetaQuery Pan et al. (2025) connects frozen MLLMs with diffusion decoders using learnable queries, enabling generation without compromising understanding capabilities. BLIP3-o Chen et al. (2025a) sequentially combines autoregressive modeling and diffusion to generate CLIP-aligned visual features, achieving state-of-the-art performance across modalities. BAGEL Deng et al. (2025) scales unified modeling through pretraining on interleaved text-image-video data, demonstrating emergent multimodal reasoning and manipulation abilities.

3 METHOD

In this section, we first introduce a unified representation of weather generation and understanding tasks, where radar nowcasting, radar inversion, and radar image / sequence understanding are formulated under a consistent sequence-to-sequence paradigm. We then present Omni-Weather, a multimodal foundation model that integrates these tasks within a shared backbone, with a detailed exposition of its architecture, modality-specific encoders, and multi-task training objectives. Finally, we describe the integration of chain-of-thought reasoning, including the construction of causal annotations and their incorporation in both training and inference, which not only enhances interpretability but also improves the perceptual quality of weather forecasts.

3.1 UNIFIED REPRESENTATION OF WEATHER GENERATION AND UNDERSTANDING TASKS

Weather modeling encompasses a wide range of objectives, from predicting future radar fields Gao et al. (2022) to generating textual assessments of forecast quality He et al. (2025a). To systematically



How would you rate the sequence on the three **dimensions, dynamic consistency, high value retaining.** Then provide ...

Give a report of the evaluated sequence. First describe the **content**... then focus on ...

What is your **description** of the observation image? ... evaluate the quality of the evaluated image on **miss, false alarm, sharpnes**...

Please assign levels ... based on the four dimensions: **miss, false alarm, sharpness and high value matching**...

You are an AI reasoning assistant for **satellite-to-radar translation**... given one pair of IR069 image frames and IR107 image frames, first extract their **cloud** and **moisture** attributes....

The task requires utilizing satellite channels IR107 and IR069 from the SEVIR dataset as input to generate radar VIL characteristics. IR107 presents a compact, blob-like low-value signature with its deepest.... IR069 simultaneously reveals an equally blob-shaped... Consequently ...

Generate a radar VIL (Vertically Integrated Liquid) data from the provided **IR069** and **IR107** satellite infrared images.

You are an AI reasoning assistant for weather nowcasting, capable of text-visual chain of thought; given 10 input VIL frames image, first extract their radar attributes, then causally infer the radar attributes for the next 12 frames...

This nowcasting exercise focuses on very short-range precipitation prediction by ingesting the last ten VIL frames as input. The forecaster must translate the information present at t0–t9 into a projection of the following twelve frames, thereby extending...

Given 10 frames VIL (Vertically Integrated Liquid), predict the next **12 frames** of VIL (Vertically Integrated Liquid) data.

{'Overall Performance': 'fair', 'Dynamic Consistency Performance': 'fair', 'Cumulative Precipitation...

In the observation sequence, the convective system shows **movement to the northeast** with a **decrease in the number of convective cells** ...

{'Overall Performance': 'fair', 'Miss Performance': 'great', 'False Alarm Performance': 'good'...

In the observation image, the precipitation area is **small** in scale with **extreme** precipitation as the maximum level observed. The distribution of...

The task requires utilizing satellite channels IR107 and IR069 from the SEVIR dataset as input to generate radar VIL characteristics. IR107 presents a compact, blob-like low-value signature with its deepest brightness-temperature depression anchored at the domain center; this quasi-circular cold core signifies towering, glaciated cloud tops produced by a vigorous convective updraft... IR069 simultaneously reveals an equally blob-shaped minimum co-located at the same central position; the pronounced depression in the 6.9 μm water-vapor channel confirms ...

This nowcasting exercise focuses on very short-range precipitation prediction by ingesting the last ten VIL frames as input. The forecaster must translate the information present at t0–t9 ... thereby extending guidance through the next hour. These temporal choices set the stage for assessing the governing motion ... the Main motion direction observed ...

Figure 3: **Framework and Task paradigm** of Omni-Weather.

organize this diversity, we categorize the tasks into two groups: *weather generation*, which focuses on producing future or cross-modal meteorological fields; *weather understanding*, which requires generating natural-language descriptions and evaluations. To support both categories, we leverage the SEVIR dataset Veillette et al. (2020), which provides time-aligned radar and satellite sequences of severe weather events. Below, we detail our task paradigm corresponding to each category.

Weather Generation. Radar nowcasting aims to predict the short-term evolution of precipitation fields. Specifically, given ten VIL frames, the model generates the subsequent twelve frames, thereby forecasting the spatio-temporal dynamics of convective systems. Radar inversion focuses on translating satellite observations into radar-derived quantities. In this task, two infrared channels (IR069 and IR107) are provided as input, and the objective is to reconstruct the corresponding VIL field, which requires a cross-modal mapping between satellite imagery and radar measurements.

Weather Understanding. Radar understanding tasks require the model to generate natural-language descriptions or structured assessments from radar observations and model forecasts. The input can be either a single VIL frame or a temporal sequence of frames, while the output is expected to cover key aspects, including storm morphology, intensity, temporal evolution, and forecast quality (e.g., misses or false alarms). This formulation follows RadarQA He et al. (2025a), where textual reports for frame and sequence are designed to support expert interpretation beyond traditional weather forecast metrics, which not only aligns more closely with real-world meteorological analysis, but also unifies image-level and sequence-level evaluation.

3.2 OMNI-WEATHER: FOUNDATION MODEL FOR WEATHER GENERATION AND UNDERSTANDING

Unified multimodal Model. Inspired by recent advances in unified multimodal foundation models such as Bagel-7B-MoT Deng et al. (2025), we design Omni-Weather to handle both *generation* and *understanding* tasks within a single architecture. Instead of training separate models for each objective, all tasks are expressed in a consistent sequence-to-sequence format. Given a task-specific prompt p_t, a radar sequence input x_t and target output y_t (e.g., future radar sequence or radar assessment report), the model learns the mapping

$$y_t = F_\theta(p_t, x_t), \tag{3.1}$$

where F_θ denotes the shared transformer backbone. This formulation allows a single model to flexibly switch across tasks by conditioning on p_t while maintaining unified training.

Architecture. As shown in Figure 3, Omni-Weather unifies generation and understanding within a single backbone by embedding all task prompts through the *text encoder*, thereby ensuring a shared textual space for conditioning across diverse tasks. In contrast, generation and understanding tasks have varying feature processing approches for vision model. Specifically, For *Radar Image / Sequence Understanding*, visual inputs (e.g., a single VIL frame or a twelve-frame VIL sequence) are encoded by the *understanding encoder* and concatenated with the corresponding text prompt; the fused tokens are subsequently processed by shared self-attention layers, and the *text decoder* produces natural-language descriptions. For *Radar Inversion*, satellite channels are embedded by the *generation encoder*, fused in the shared self attention layers, and decoded by the *VAE decoder* to reconstruct the VIL field. For *Radar Nowcasting* tasks, ten input VIL frames are encoded by *radar sequence encoder*. Specifically, we instantiate this temporal encoder with EarthFormer Gao et al. (2022) to provide motion-aware tokens that stabilize long-horizon dynamics and improve temporal coherence. Since directly forcing the backbone to learn multi-frame evolution with Gen Encoder proved less stable, conditioning the shared attention layers on EarthFormer's temporally aggregated tokens preserves the unified pipeline while injecting reliable temporal structure. Conditioned on the fused representation, the *VAE decoder* outputs the forecast sequence of the next twelve VIL frames. For radar inversion, we directly reuse the pretrained weights of the FLUX.1-schnell general-purpose visual VAE as the continuous latent representation and decoder for the VIL field. This VAE, trained on large-scale natural images, maps 256×256 inputs into a compact latent space and reconstructs them with high fidelity, enabling us to leverage its strong generic visual tokenizer capacity without designing task-specific components. In this work we keep the VAE entirely frozen and only train the unified backbone to map satellite observations into radar latent codes, which already suffices to obtain stable and high-quality radar inversion.

Training Objectives We initialize Omni-Weather from the pretrained Bagel-7B-MoT, which provides a strong multimodal backbone trained on large-scale general data. Building on this foundation, we conduct domain-specific supervised finetuning jointly across all weather tasks.

Formally, let $\tau_t(\cdot)$ be the modality-specific encoder for task t. The model input sequence is defined as

$$X_t \;=\; \big[\,\tau_{\text{text}}(p_t)\,;\;\tau_t(x_t)\,;\;\kappa_t\,\big], \tag{3.2}$$

where κ_t are optional conditioning tokens (e.g., temporal embeddings produced by the nowcasting encoder). The shared backbone produces contextualized tokens

$$\hat{y}_t = \begin{cases} G_\phi(f_\theta(X_t))\,, & t \in \mathcal{T}_{\text{gen}}, \\ L_\psi(f_\theta(X_t))\,, & t \in \mathcal{T}_{\text{under}}. \end{cases} \tag{3.3}$$

Here, $f_\theta(\cdot)$ is the shared encoder/backbone that produces task representations; G_ϕ is the VAE decoder for generation tasks $\mathcal{T}_{\text{gen}}$; L_ψ is the text decoder for understanding tasks $\mathcal{T}_{\text{under}}$. For understanding tasks $t \in \mathcal{T}_{\text{under}}$, we train the text decoder with a standard autoregressive language-modeling objective:

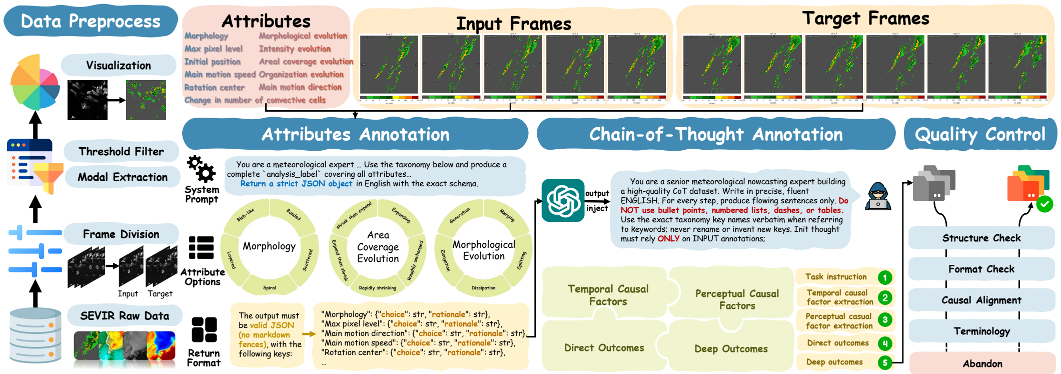
$$\mathcal{L}_{\text{under}} = \sum_{t \in \mathcal{T}_{\text{under}}} \lambda_t \left(-\sum_{k=1}^{n_t} \log p_\psi\big(y_{t,k} \mid y_{t,<k}, f_\theta(X_t)\big) \right), \tag{3.4}$$

where n_t is the target text length and p_ψ denotes the token distribution produced by L_ψ. Here, λ_t is a scalar task-level weight that balances the contribution of different understanding tasks *within* this loss family.

Analogously, for generation tasks $t \in \mathcal{T}_{\text{gen}}$, we adopt a mean-squared reconstruction loss over pixels / frames:

$$\mathcal{L}_{\text{gen}} = \sum_{t \in \mathcal{T}_{\text{gen}}} \lambda_t \, \frac{1}{|\Omega_t|} \, \big\| \hat{y}_t - y_t \big\|_2^2, \tag{3.5}$$

where Ω_t indexes the target for task t. In both cases, λ_t is only a per-task scalar weight and *not* a multiplicative term between $\mathcal{L}_{\text{gen}}$ and $\mathcal{L}_{\text{under}}$; the overall supervised fine-tuning objective is the sum $\mathcal{L} = \mathcal{L}_{\text{gen}} + \mathcal{L}_{\text{under}}$. Since each task loss is already normalized by $|\Omega_t|$ or n_t so that their magnitudes are comparable, we set all $\lambda_t = 1$ in our experiments. This SFT objective unifies generation and understanding under shared representation while preserving task-specific decoders.



You are a meteorological expert ... Use the taxonomy below and produce a complete `analysis_label` covering all attributes... Return a strict JSON object in English with the exact schema.

The output must be valid JSON (no markdown fences), with the following keys:

```
"Morphology": {"choice": str, "rationale": str},
"Max pixel level": {"choice": str, "rationale": str},
"Main motion direction": {"choice": str, "rationale": str},
"Main motion speed": {"choice": str, "rationale": str},
"Rotation center": {"choice": str, "rationale": str},
...
```

You are a senior meteorological nowcasting expert building a high-quality CoT dataset. Write in precise, fluent ENGLISH. For every step, produce flowing sentences only. Do NOT use bullet points, numbered lists, dashes, or tables. Use the exact taxonomy key names verbatim when referring to keywords; never rename or invent new keys. Init thought must rely ONLY on INPUT annotations;

Figure 4: **Construction of our CoT data**. First, we preprocess the raw SEVIR data to obtain high-quality input / output frame pairs. Second, we carefully design prompts and leverage GPT-4o for attributes annotation. Third, the annotated attributes are incorporated into CoT prompts to generate CoT annotations, followed by a quality verification step to produce the final CoT dataset.

3.3 Chain-of-Thought Construction for Weather Generation Reasoning

While unified training enables multi-task learning across generation and understanding, the resulting models still behave as black boxes, lacking explicit reasoning. To enhance interpretability and causal inference, we introduce *Chain-of-Thought (CoT)* supervision as an auxiliary instruction layer for generation tasks. The CoT explicitly captures causal and perceptual factors underlying meteorological evolution, thereby guiding the model toward more structured reasoning about storm dynamics.

Chain-of-Thought for Causal Reasoning in Weather. Our CoT formulation is tailored to the weather domain, where reasoning is framed as causal inference over storm dynamics. To operationalize this, we design a taxonomy of causal elements with expert-defined keywords, refined via GPT-based annotation. The taxonomy is adapted from RadarQA but restructured according to annotation difficulty: *causal factors* (e.g., morphology, intensity, motion) are relatively direct to extract, whereas *outcome indicators* (e.g., storm evolution patterns) require higher-level inference. For nowcasting, causal factors are first derived from the input VIL sequence, then combined with projected causal factors of the forecast frames to infer the more difficult outcome indicators describing future storm behavior. For satellite-to-radar inversion, reasoning involves only causal factors, enabling a direct projection from satellite channels to a single VIL frame. Based on this structure, we construct CoT traces in a three-stage pipeline (Figure 4): attribute annotation with GPT-4o, task-specific reasoning generation with GPT-o3, and automated verification for structural consistency, causal alignment, and terminology normalization. Detailed taxonomy design, reasoning procedures, and prompt template are provided in Appendix A.4.

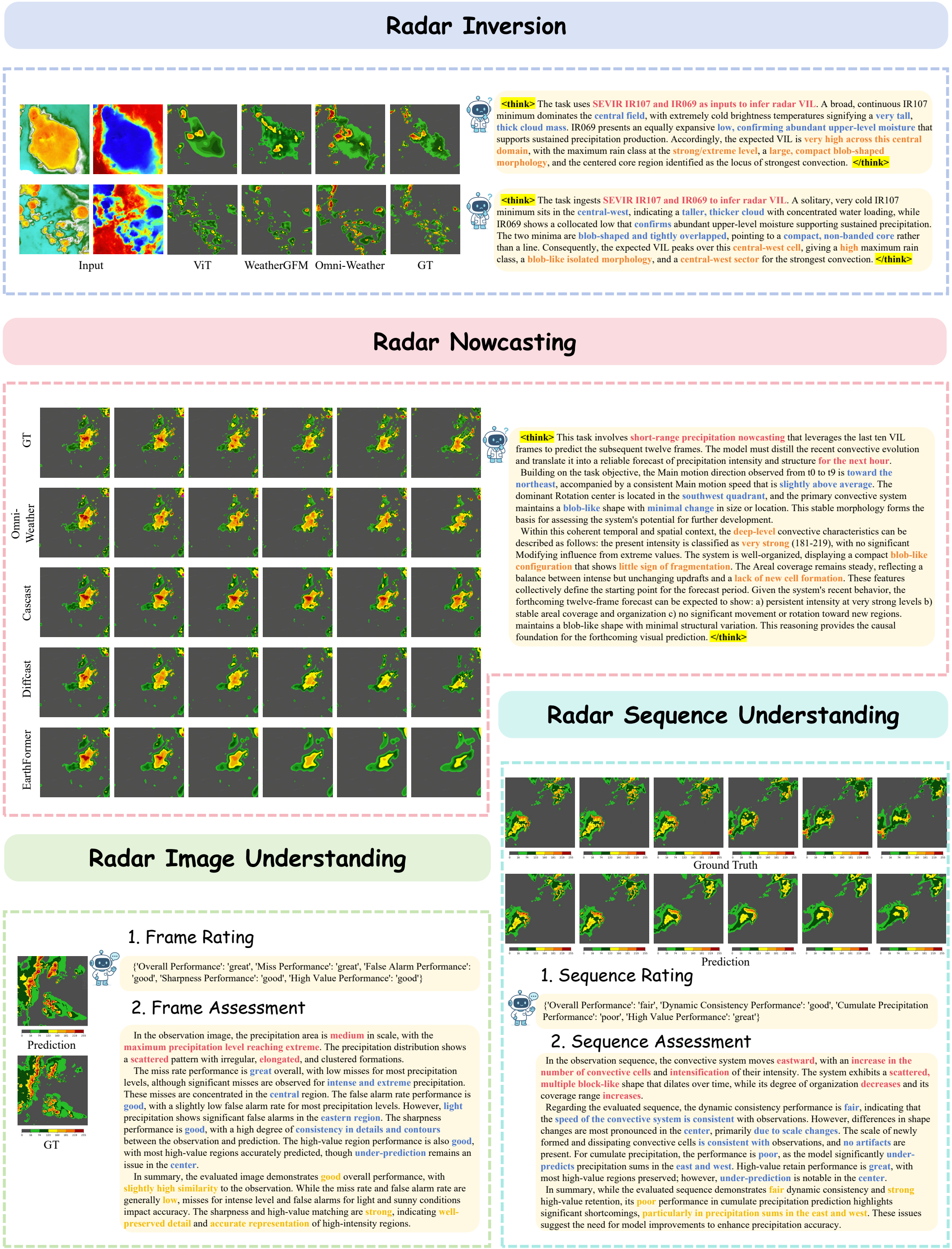
Integration for unified framework. We incorporate CoT reasoning into Omni-Weather from two complementary perspectives. First, during the training phase, CoT serves as auxiliary supervision, requiring the model to generate both intermediate reasoning text and the final prediction, which guides the backbone toward causal interpretability. Second, during inference, CoT is utilized as a reasoning prompt, concatenated with task-specific instructions and inputs to steer the model toward more structured and explainable outputs. This integration allows CoT to not only enhance interpretability but also improve perceptual fidelity and qualitative consistency of generation tasks.

4 Experiments

4.1 Implementation and Evaluation

We train our model end-to-end on a node with 8× H200 GPUs for 20k steps, using packed sequences and the AdamW optimizer Loshchilov & Hutter (2017) with a base learning rate of 2×10^{-4}, weight decay of 0.05, and cosine decay scheduling with a 2k-step warm-up. All images are capped at a resolution of 256×256, resulting in approximately 256 visual tokens per image.

For generation tasks, we report pixel-level metrics (e.g., CSI, CRPS) to evaluate radar accuracy and perceptual metrics (e.g., LPIPS, RadarQA score) to capture structural and semantic consistency. For



<think> The task uses SEVIR IR107 and IR069 as inputs to infer radar VIL. A broad, continuous IR107 minimum dominates the central field, with extremely cold brightness temperatures signifying a very tall, thick cloud mass. IR069 presents an equally expansive low, confirming abundant upper-level moisture that supports sustained precipitation production. Accordingly, the expected VIL is very high across this central domain, with the maximum rain class at the strong/extreme level, a large, compact blob-shaped morphology, and the centered core region identified as the locus of strongest convection. </think>

<think> The task ingests SEVIR IR107 and IR069 to infer radar VIL. A solitary, very cold IR107 minimum sits in the central-west, indicating a taller, thicker cloud with concentrated water loading, while IR069 shows a collocated low that confirms abundant upper-level moisture supporting sustained precipitation. The two minima are blob-shaped and tightly overlapped, pointing to a compact, non-banded core rather than a line. Consequently, the expected VIL peaks over this central-west cell, giving a high maximum rain class, a blob-like isolated morphology, and a central-west sector for the strongest convection. </think>

<think> This task involves short-range precipitation nowcasting that leverages the last ten VIL frames to predict the subsequent twelve frames. The model must distill the recent convective evolution and translate it into a reliable forecast of precipitation intensity and structure for the next hour.

Building on the task objective, the Main motion direction observed from t0 to t9 is toward the northeast, accompanied by a consistent Main motion speed that is slightly above average. The dominant Rotation center is located in the southwest quadrant, and the primary convective system maintains a blob-like shape with minimal change in size or location. This stable morphology forms the basis for assessing the system's potential for further development.

Within this coherent temporal and spatial context, the deep-level convective characteristics can be described as follows: the present intensity is classified as very strong (181-219), with no significant Modifying influence from extreme values. The system is well-organized, displaying a compact blob-like configuration that shows little sign of fragmentation. The Areal coverage remains steady, reflecting a balance between intense but unchanging updrafts and a lack of new cell formation. These features collectively define the starting point for the forecast period. Given the system's recent behavior, the forthcoming twelve-frame forecast can be expected to show: a) persistent intensity at very strong levels b) stable areal coverage and organization c) no significant movement or rotation toward new regions. maintains a blob-like shape with minimal structural variation. This reasoning provides the causal foundation for the forthcoming visual prediction. </think>

{'Overall Performance': 'fair', 'Dynamic Consistency Performance': 'good', 'Cumulate Precipitation Performance': 'poor', 'High Value Performance': 'great'}

In the observation sequence, the convective system moves eastward, with an increase in the number of convective cells and intensification of their intensity. The system exhibits a scattered, multiple block-like shape that dilates over time, while its degree of organization decreases and its coverage range increases.

Regarding the evaluated sequence, the dynamic consistency performance is fair, indicating that the speed of the convective system is consistent with observations. However, differences in shape changes are most pronounced in the center, primarily due to scale changes. The scale of newly formed and dissipating convective cells is consistent with observations, and no artifacts are present. For cumulate precipitation, the performance is poor, as the model significantly under-predicts precipitation sums in the east and west. High-value retain performance is great, with most high-value regions preserved; however, under-prediction is notable in the center.

In summary, while the evaluated sequence demonstrates fair dynamic consistency and strong high-value retention, its poor performance in cumulate precipitation prediction highlights significant shortcomings, particularly in precipitation sums in the east and west. These issues suggest the need for model improvements to enhance precipitation accuracy.

{'Overall Performance': 'great', 'Miss Performance': 'great', 'False Alarm Performance': 'good', 'Sharpness Performance': 'good', 'High Value Performance': 'good'}

In the observation image, the precipitation area is medium in scale, with the maximum precipitation level reaching extreme. The precipitation distribution shows a scattered pattern with irregular, elongated, and clustered formations.

The miss rate performance is great overall, with low misses for most precipitation levels, although significant misses are observed for intense and extreme precipitation. These misses are concentrated in the central region. The false alarm rate performance is good, with a slightly low false alarm rate for most precipitation levels. However, light precipitation shows significant false alarms in the eastern region. The sharpness performance is good, with a high degree of consistency in details and contours between the observation and prediction. The high-value region performance is also good, with most high-value regions accurately predicted, though under-prediction remains an issue in the center.

In summary, the evaluated image demonstrates good overall performance, with slightly high similarity to the observation. While the miss rate and false alarm rate are generally low, misses for intense level and false alarms for light and sunny conditions impact accuracy. The sharpness and high-value matching are strong, indicating well-preserved detail and accurate representation of high-intensity regions.

Figure 5: **A set of qualitative results.** We show two radar inversion examples with think traces, a nowcasting case where Omni-Weather (with think trace) is compared against CasCast, DiffCast, and EarthFormer, and one example each of radar image and sequence understanding with attribute scores and textual evaluations. Omni-Weather surpasses all baselines.

understanding tasks, evaluation follows RadarQA protocols, considering both prediction–reference alignment and LLM-based external judgments. To ensure fair comparison, we benchmark against the strongest available models: CasCast, DiffCast, and EarthFormer for nowcasting; WeatherGFM, UNet, and ViT for satellite-to-radar inversion; and GPT4-Score as well as the domain-specialized RadarQA for understanding. Full details of metrics and baselines are provided in Appendix A.3.

Table 1: **Quantitative results on Weather Generation and Weather Understanding tasks.**The best results are highlighted in bold, and the second-best results are underscored. Abbreviations: CSI-M - CSI-Mean, R.S - Radar Score, CSI-P4 - CSI-Pool4, CSI-P16 - CSI-Pool16, C-16 - CSI@16, C-74 - CSI@74, C-160 - CSI@160, C-181 - CSI@181, C-219 - CSI@219, Dyn. - Dynamic Consistency, Cum. - Cumulate Precipitation, H. Val. - High. Value, R._L - Rouge_L, B.S - BertScore, Sharp. - Sharpness. Metrics marked with ↓ denote lower-is-better objectives, whereas metrics without such notation should be interpreted as higher-is-better.

Weather Generation

Method	Radar Nowcasting							Method	Radar Inversion						
	CSI-M	R.S	CSI-P4	CSI-P16	CRPS ↓	SSIM	LPIPS ↓		R.S	RMSE ↓	C-16	C-74	C-160	C-181	C-219
Earthformer	**0.389**	1.92	0.401	0.387	0.037	0.729	0.322	UNet	1.75	0.821	0.222	0.370	0.180	0.153	0.079
Diffcast	0.375	2.43	0.407	0.511	0.033	0.739	0.235	ViT	2.01	0.445	0.602	0.436	0.180	0.131	0.042
Cascast	0.384	2.72	0.414	0.518	0.031	0.746	0.207	WeatherGFM	2.28	**0.436**	0.619	0.447	0.208	0.157	0.053
Omni-Weather	0.384	2.69	**0.427**	0.539	**0.026**	0.746	0.179	-	2.42	0.514	**0.622**	0.469	0.263	0.221	0.118
Omni-Weather-thinking	0.353	**2.86**	0.421	**0.542**	0.028	**0.751**	**0.166**	-	**2.51**	0.507	0.621	**0.473**	**0.277**	**0.230**	**0.129**

Weather Understanding

Method	Radar Sequence Understanding							Radar Image Understanding							
	Overall	Dyn.	Cum.	H. Val.	R._L	B.S	GPT4	Overall	Miss	FAR	H. Val.	Sharp.	R._L	B.S	GPT4
Claude-sonnet-4	20.79	20.79	20.79	21.78	0.287	0.745	5.73	32.79	32.56	34.19	24.77	46.05	0.368	0.743	5.18
Gemini-2.5-pro	27.59	28.34	26.72	22.47	0.254	0.739	5.77	21.40	31.16	29.65	29.30	40.58	0.348	0.741	5.63
GPT-5	49.50	36.63	35.64	30.69	0.213	0.690	6.85	56.05	21.74	32.79	40.81	48.49	0.297	0.702	6.31
RadarQA	**66.17**	53.31	**48.94**	**80.52**	0.436	**0.815**	6.87	61.51	67.67	65.35	69.19	78.60	0.512	**0.809**	**6.58**
Omni-Weather	61.79	**64.05**	45.19	67.29	**0.446**	0.810	**7.48**	**64.30**	**92.21**	**88.72**	**91.4**	**91.74**	**0.543**	0.760	6.03

4.2 Experimental Results

Currently, there exists no unified model capable of simultaneously handling both weather generation and weather understanding tasks. Existing approaches are typically specialized, such as Cascast or DiffCast for forecasting, or understanding-only models such as RadarQA for evaluation. In contrast, Omni-Weather is designed as a single framework to integrate generation and understanding. While aiming for strong quantitative performance across generation and understanding tasks, our experiments further investigate how a unified framework supports mutual gains between these tasks, reveals trade-offs between perceptual reasoning and pixel-level accuracy, and leverages both scientific and general-domain data for improved learning.

Omni-Weather achieves superior performance in weather generation. As shown in Table 1, Omni-Weather improves both deterministic accuracy and perceptual quality in nowcasting. Compared with single-task baselines, our model reduces CRPS by over 15% and improves LPIPS by more than 25%, while maintaining similar CSI and SSIM. On the radar inversion task, Omni-Weather consistently surpasses both specialized (i.e., WeatherGFM) and generalist (i.e., UNet and ViT) models, achieving higher CSI scores across all thresholds, with gains up to 20% at high-value levels. Furthermore, when augmented with *thinking* inference, Omni-Weather achieves clear improvements in perceptual quality, LPIPS decreases by nearly 10% while showing minor reductions on pixel-level metrics such as CSI-Mean. This highlights that explicit reasoning can enhance visual fidelity and interpretability with limited cost to deterministic accuracy.

Omni-Weather delivers strong results in weather understanding. Table 1 also presents results for weather understanding tasks. Closed-source LLMs fail to adapt to this task, often achieving accuracies below 30%. While RadarQA serves as a competitive benchmark, Omni-Weather surpasses it: on radar image understanding, accuracy on key attributes (e.g., *Miss*, *False Alarm*) exceeds RadarQA by 20–25 points, and on radar sequence understanding, *Dynamic Consistency* improves by over 10 points with a 5% overall gain. These results highlight that Omni-Weather attains strong capability in understanding both weather sequences and single frames data.

Table 2: Training only Understanding (U), only Generation (G), or Joint (U+G). Frame / Sequence tasks are evaluated in accuracy, GPT4-Score for understanding, CSI-M, RMSE for generation.

Setting	Understanding		Generation	
	Accuracy ↑	GPT4-score ↑	CSI-M ↑	RMSE ↓
Und-only	81.95 / 54.34	5.78 / **6.03**	-	-
Gen-only	-	-	0.303 / 0.323	0.590 / 19.01
Joint (U+G)	**86.65 / 59.58**	**7.48 / 6.03**	**0.338 / 0.347**	**0.514 / 17.11**

Table 3: **Effect of CoT finetuning and thinking inference.** ↑ higher is better, ↓ lower is better. Training both with CoT finetuning and thinking inference achieves the best result in most metrics. Abbreviation: R.S - Radar-Score.

CoT FT	Think Inf.	CSI-M ↑	CRPS ↓	R.S ↑	LPIPS ↓	GPT4-Score ↑
✓	✗	**0.347**	**0.023**	2.423	0.182	-
✗	✓	0.237	0.042	2.032	0.213	4.21
✓	✓	0.335	**0.023**	**2.657**	**0.163**	**7.82**

Omni-Weather demonstrates versatile qualitative performance across tasks. Figure 5 illustrates Omni-Weather's outputs across both generation and understanding tasks. In the radar inversion task, Omni-Weather generates VIL fields with richer high-value structures and reasoning traces linking satellite cues to radar responses. In the radar nowcasting task, forecasts exhibit fine-grained storm details with improved spatial reflected by higher Radar-Score, and the think traces offer interpretable accounts of storm evolution. For radar understanding tasks, Omni-Weather delivers expert-like outputs, combining attribute-level ratings with detailed textual evaluations that provide domain-specific insights.

4.3 Ablation Studies and Analysis

Impact of joint training on generation and understanding. To examine whether generation and understanding tasks can benefit each other, we performed supervised fine-tuning on Bagel-7B-MoT using only understanding data (U-only), only generation data (G-only), or both jointly (U+G), and evaluated on 200 randomly sampled validation examples per task. As shown in Table 2, joint training improves performance across both understanding and generation: understanding achieves higher overall scores and better consistency, while generation gains in both accuracy and perceptual quality. These results indicate that unified training enables the model to both learn and interpret weather data more effectively, with generation and understanding tasks mutually enhancing each other.

Effectiveness of Mixed Scientific and General Data. We conducted an additional experiment by finetuning Bagel-7B-MoT on SEVIR alone versus SEVIR combined with 20k samples from the general metaquery Pan et al. (2025) dataset. As shown in Figure 6, the inclusion of general data consistently improves model performance, particularly in deterministic metrics and perceptual quality indicators. These findings suggest that while scientific data anchors domain-specific fidelity, general data provides auxiliary coverage of diverse patterns, enabling the model to learn more robust cross-modal information for better representations.

Perceptual Trade-offs in Reasoning. We investigate the impact of CoT-annotated supervision and reasoning-based inference on radar nowcasting, evaluating 200 carefully sampled test cases. As shown in Table 3, reasoning introduces a clear trade-off: richer prompts yield more detailed generations and improvements in perceptual image metrics (e.g., LPIPS, Radar-Score), while pixel-level measures such as CSI moderately decline. Beyond images, we also evaluate the generated textual explanations with a GPT4-Score, which assigns higher scores to CoT-enhanced outputs, confirming gains in interpretability. A case study in Figure 7 further illustrates this effect, where reasoning produces sharper storm structures and more coherent temporal evolution despite lower CSI, suggesting that it prioritizes semantic and structural fidelity over pixel-wise alignment. Qualitative comparisons of reasoning content are provided in Appendix A.5.

5 Conclusion

We introduce Omni-Weather, a unified foundation model for weather generation and understanding. Through a shared backbone, it supports both generation and understanding tasks. This design allows the model to surpass task-specific baselines and enables reasoning analysis. The reasoning ability further enhances interpretability and improves the perceptual quality of radar sequences, highlighting the potential of Omni-Weather as a generalist foundation model for future weather applications.

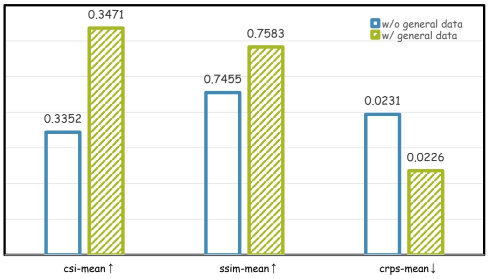


Figure 6: **Effect of mixed data.**

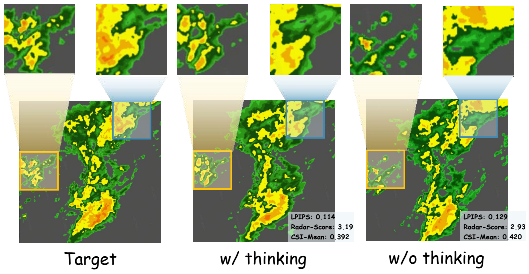


Figure 7: **Case study with thinking.**

Limitations. First, Omni-Weather cannot yet adapt to general-domain VAEs. Second, broader validation across diverse weather tasks, such as medium-range forecasting and typhoon track prediction, remains necessary. Addressing these limitations will be crucial for advancing foundation models toward more robust and universally applicable weather intelligence.

Ethics and Reproducibility Statement

This work develops a unified model for weather generation and understanding, aiming to improve interpretability and perceptual quality in meteorological forecasting. The datasets used (e.g., SEVIR, RadarQA) are publicly available, and all experiments follow their licenses without involving sensitive information. We acknowledge that unified models may be misused without expert oversight; thus, our method is intended for research purposes and should be complemented by professional interpretation. To support reproducibility, we provide detailed descriptions of tasks, model design, training objectives, and evaluation protocols, and we release a Chain-of-Thought dataset for causal reasoning at `https://anonymous.4open.science/r/cot-data-E3D4/` along with its construction code. Upon acceptance, we will further release the complete codebase, model checkpoints, and data-processing scripts to ensure transparency and verifiability.

References

Kaifeng Bi, Lingxi Xie, Hengheng Zhang, Xin Chen, Xiaotao Gu, and Qi Tian. Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast. *arXiv preprint arXiv:2211.02556*, 2022.

Cristian Bodnar, Wessel P Bruinsma, Ana Lucic, Megan Stanley, Anna Allen, Johannes Brandstetter, Patrick Garvan, Maik Riechert, Jonathan A Weyn, Haiyu Dong, et al. A foundation model for the earth system. *Nature*, pp. 1–8, 2025.

Jiuhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyi Zhou, Saining Xie, Silvio Savarese, et al. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. *arXiv preprint arXiv:2505.09568*, 2025a.

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.

Xiangyu Chen, Zheyuan Li, Yuandong Pu, Yihao Liu, Jiantao Zhou, Yu Qiao, and Chao Dong. A comparative study of image restoration networks for general backbone network design. In *European Conference on Computer Vision*, pp. 74–91. Springer, 2024a.

Xiangyu Chen, Yihao Liu, Yuandong Pu, Wenlong Zhang, Jiantao Zhou, Yu Qiao, and Chao Dong. Learning a low-level vision generalist via visual task prompt. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 2671–2680, 2024b.

Xiangyu Chen, Kaiwen Zhu, Yuandong Pu, Shuo Cao, Xiaohui Li, Wenlong Zhang, Yihao Liu, Yu Qiao, Jiantao Zhou, and Chao Dong. Exploring scalable unified modeling for general low-level vision. *arXiv preprint arXiv:2507.14801*, 2025b.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024c.

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

Zhihan Gao, Xingjian Shi, Hao Wang, Yi Zhu, Yuyang Bernie Wang, Mu Li, and Dit-Yan Yeung. Earthformer: Exploring space-time transformers for earth system forecasting. *Advances in Neural Information Processing Systems*, 35:25390–25403, 2022.

Zhihan Gao, Xingjian Shi, Boran Han, Hao Wang, Xiaoyong Jin, Danielle Maddix, Yi Zhu, Mu Li, and Yuyang Bernie Wang. Prediff: Precipitation nowcasting with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:78621–78656, 2023.

Junchao Gong, Lei Bai, Peng Ye, Wanghan Xu, Na Liu, Jianhua Dai, Xiaokang Yang, and Wanli Ouyang. Cascast: Skillful high-resolution precipitation nowcasting via cascaded modelling. *arXiv preprint arXiv:2402.04290*, 2024.

Tao Han, Song Guo, Fenghua Ling, Kang Chen, Junchao Gong, Jingjia Luo, Junxia Gu, Kan Dai, Wanli Ouyang, and Lei Bai. Fengwu-ghr: Learning the kilometer-scale medium-range global weather forecasting. *arXiv preprint arXiv:2402.00059*, 2024.

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. *arXiv preprint arXiv:2508.12291*, 2025a.

Xuming He, Zhiwang Zhou, Wenlong Zhang, Xiangyu Zhao, Hao Chen, Shiqi Chen, and Lei Bai. Diffsr: Learning radar reflectivity synthesis via diffusion model from satellite observations. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025b.

Ajay Krishnan and VS Anoop. Climatenlp: Analyzing public sentiment towards climate change using natural language processing. *arXiv preprint arXiv:2310.08099*, 2023.

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.

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.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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.

Tung Nguyen, Johannes Brandstetter, Ashish Kapoor, Jayesh K Gupta, and Aditya Grover. Climax: A foundation model for weather and climate. *arXiv preprint arXiv:2301.10343*, 2023.

Tung Nguyen, Rohan Shah, Hritik Bansal, Troy Arcomano, Romit Maulik, Rao Kotamarthi, Ian Foster, Sandeep Madireddy, and Aditya Grover. Scaling transformer neural networks for skillful and reliable medium-range weather forecasting. *Advances in Neural Information Processing Systems*, 37:68740–68771, 2024.

Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Jiuhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with metaqueries. *arXiv preprint arXiv:2504.06256*, 2025.

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. Gencast: Diffusion-based ensemble forecasting for medium-range weather. *arXiv preprint arXiv:2312.15796*, 2023.

Yuandong Pu, Le Zhuo, Kaiwen Zhu, Liangbin Xie, Wenlong Zhang, Xiangyu Chen, Peng Gao, Yu Qiao, Chao Dong, and Yihao Liu. Lumina-omnilv: A unified multimodal framework for general low-level vision. *arXiv preprint arXiv:2504.04903*, 2025.

Tobias Schimanski, Julia Bingler, Camilla Hyslop, Mathias Kraus, and Markus Leippold. Climatebert-netzero: Detecting and assessing net zero and reduction targets. *arXiv preprint arXiv:2310.08096*, 2023.

Johannes Schmude, Sujit Roy, Will Trojak, Johannes Jakubik, Daniel Salles Civitarese, Shraddha Singh, Julian Kuehnert, Kumar Ankur, Aman Gupta, Christopher E Phillips, et al. Prithvi wxc: Foundation model for weather and climate. *arXiv preprint arXiv:2409.13598*, 2024.

Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. Lmfusion: Adapting pretrained language models for multimodal generation. *arXiv preprint arXiv:2412.15188*, 2024.

Rui Tian, Mingfei Gao, Mingze Xu, Jiaming Hu, Jiasen Lu, Zuxuan Wu, Yinfei Yang, and Afshin Dehghan. Unigen: Enhanced training & test-time strategies for unified multimodal understanding and generation. *arXiv preprint arXiv:2505.14682*, 2025.

Shengbang Tong, David Fan, Jiachen Zhu, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. *arXiv preprint arXiv:2412.14164*, 2024.

Sumanth Varambally, Veeramakali Vignesh Manivannan, Yasaman Jafari, Luyu Han, Zachary Novack, Zhirui Xia, Salva Rühling Cachay, Srikar Eranky, Ruijia Niu, Taylor Berg-Kirkpatrick, et al. Aquilon: Towards building multimodal weather llms. In *ICML 2025 Workshop on Assessing World Models*.

Mark Veillette, Siddharth Samsi, and Chris Mattioli. Sevir: A storm event imagery dataset for deep learning applications in radar and satellite meteorology. *Advances in Neural Information Processing Systems*, 33:22009–22019, 2020.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. Climatebert: A pretrained language model for climate-related text. *arXiv preprint arXiv:2110.12010*, 2021.

Walter F Wiggins and Ali S Tejani. On the opportunities and risks of foundation models for natural language processing in radiology. *Radiology: Artificial Intelligence*, 4(4):e220119, 2022.

Demin Yu, Xutao Li, Yunming Ye, Baoquan Zhang, Chuyao Luo, Kuai Dai, Rui Wang, and Xunlai Chen. Diffcast: A unified framework via residual diffusion for precipitation nowcasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27758–27767, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

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

Xiangyu Zhao, Zhiwang Zhou, Wenlong Zhang, Yihao Liu, Xiangyu Chen, Junchao Gong, Hao Chen, Ben Fei, Shiqi Chen, Wanli Ouyang, et al. Weathergfm: Learning a weather generalist foundation model via in-context learning. *arXiv preprint arXiv:2411.05420*, 2024.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. 2024. URL `https://api.semanticscholar.org/CorpusID:271909855`.

A APPENDIX

A.1 OVERVIEW

This Appendix is structured as follows:

- Sec. A.2: Details of task paradigm and dataset.
- Sec. A.3: Details of evaluation metrics.
- Sec. A.4: Data construction of Chain-of-thought dataset.
- Sec. A.5 : Details about Chain-of-though reasoning.
- Sec. A.6: More experimental results of Omni-Weather.
- Sec. A.7: Qualitative results of Omni-Weather.
- Sec. A.8: Usage of LLM.

A.2 MORE DETAILS ABOUT DATASETS AND TASK

A.2.1 DATA DETAIL

SEVIR. The Storm EVent ImagRy (SEVIR) dataset is a large-scale collection of temporally aligned weather observations covering the continental United States. It integrates multiple sensing modalities, including visible and infrared satellite imagery, lightning event records, and mosaics of Vertically Integrated Liquid (VIL) derived from NEXRAD radar. In this work, we focus on the radar-based VIL product, which provides a spatio-temporal representation of convective storm structures. SEVIR contains over 20,000 storm events sampled between 2017 and 2020, each spanning a 4-hour window at 5-minute resolution and covering approximately $384\,\text{km} \times 384\,\text{km}$ regions. To support short-range forecasting tasks, sequences are typically arranged as input–output pairs, where a set of observed frames is used to predict future VIL evolution. The images are normalized to the range $[0, 255]$ and evaluated against threshold-based metrics (e.g., CSI, HSS) following established protocols. This combination of multi-sensor coverage, temporal alignment, and standardized evaluation makes SEVIR a widely adopted benchmark for data-driven weather prediction.

RadarQA Dataset. RQA-70K from RadarQA is a large-scale forecast quality analysis dataset encompassing four tasks: frame rating, frame assessment, sequence rating, and sequence assessment. RQA-70K is constructed through a combination of human annotation and automated labeling. By integrating traditional forecasting metrics with expert knowledge, the dataset provides a comprehensive benchmark for the assessment of weather radar forecasting.

A.2.2 TASK DETAIL

Weather Generation. For radar nowcasting, the model is trained to predict the short-term spatio-temporal evolution of precipitation. Specifically, we use sequences of 10 observed VIL frames as input and require the model to generate the subsequent 12 frames. All frames are preprocessed to a spatial resolution of 256×256, which balances coverage of mesoscale convective features with computational efficiency. This setting follows standard short-range nowcasting protocols but is tailored to emphasize fine-scale storm structures, ensuring that the model learns both spatial coherence and temporal continuity in convective system development.

Weather Understanding. Weather understanding tasks require the model to produce natural-language descriptions and evaluations based on radar observations or forecast sequences. The input can be either a single VIL frame or a sequence of frames, and the output is a structured report covering key meteorological aspects such as storm morphology, intensity, temporal evolution, and forecast quality (e.g., hits, misses, or false alarms). Following the RadarQA benchmark, model responses are rated along multiple dimensions using a four-level ordinal scale (*fair*, *poor*, *good*, *great*), which are mapped to numerical values 1–4. Scores are then averaged across dimensions to obtain the *Radar Score*, providing a comprehensive indicator of diagnostic quality. This task formulation bridges language modeling with domain-specific evaluation, enabling models not only to assess physical forecasts but also to generate expert-like reasoning aligned with meteorological practice.

A.3 Evaluation Protocols and Metrics

To enable a more comprehensive and accurate assessment, we employ a diverse set of evaluation metrics across different tasks. The detailed definitions of these metrics are provided below.

CSI. Critical Success Index (CSI) is widely used in the evaluation for weather forecasting tasks. Formally, it is defined as:

$$CSI = \frac{TP}{TP + FN + FP} \tag{A.1}$$

where TP, FP, and FN denote the number of true positives, false positives, and false negatives, respectively. Following Cascast Gong et al. (2024), we apply thresholds at 16, 74, 133, 160, 181, and 219 to evaluate model performance across different VIL intensity ranges.

SSIM Wang et al. (2004). Structural Similarity Index Measure (SSIM) is a perceptual metric that quantifies the similarity between two images by comparing their contrast, luminance, and structure. Formally, it is defined as:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{A.2}$$

where μ_x and μ_y are the means of x and y, σ_x^2 and σ_y^2 are the variances, σ_{xy} is the covariance, and C_1, C_2 are small constants to stabilize the division. Higher SSIM values indicate stronger similarity between the prediction and observation.

CRPS. Continuous Ranked Probability Score (CRPS) evaluates the accuracy of probabilistic forecasts by comparing Cumulative Distribution Function (CDF) with observation x. Formally, it is defined as:

$$CRPS(F, x) = \int_{-\infty}^{+\infty} \big(F(y) - \mathbf{1}\{y \geq x\}\big)^2 dy \tag{A.3}$$

where $\mathbf{1}\{y \geq x\}$ is the indicator function.

LPIPS Zhang et al. (2018). Learned Perceptual Image Patch Similarity (LPIPS) is a perceptual metric designed to evaluate similarity between images in a manner aligned with human judgment. LPIPS leverages neural networks to compute differences in deep feature-based representations.

Rouge_L Lin (2004). Rouge_L is widely used for evaluating the quality of generated text by measuring the Longest Common Subsequence (LCS) between a candidate and reference sequence. Rouge_L accounts for sentence-level similarity, capturing both content and fluency.

BertScore Zhang et al. (2019). BertScore is a learned mertic for evaluating text generation that leverages text embeddings from pretrained language models (e.g., Bert) to compute similarity between candidate and reference sentences.

GPT4-Score. GPT4-Score leverages the reasoning and understanding capabilities of GPT4 to assess generated outputs. Specifically, both the ground truth and the prediction are provided to GPT-4, which evaluates the prediction based on overall accuracy, content richness, and fidelity to the reference.

Radar-Score.RadarQA formulates radar understanding as a rating task, where model-generated diagnostic reports are assessed along multiple meteorologically relevant dimensions, such as storm morphology, intensity, temporal evolution, and forecast quality. Each dimension is rated on a four-level ordinal scale {fair, poor, good, great}, which are mapped to numerical values 1 to 4. The *Radar Score* is then obtained by averaging these ratings across all evaluated dimensions, producing a single interpretable measure that reflects the overall diagnostic quality of the model's output.

A.4 Detail of Chain-of-Thought Construction

To enable causal reasoning over storm dynamics in the weather domain, we carefully design a CoT data construction pipeline to generate high-quality CoT data.

Data Preprocess. First, we extract the raw data from the SEVIR dataset and segment each event into three pairs of 10-frame inputs and 12-frame outputs. Second, we filter out samples with limited informative content and visualize the retained frames using SEVIR's colormap.

Attributes Annotation. For the constructed input / output frame pairs, we perform attributes annotation by designing a structured prompt, which encompasses four components: system prompt, attribute options, return format, and caution instructions. The prompt is then provided to a large language model to generate structured JSON outputs that serve as inputs for the subsequent CoT annotation process.

System Prompt for Attribute Annotation

You are a meteorological expert. You will be given an animated GIF containing the 10/12 LABEL frames to generate the final conclusions. Use the taxonomy below and produce a complete "analysis_label" covering all attributes, consistent with the LABEL sequence. Use only the exact option strings provided for each attribute; if an attribute is not apparent from the LABEL frames, set "choice" to "not apparent". In "rationale", use probabilistic language when evidence is weak (e.g., "likely", "possible"). Do not reference or speculate about INPUT frames.
Orientation & mapping (do not ignore):
Images are rendered with ‘origin=’upper’‘. Treat the top of the image as North (N), the bottom as South (S), the left as West (W), and the right as East (E). - Up = North, Down = South, Left = West, Right = East. - Diagonals: Up-Left = Northwest, Up-Right = Northeast, Down-Left = Southwest, Down-Right = Southeast. Do not rotate or flip the images. Determine ‘Main motion direction‘ strictly using this mapping.
Definitions (used consistently throughout):
- Convective cell: a contiguous region whose pixel value exceeds a reflectivity/intensity threshold (>32 on a 0–255 scale; colorbar green and above). Very small isolated speckles can be ignored when they are unlikely to influence scene-level judgment. - Main convective system: the dominant connected system within the scene (the most spatially prominent/organized aggregate of cells). When multiple candidates exist, use the overall/aggregate pattern that best represents the sequence. Taxonomy (attributes and allowed options):{enum_text} Return a strict JSON object in English with the exact schema:{schema_text}

Attribute Options

```
"Morphology": (
    "Spatial organization pattern at the start window (or across the input window); "
    "scattered = many small isolated cells distributed broadly (low organization); "
    "banded = elongated/narrow rainband aligned along a main axis; "
    "blob-like = one or a few dominant compact clusters with larger areal extent; "
    "spiral = curved/spiraling rainbands (e.g., tropical cyclones/mesoscale vortices); "
    "layered = sheet-like, wide coverage with uniform texture and less distinct boundaries; "
    "bow-shaped = an arced/bowing reflectivity segment (e.g., bow echo)."
),
"Max pixel level": (
    "Peak intensity represented by the maximum pixel value across the scene (0–255). "
    "Use the provided bins: no significant (0–31), very weak (31–73), weak (74–132), "
    "moderate (133–159), strong (160–180), very strong (181–218), extreme (219–255)."
    " Pixel intensity values map directly to the provided colorbar."
),
"Initial position of the main convective system": (
    "Coarse location of the dominant connected system at t0 in image coordinates; "
    "Centered / N / S / W / E / four quadrants; or ’no clear main system’ if dominance is
ambiguous."
),
"Main motion direction": (
    "Dominant displacement of the MAIN system (aggregate of cells above threshold >32)
across frames. "
    "If multiple clusters move differently or displacement is unclear, choose ’no obvious
motion’."
),
"Main motion speed": (
    "Qualitative speed class based on normalized displacement per frame relative to image size: "
    "near-stationary / slow / moderate / fast / very fast."
),
"Rotation center": (
    "Approximate sector of the rotation center if clear cyclonic/anticyclonic rotation is present (N/NE/E/SE/S/SW/W/NW); "
    "otherwise 'no rotation' or 'location uncertain'."
),
"Change in number of convective cells": (
    "Trend in the count of distinct convective cells. "
    "Ignore tiny isolated speckles when appropriate."
),
"Morphological evolution of the main system": (
    "Primary structural change of the MAIN system over time: elongation, shrinkage, expansion, "
    "merging, splitting, dissipation, generation."
),
"Intensity evolution": (
    "Overall trend of reflectivity/brightness of the MAIN system: strengthening / weakening / roughly unchanged."
),
"Areal coverage evolution": (
    "Trend in the areal extent of above-threshold pixels (>32): "
    "expanding / roughly unchanged / rapidly shrinking / expand then shrink / shrink then gradually expand."
),
"Organization evolution": (
    "Change in connectedness/ordering of the system: becoming connected (fragmented → more coherent), becoming fragmented (coherent → more broken), "
    "connected then gradually weakening (coherent but loosening/fading), fragmented then becoming connected (consolidation), no obvious change."
),
```

Return Format for Attribute Annotation

The output must be valid JSON (no markdown fences), with the following keys:

```
{
    "sample_id": str,
    "analysis_input": {
        "annotations": {
            "Morphology": {"choice": str, "rationale": str},
            "Max pixel level": {"choice": str, "rationale": str},
            "Initial position of the main convective system": {"choice": str, "rationale": str},
            "Main motion direction": {"choice": str, "rationale": str},
            "Main motion speed": {"choice": str, "rationale": str},
            "Rotation center": {"choice": str, "rationale": str},
            "Change in number of convective cells": {"choice": str, "rationale": str},
            "Morphological evolution of the main system": {"choice": str, "rationale": str},
            "Intensity evolution": {"choice": str, "rationale": str},
            "Areal coverage evolution": {"choice": str, "rationale": str},
            "Organization evolution": {"choice": str, "rationale": str}
        },
        "global_rationale": str
    }
```

}

Additional Rules for Attribute Annotation

Additional rules:

- Consider only INPUT frames (t0..t9); do not mention LABEL frames.
- Cover all attributes above.
- For each attribute, ‘choice‘ must be one of the listed options verbatim, or the literal string "not apparent" when evidence is insufficient.
- Keep explanations concise, using terms like "likely" / "possible" when appropriate.
- Orientation & mapping (do not ignore): Images are rendered with ‘origin=’upper’‘. Treat the top of the image as North (N), the bottom as South (S), the left as West (W), and the right as East (E). For diagonals: up-left=Northwest, up-right=Northeast, down-left=Southwest, down-right=Southeast. Do not rotate or flip the images.
- Consistency check for "Main motion direction": Verify that the apparent displacement on the image (screen coordinates) matches the above mapping before choosing. If uncertain, choose "no obvious motion".
- Definition of "Main motion direction": Determine motion using the contiguous regions where pixel value > 31 (i.e., colorbar green and above). Track the overall trajectory of these regions from t0 to t9 (or t10 to t11 for LABEL). If multiple such regions exist, report their general/aggregate motion direction. If regions are too fragmented or exhibit multiple inconsistent directions, set the direction to "no obvious motion".

CoT Annotation. After obtaining the annotated attribute data, the large language model is prompted to generate outputs in a predefined sequence: task instruction, Temporal causal factor, perceptual causal factor, direct outcomes, and deep outcomes. The system prompt, cautions, return format, and instruction for each step are detailed as follows.

System Prompt for CoT Annotation

You are a senior meteorological nowcasting expert building a high-quality CoT dataset. Write in precise, fluent ENGLISH. For every step, produce flowing sentences only. Do NOT use bullet points, numbered lists, dashes, or tables. Use the exact taxonomy key names verbatim when referring to keywords; never rename or invent new keys. Init thought must rely ONLY on INPUT annotations; Summary thought must rely ONLY on LABEL annotations. Keep the reasoning scientific and case-specific.

Instructions and Examples for each Step

INIT THOUGHT (INPUT-based; do not use LABEL):
Step 1 — Task instruction:
Explain in 2–3 sentences that this is a short-range precipitation nowcasting task based on the last 10 VIL frames, and the goal is to forecast the next 12 frames.
Step 2 — Temporal causal factor extraction (merged with temporal description):
Write a single descriptive paragraph that uses ONLY the temporal causal factor keywords. Explicitly restate the choices for "Main motion direction", "Main motion speed", and "Rotation center" as observed from t0–t9, and keep the prose strictly descriptive. Do NOT include meta-summaries such as “this establishes a premise”, avoid invented terminology, mechanisms, or evaluative language, and do not reference perceptual or outcome keywords. Use present/immediate past tense and clear compass terms.
Example (style only): "Across t0–t9, ’Main motion direction’ is east, ’Main motion speed’ is slow, and ’Rotation center’ is no rotation, so the motion is a slow eastward translation without rotational signatures."
Step 3 — Perceptual causal factor extraction (merged with spatial description):
Write a single descriptive paragraph that uses ONLY the perceptual causal factor keywords. Explicitly restate the choices for "Morphology", "Max pixel level", and "Initial position of

the main convective system" as observed from t0–t9, and keep the prose strictly descriptive. Do NOT include meta-summaries such as “this defines a baseline”, avoid mechanisms or evaluative language, and do not reference temporal or outcome keywords. Use the present/immediate past tense.
Example (style only): "The ’Morphology’ is blob-like, the ’Max pixel level’ is strong, and the ’Initial position of the main convective system’ is centered, yielding compact high-intensity echoes near the center across t0–t9."
Step 4 — Direct outcomes (1st-Tier Outcome):
Analyze ONLY the direct outcome keyword "Intensity evolution". Write exactly ONE concise sentence that delivers the conclusion for "Intensity evolution", justified solely by the temporal and perceptual paragraphs (Steps 2–3). The sentence must explicitly name "Intensity evolution". Do NOT reference deep outcomes, mechanisms, grids, coordinates, or any terms not present in the keyword taxonomy. Avoid generic phrases like "structural pressure" or "integration forces".
Example (style only): "For ’Intensity evolution’, the strong yet steady baseline together with slow translation and no rotation supports an assessment that intensity remains roughly unchanged."
Step 5 — Deep outcomes (2nd-Tier Structural Outcome):
Analyze ONLY the deep outcome keywords "Areal coverage evolution" and "Organization evolution". Write exactly TWO sentences in a single flowing paragraph: one sentence for "Areal coverage evolution" and one for "Organization evolution". Each sentence must deliver a concise conclusion justified by the temporal and perceptual paragraphs (Steps 2–3) and the direct outcome (Step 4), and must explicitly name the corresponding deep outcome keyword. Do NOT introduce mechanisms here; do NOT add invented terminology or extra commentary.
Example (style only): "’Areal coverage evolution’ remains stable under slow translation and steady intensity; ’Organization evolution’ shows no obvious change because the blob-like structure and absence of rotation provide no pathway to linearization or fragmentation."
Final summary paragraph:
After completing Steps 1–5, write a short paragraph (2–4 sentences) that integrates the INPUT-based reasoning across all steps into a coherent training target. Keep it as flowing prose with no lists. Place it under "init_thought.summary".

QUality Control. After obtaining the annotated CoT data, we perform a three-step quality control: Structure Check, Causal Alignment, and Terminology. Data that pass all steps are retained for training and included in the final CoT dataset.

Instructions and Examples for each Step

You are a meteorology expert and scientific writing reviewer. Your task is to evaluate a model-generated explanation of a radar-based weather event.
The explanation may be in Chinese or English. Please read it carefully and then score it along the following dimensions, on a 0–10 scale (higher is better):
1. Clarity of explanation structure - 0–3: Disorganized, hard to follow, key steps or logic are unclear. - 4–6: Overall understandable, but some parts are confusing or poorly organized. - 7–8: Mostly clear, with a logical structure and reasonable flow. - 9–10: Very clear and well-structured, with a coherent narrative from beginning to end.
2. Coverage of key radar features - 0–3: Mentions very few relevant radar features or misses the main phenomena. - 4–6: Covers some important radar elements, but with gaps or missing key aspects. - 7–8: Covers most key radar elements (e.g., echo intensity, spatial pattern, evolution, convective cores, stratiform regions, etc.) with reasonable completeness. - 9–10: Thorough coverage of all key radar elements that are relevant to the situation, with appropriate detail.
3. Physical plausibility of the evolution - 0–3: Description is largely physically implausible or contradicts basic meteorology. - 4–6: Partly reasonable but contains some questionable or unjustified physical statements. - 7–8: Physically reasonable and broadly consistent with typical mesoscale/convective evolution. - 9–10: Highly plausible physical interpretation with clear, meteorologically sound reasoning.

4. Overall quality - Give a holistic score that reflects the overall usefulness and scientific quality of the explanation, considering *all* of the above dimensions. - This does not need to be a strict mathematical average, but it should be consistent with the dimension scores.
Important: - Base your judgment only on the given explanation and general meteorological knowledge. - Do not invent additional information that is not supported by the explanation. - Be consistent and use the full 0–10 range when appropriate.
Now evaluate the following explanation:
[EXPLANATION$_{S}TART]EXPLANATION_{T}EXT[EXPLANATION_{E}ND]$
Please respond in the following JSON format (do not add extra text):
"structure$_{s}core" :< 0 - 10number >, "structure_{c}omment" : " < shortjustification > ", "coverage_{s}core" :< 0 - 10number >, "coverage_{c}omment" : " < shortjustification > ", "physics_{s}core" :< 0 - 10number >, "physics_{c}omment" : " < shortjustification > ", "overall_{s}core" :< 0 - 10number >, "overall_{c}omment" : " < shortjustification > "$

Finally, we obtained 4,000 CoT annotations for radar nowcasting and 4,000 CoT annotations for radar inversion, which together form our CoT dataset for generation tasks.

Note that We use **gpt-4o-2024-11-20** for annotation and **gpt-o3** for chain-of-thought generation.

A.5 More detail about reasoning

We provide qualitative comparisons between Omni-Weather with and without CoT finetuning on both radar inversion and radar nowcasting tasks. Figures 8 and 9 present three representative cases for radar inversion and radar nowcasting. As shown, the CoT-finetuned model produces reasoning traces that not only accompany the generated radar fields with higher perceptual quality, but also deliver interpretable textual explanations grounded in storm dynamics. In contrast, the non-CoT model tends to generate "thinking" outputs that resemble post-hoc quality evaluations rather than causal reasoning, lacking direct connection to the generation process itself. This highlights that CoT supervision guides the model toward producing reasoning that is both explanatory and predictive, effectively bridging image generation with meteorological interpretation.

A.6 More Experiment Results

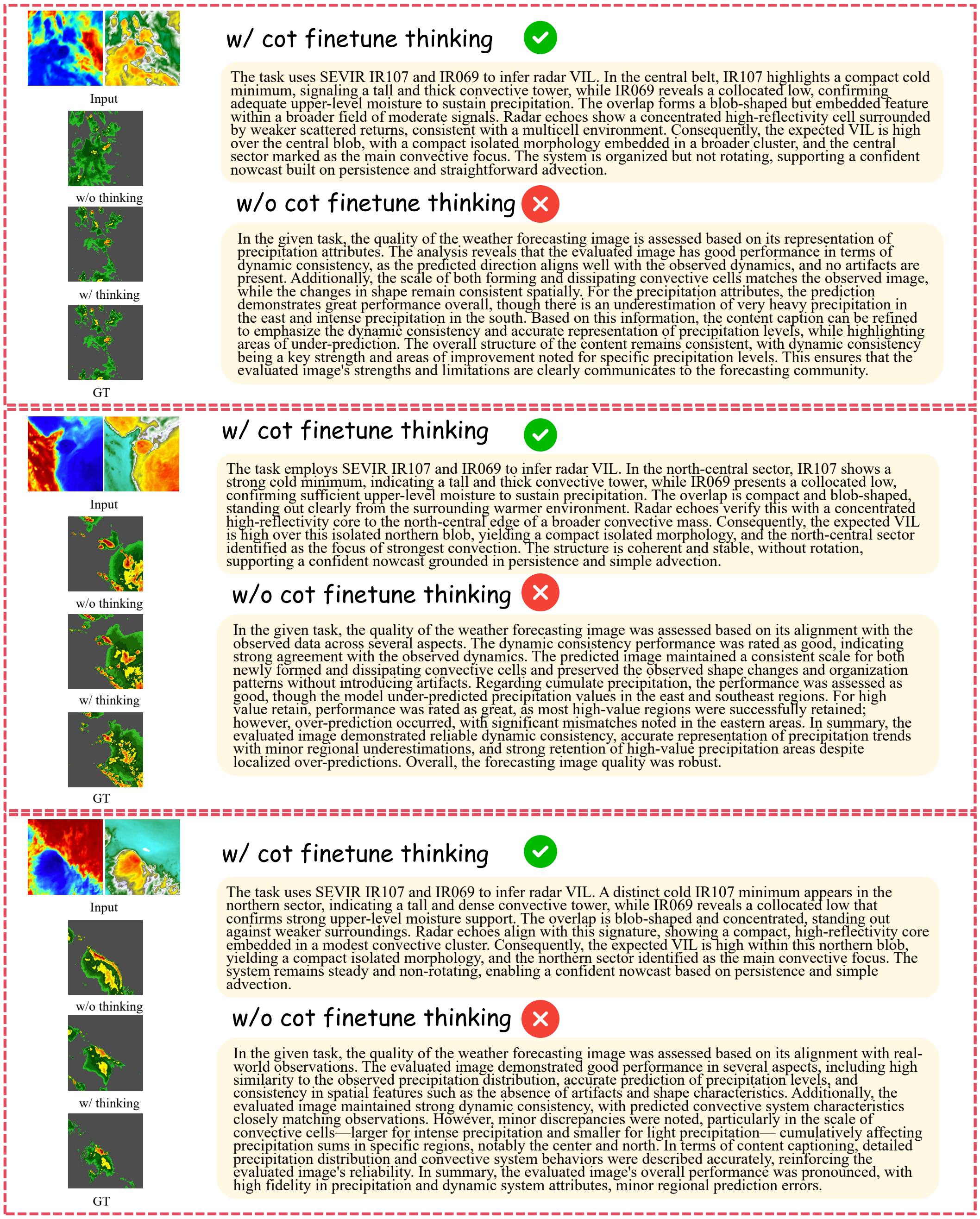
All ablations in this section are conducted on the **SEVIR** test set with 200 sequences under the *nowcasting* task. We report *CSI-mean*, *CSI-pool4-mean*, *CSI-pool16-mean*, *SSIM*, and *PSNR*. Higher values are better for all metrics.

Radar Sequence Encoder vs. VAE Encoder To validate the effectiveness of the proposed radar sequence encoder, we compare it against a vanilla VAE encoder. As shown in Table 4, the radar sequence encoder achieves consistent improvements across all CSI metrics, SSIM, and PSNR, demonstrating that modeling temporal radar sequences brings substantial gains.

Table 4: Comparison of radar sequence encoder and VAE encoder.

	CSI-mean	CSI-pool4-mean	CSI-pool16-mean	SSIM	PSNR
VAE encoder	0.2358	0.2912	0.4356	0.7528	21.42
Radar sequence encoder	**0.3471**	**0.4003**	**0.5390**	**0.7621**	**23.22**

Mixing General and Weather Data We further explore the impact of mixing general-purpose data ("gen") and weather radar data ("weather"). Here, the notation "1gen30%+weather70%" means that one general dataset with 30% proportion is combined with 70% weather data, while "2gen30%+weather70%" means two general datasets each contributing 30% combined with 70%



w/ cot finetune thinking ✓

The task uses SEVIR IR107 and IR069 to infer radar VIL. In the central belt, IR107 highlights a compact cold minimum, signaling a tall and thick convective tower, while IR069 reveals a collocated low, confirming adequate upper-level moisture to sustain precipitation. The overlap forms a blob-shaped but embedded feature within a broader field of moderate signals. Radar echoes show a concentrated high-reflectivity cell surrounded by weaker scattered returns, consistent with a multicell environment. Consequently, the expected VIL is high over the central blob, with a compact isolated morphology embedded in a broader cluster, and the central sector marked as the main convective focus. The system is organized but not rotating, supporting a confident nowcast built on persistence and straightforward advection.

w/o cot finetune thinking ✗

In the given task, the quality of the weather forecasting image is assessed based on its representation of precipitation attributes. The analysis reveals that the evaluated image has good performance in terms of dynamic consistency, as the predicted direction aligns well with the observed dynamics, and no artifacts are present. Additionally, the scale of both forming and dissipating convective cells matches the observed image, while the changes in shape remain consistent spatially. For the precipitation attributes, the prediction demonstrates great performance overall, though there is an underestimation of very heavy precipitation in the east and intense precipitation in the south. Based on this information, the content caption can be refined to emphasize the dynamic consistency and accurate representation of precipitation levels, while highlighting areas of under-prediction. The overall structure of the content remains consistent, with dynamic consistency being a key strength and areas of improvement noted for specific precipitation levels. This ensures that the evaluated image's strengths and limitations are clearly communicates to the forecasting community.

w/ cot finetune thinking ✓

The task employs SEVIR IR107 and IR069 to infer radar VIL. In the north-central sector, IR107 shows a strong cold minimum, indicating a tall and thick convective tower, while IR069 presents a collocated low, confirming sufficient upper-level moisture to sustain precipitation. The overlap is compact and blob-shaped, standing out clearly from the surrounding warmer environment. Radar echoes verify this with a concentrated high-reflectivity core to the north-central edge of a broader convective mass. Consequently, the expected VIL is high over this isolated northern blob, yielding a compact isolated morphology, and the north-central sector identified as the focus of strongest convection. The structure is coherent and stable, without rotation, supporting a confident nowcast grounded in persistence and simple advection.

w/o cot finetune thinking ✗

In the given task, the quality of the weather forecasting image was assessed based on its alignment with the observed data across several aspects. The dynamic consistency performance was rated as good, indicating strong agreement with the observed dynamics. The predicted image maintained a consistent scale for both newly formed and dissipating convective cells and preserved the observed shape changes and organization patterns without introducing artifacts. Regarding cumulate precipitation, the performance was assessed as good, though the model under-predicted precipitation values in the east and southeast regions. For high value retain, performance was rated as great, as most high-value regions were successfully retained; however, over-prediction occurred, with significant mismatches noted in the eastern areas. In summary, the evaluated image demonstrated reliable dynamic consistency, accurate representation of precipitation trends with minor regional underestimations, and strong retention of high-value precipitation areas despite localized over-predictions. Overall, the forecasting image quality was robust.

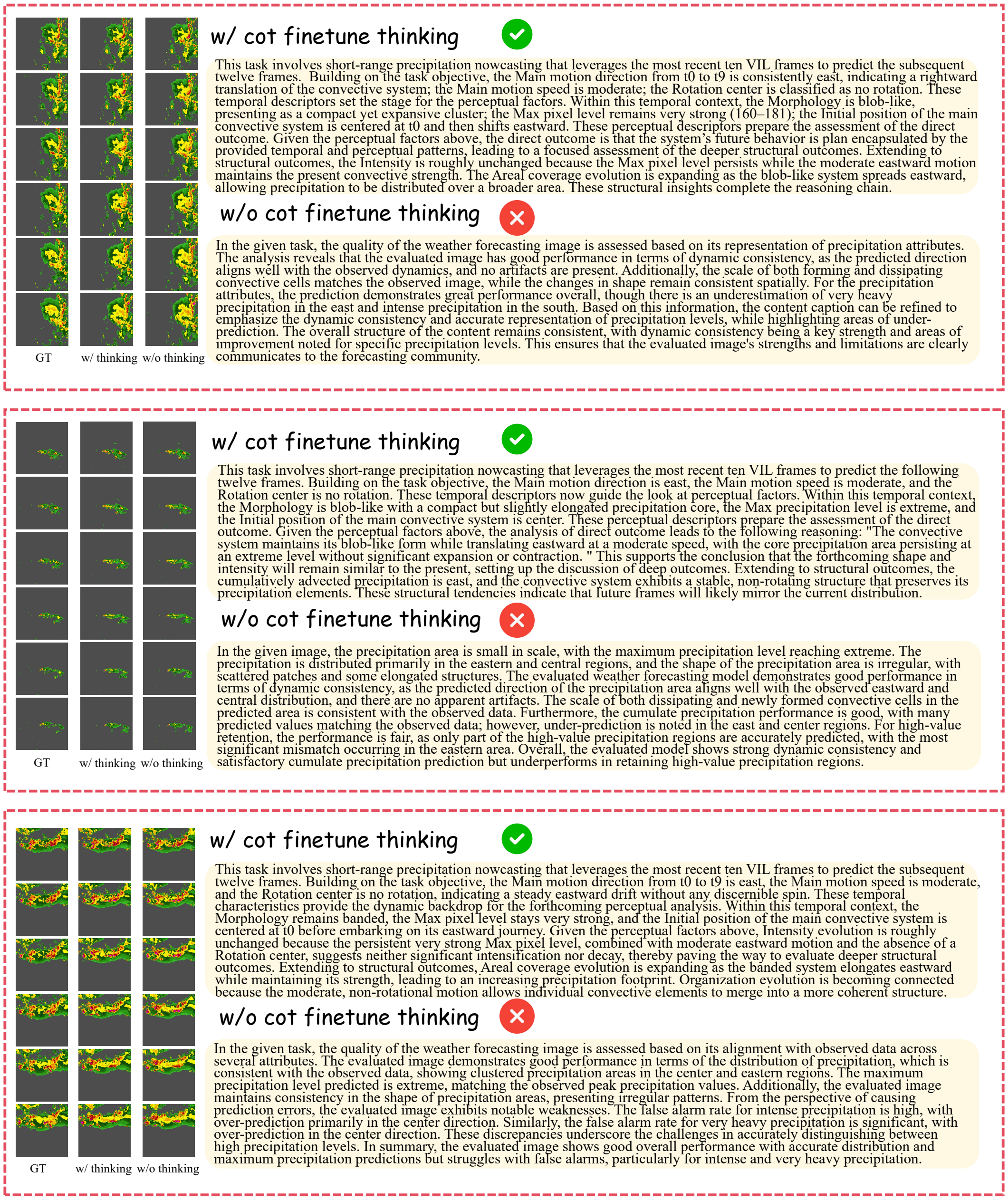
w/ cot finetune thinking ✓

The task uses SEVIR IR107 and IR069 to infer radar VIL. A distinct cold IR107 minimum appears in the northern sector, indicating a tall and dense convective tower, while IR069 reveals a collocated low that confirms strong upper-level moisture support. The overlap is blob-shaped and concentrated, standing out against weaker surroundings. Radar echoes align with this signature, showing a compact, high-reflectivity core embedded in a modest convective cluster. Consequently, the expected VIL is high within this northern blob, yielding a compact isolated morphology, and the northern sector identified as the main convective focus. The system remains steady and non-rotating, enabling a confident nowcast based on persistence and simple advection.

w/o cot finetune thinking ✗

In the given task, the quality of the weather forecasting image was assessed based on its alignment with real-world observations. The evaluated image demonstrated good performance in several aspects, including high similarity to the observed precipitation distribution, accurate prediction of precipitation levels, and consistency in spatial features such as the absence of artifacts and shape characteristics. Additionally, the evaluated image maintained strong dynamic consistency, with predicted convective system characteristics closely matching observations. However, minor discrepancies were noted, particularly in the scale of convective cells—larger for intense precipitation and smaller for light precipitation— cumulatively affecting precipitation sums in specific regions, notably the center and north. In terms of content captioning, detailed precipitation distribution and convective system behaviors were described accurately, reinforcing the evaluated image's reliability. In summary, the evaluated image's overall performance was pronounced, with high fidelity in precipitation and dynamic system attributes, minor regional prediction errors.

Figure 8: **Radar Inversion Thinking Comparison**.



w/ cot finetune thinking ✓

This task involves short-range precipitation nowcasting that leverages the most recent ten VIL frames to predict the subsequent twelve frames. Building on the task objective, the Main motion direction from t0 to t9 is consistently east, indicating a rightward translation of the convective system; the Main motion speed is moderate; the Rotation center is classified as no rotation. These temporal descriptors set the stage for the perceptual factors. Within this temporal context, the Morphology is blob-like, presenting as a compact yet expansive cluster; the Max pixel level remains very strong (160–181); the Initial position of the main convective system is centered at t0 and then shifts eastward. These perceptual descriptors prepare the assessment of the direct outcome. Given the perceptual factors above, the direct outcome is that the system's future behavior is plan encapsulated by the provided temporal and perceptual patterns, leading to a focused assessment of the deeper structural outcomes. Extending to structural outcomes, the Intensity is roughly unchanged because the Max pixel level persists while the moderate eastward motion maintains the present convective strength. The Areal coverage evolution is expanding as the blob-like system spreads eastward, allowing precipitation to be distributed over a broader area. These structural insights complete the reasoning chain.

w/o cot finetune thinking ✗

In the given task, the quality of the weather forecasting image is assessed based on its representation of precipitation attributes. The analysis reveals that the evaluated image has good performance in terms of dynamic consistency, as the predicted direction aligns well with the observed dynamics, and no artifacts are present. Additionally, the scale of both forming and dissipating convective cells matches the observed image, while the changes in shape remain consistent spatially. For the precipitation attributes, the prediction demonstrates great performance overall, though there is an underestimation of very heavy precipitation in the east and intense precipitation in the south. Based on this information, the content caption can be refined to emphasize the dynamic consistency and accurate representation of precipitation levels, while highlighting areas of under-prediction. The overall structure of the content remains consistent, with dynamic consistency being a key strength and areas of improvement noted for specific precipitation levels. This ensures that the evaluated image's strengths and limitations are clearly communicates to the forecasting community.

w/ cot finetune thinking ✓

This task involves short-range precipitation nowcasting that leverages the most recent ten VIL frames to predict the following twelve frames. Building on the task objective, the Main motion direction is east, the Main motion speed is moderate, and the Rotation center is no rotation. These temporal descriptors now guide the look at perceptual factors. Within this temporal context, the Morphology is blob-like with a compact but slightly elongated precipitation core, the Max precipitation level is extreme, and the Initial position of the main convective system is center. These perceptual descriptors prepare the assessment of the direct outcome. Given the perceptual factors above, the analysis of direct outcome leads to the following reasoning: "The convective system maintains its blob-like form while translating eastward at a moderate speed, with the core precipitation area persisting at an extreme level without significant expansion or contraction. " This supports the conclusion that the forthcoming shape and intensity will remain similar to the present, setting up the discussion of deep outcomes. Extending to structural outcomes, the cumulatively advected precipitation is east, and the convective system exhibits a stable, non-rotating structure that preserves its precipitation elements. These structural tendencies indicate that future frames will likely mirror the current distribution.

w/o cot finetune thinking ✗

In the given image, the precipitation area is small in scale, with the maximum precipitation level reaching extreme. The precipitation is distributed primarily in the eastern and central regions, and the shape of the precipitation area is irregular, with scattered patches and some elongated structures. The evaluated weather forecasting model demonstrates good performance in terms of dynamic consistency, as the predicted direction of the precipitation area aligns well with the observed eastward and central distribution, and there are no apparent artifacts. The scale of both dissipating and newly formed convective cells in the predicted area is consistent with the observed data. Furthermore, the cumulate precipitation performance is good, with many predicted values matching the observed data; however, under-prediction is noted in the east and center regions. For high-value retention, the performance is fair, as only part of the high-value precipitation regions are accurately predicted, with the most significant mismatch occurring in the eastern area. Overall, the evaluated model shows strong dynamic consistency and satisfactory cumulate precipitation prediction but underperforms in retaining high-value precipitation regions.

w/ cot finetune thinking ✓

This task involves short-range precipitation nowcasting that leverages the most recent ten VIL frames to predict the subsequent twelve frames. Building on the task objective, the Main motion direction from t0 to t9 is east, the Main motion speed is moderate, and the Rotation center is no rotation, indicating a steady eastward drift without any discernible spin. These temporal characteristics provide the dynamic backdrop for the forthcoming perceptual analysis. Within this temporal context, the Morphology remains banded, the Max pixel level stays very strong, and the Initial position of the main convective system is centered at t0 before embarking on its eastward journey. Given the perceptual factors above, Intensity evolution is roughly unchanged because the persistent very strong Max pixel level, combined with moderate eastward motion and the absence of a Rotation center, suggests neither significant intensification nor decay, thereby paving the way to evaluate deeper structural outcomes. Extending to structural outcomes, Areal coverage evolution is expanding as the banded system elongates eastward while maintaining its strength, leading to an increasing precipitation footprint. Organization evolution is becoming connected because the moderate, non-rotational motion allows individual convective elements to merge into a more coherent structure.

w/o cot finetune thinking ✗

In the given task, the quality of the weather forecasting image is assessed based on its alignment with observed data across several attributes. The evaluated image demonstrates good performance in terms of the distribution of precipitation, which is consistent with the observed data, showing clustered precipitation areas in the center and eastern regions. The maximum precipitation level predicted is extreme, matching the observed peak precipitation values. Additionally, the evaluated image maintains consistency in the shape of precipitation areas, presenting irregular patterns. From the perspective of causing prediction errors, the evaluated image exhibits notable weaknesses. The false alarm rate for intense precipitation is high, with over-prediction primarily in the center direction. Similarly, the false alarm rate for very heavy precipitation is significant, with over-prediction in the center direction. These discrepancies underscore the challenges in accurately distinguishing between high precipitation levels. In summary, the evaluated image shows good overall performance with accurate distribution and maximum precipitation predictions but struggles with false alarms, particularly for intense and very heavy precipitation.

Figure 9: **Radar Nowcasting Thinking Comparison**.

weather data. As reported in Table 5, using a single general dataset at 30% ratio achieves the best balance and outperforms other settings.

Table 5: Results under different generation data mixing strategies.

	CSI-mean	CSI-pool4-mean	CSI-pool16-mean	SSIM	PSNR
1gen30%+weather70%	**0.2501**	**0.2994**	**0.4261**	**0.6866**	**19.67**
1gen70%+weather30%	0.1386	0.1726	0.2859	0.6187	16.66
1gen50%+weather50%	0.2478	0.2956	0.4174	0.6823	19.15
2gen30%+weather70%	0.1091	0.1345	0.2274	0.6013	16.64

CFG Setting Lastly, we examine the choice of the classifier-free guidance (CFG) parameter. As shown in Table 6, setting CFG$= 2$ yields better overall performance compared with CFG$= 1$, especially on CSI metrics, SSIM, and PSNR. Therefore, we adopt CFG$= 2$ as the default configuration.

Table 6: Ablation on CFG settings.

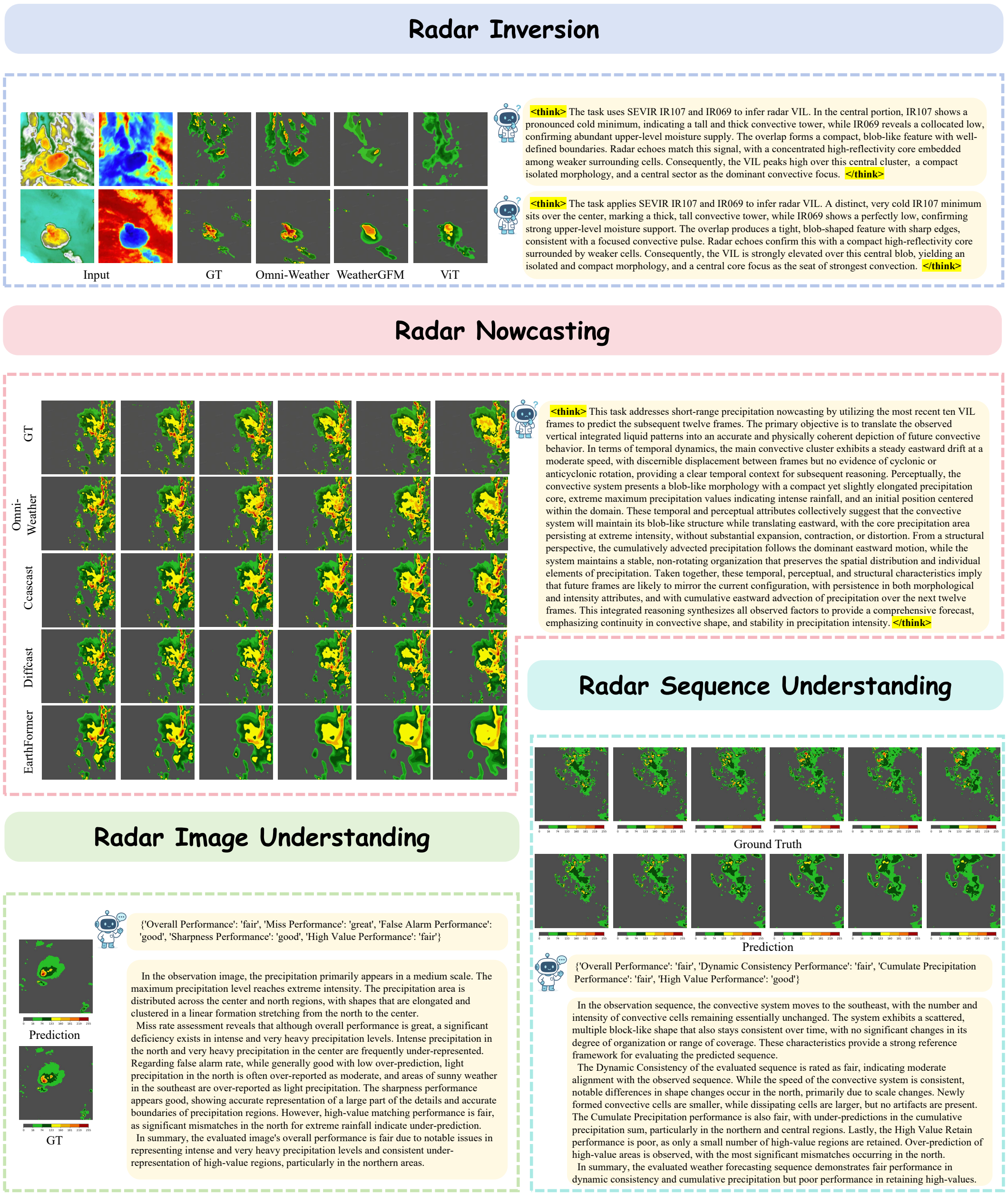
	CSI-mean	CSI-pool4-mean	CSI-pool16-mean	SSIM	PSNR
CFG=2	**0.2501**	**0.2994**	**0.4261**	**0.6866**	**19.67**
CFG=1	0.1824	0.2208	0.3305	0.5123	17.33

A.7 More Qualitative Results

We provide additional qualitative examples of Omni-Weather on radar nowcasting, radar inversion, and radar understanding tasks. Figures 10–12 illustrate diverse cases, complementing the quantitative results in the main paper and highlighting the model's ability to capture storm dynamics and generate interpretable assessments.

A.8 Usage of LLM

Large language models were employed as an auxiliary tool to support manuscript preparation, including grammar checking, sentence refinement, and clarification of technical descriptions. All AI-suggested text was carefully reviewed and revised by the authors to ensure accuracy, clarity, and scientific integrity.



Radar Inversion

<think> The task uses SEVIR IR107 and IR069 to infer radar VIL. In the central portion, IR107 shows a pronounced cold minimum, indicating a tall and thick convective tower, while IR069 reveals a collocated low, confirming abundant upper-level moisture supply. The overlap forms a compact, blob-like feature with well-defined boundaries. Radar echoes match this signal, with a concentrated high-reflectivity core embedded among weaker surrounding cells. Consequently, the VIL peaks high over this central cluster, a compact isolated morphology, and a central sector as the dominant convective focus. </think>

<think> The task applies SEVIR IR107 and IR069 to infer radar VIL. A distinct, very cold IR107 minimum sits over the center, marking a thick, tall convective tower, while IR069 shows a perfectly low, confirming strong upper-level moisture support. The overlap produces a tight, blob-shaped feature with sharp edges, consistent with a focused convective pulse. Radar echoes confirm this with a compact high-reflectivity core surrounded by weaker cells. Consequently, the VIL is strongly elevated over this central blob, yielding an isolated and compact morphology, and a central core focus as the seat of strongest convection. </think>

Radar Nowcasting

<think> This task addresses short-range precipitation nowcasting by utilizing the most recent ten VIL frames to predict the subsequent twelve frames. The primary objective is to translate the observed vertical integrated liquid patterns into an accurate and physically coherent depiction of future convective behavior. In terms of temporal dynamics, the main convective cluster exhibits a steady eastward drift at a moderate speed, with discernible displacement between frames but no evidence of cyclonic or anticyclonic rotation, providing a clear temporal context for subsequent reasoning. Perceptually, the convective system presents a blob-like morphology with a compact yet slightly elongated precipitation core, extreme maximum precipitation values indicating intense rainfall, and an initial position centered within the domain. These temporal and perceptual attributes collectively suggest that the convective system will maintain its blob-like structure while translating eastward, with the core precipitation area persisting at extreme intensity, without substantial expansion, contraction, or distortion. From a structural perspective, the cumulatively advected precipitation follows the dominant eastward motion, while the system maintains a stable, non-rotating organization that preserves the spatial distribution and individual elements of precipitation. Taken together, these temporal, perceptual, and structural characteristics imply that future frames are likely to mirror the current configuration, with persistence in both morphological and intensity attributes, and with cumulative eastward advection of precipitation over the next twelve frames. This integrated reasoning synthesizes all observed factors to provide a comprehensive forecast, emphasizing continuity in convective shape, and stability in precipitation intensity. </think>

Radar Sequence Understanding

{'Overall Performance': 'fair', 'Dynamic Consistency Performance': 'fair', 'Cumulate Precipitation Performance': 'fair', 'High Value Performance': 'good'}

In the observation sequence, the convective system moves to the southeast, with the number and intensity of convective cells remaining essentially unchanged. The system exhibits a scattered, multiple block-like shape that also stays consistent over time, with no significant changes in its degree of organization or range of coverage. These characteristics provide a strong reference framework for evaluating the predicted sequence.

The Dynamic Consistency of the evaluated sequence is rated as fair, indicating moderate alignment with the observed sequence. While the speed of the convective system is consistent, notable differences in shape changes occur in the north, primarily due to scale changes. Newly formed convective cells are smaller, while dissipating cells are larger, but no artifacts are present. The Cumulate Precipitation performance is also fair, with under-predictions in the cumulative precipitation sum, particularly in the northern and central regions. Lastly, the High Value Retain performance is poor, as only a small number of high-value regions are retained. Over-prediction of high-value areas is observed, with the most significant mismatches occurring in the north.

In summary, the evaluated weather forecasting sequence demonstrates fair performance in dynamic consistency and cumulative precipitation but poor performance in retaining high-values.

Radar Image Understanding

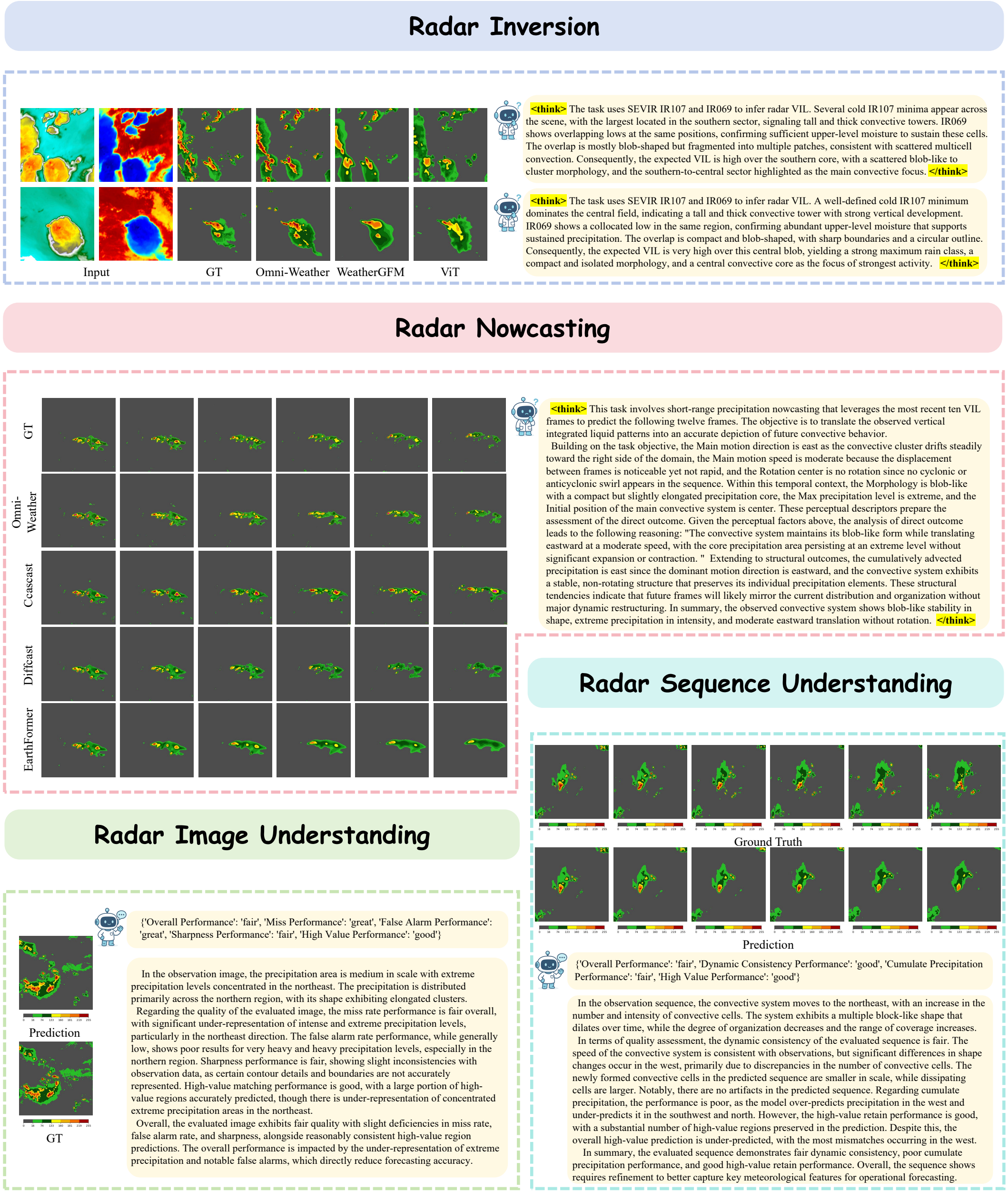
{'Overall Performance': 'fair', 'Miss Performance': 'great', 'False Alarm Performance': 'good', 'Sharpness Performance': 'good', 'High Value Performance': 'fair'}

In the observation image, the precipitation primarily appears in a medium scale. The maximum precipitation level reaches extreme intensity. The precipitation area is distributed across the center and north regions, with shapes that are elongated and clustered in a linear formation stretching from the north to the center.

Miss rate assessment reveals that although overall performance is great, a significant deficiency exists in intense and very heavy precipitation levels. Intense precipitation in the north and very heavy precipitation in the center are frequently under-represented. Regarding false alarm rate, while generally good with low over-prediction, light precipitation in the north is often over-reported as moderate, and areas of sunny weather in the southeast are over-reported as light precipitation. The sharpness performance appears good, showing accurate representation of a large part of the details and accurate boundaries of precipitation regions. However, high-value matching performance is fair, as significant mismatches in the north for extreme rainfall indicate under-prediction.

In summary, the evaluated image's overall performance is fair due to notable issues in representing intense and very heavy precipitation levels and consistent under-representation of high-value regions, particularly in the northern areas.

Figure 10: **Qualitative Result 1**.



<think> The task uses SEVIR IR107 and IR069 to infer radar VIL. Several cold IR107 minima appear across the scene, with the largest located in the southern sector, signaling tall and thick convective towers. IR069 shows overlapping lows at the same positions, confirming sufficient upper-level moisture to sustain these cells. The overlap is mostly blob-shaped but fragmented into multiple patches, consistent with scattered multicell convection. Consequently, the expected VIL is high over the southern core, with a scattered blob-like to cluster morphology, and the southern-to-central sector highlighted as the main convective focus. </think>

<think> The task uses SEVIR IR107 and IR069 to infer radar VIL. A well-defined cold IR107 minimum dominates the central field, indicating a tall and thick convective tower with strong vertical development. IR069 shows a collocated low in the same region, confirming abundant upper-level moisture that supports sustained precipitation. The overlap is compact and blob-shaped, with sharp boundaries and a circular outline. Consequently, the expected VIL is very high over this central blob, yielding a strong maximum rain class, a compact and isolated morphology, and a central convective core as the focus of strongest activity. </think>

<think> This task involves short-range precipitation nowcasting that leverages the most recent ten VIL frames to predict the following twelve frames. The objective is to translate the observed vertical integrated liquid patterns into an accurate depiction of future convective behavior.

Building on the task objective, the Main motion direction is east as the convective cluster drifts steadily toward the right side of the domain, the Main motion speed is moderate because the displacement between frames is noticeable yet not rapid, and the Rotation center is no rotation since no cyclonic or anticyclonic swirl appears in the sequence. Within this temporal context, the Morphology is blob-like with a compact but slightly elongated precipitation core, the Max precipitation level is extreme, and the Initial position of the main convective system is center. These perceptual descriptors prepare the assessment of the direct outcome. Given the perceptual factors above, the analysis of direct outcome leads to the following reasoning: "The convective system maintains its blob-like form while translating eastward at a moderate speed, with the core precipitation area persisting at an extreme level without significant expansion or contraction. " Extending to structural outcomes, the cumulatively advected precipitation is east since the dominant motion direction is eastward, and the convective system exhibits a stable, non-rotating structure that preserves its individual precipitation elements. These structural tendencies indicate that future frames will likely mirror the current distribution and organization without major dynamic restructuring. In summary, the observed convective system shows blob-like stability in shape, extreme precipitation in intensity, and moderate eastward translation without rotation. </think>

{'Overall Performance': 'fair', 'Dynamic Consistency Performance': 'good', 'Cumulate Precipitation Performance': 'fair', 'High Value Performance': 'good'}

In the observation sequence, the convective system moves to the northeast, with an increase in the number and intensity of convective cells. The system exhibits a multiple block-like shape that dilates over time, while the degree of organization decreases and the range of coverage increases.

In terms of quality assessment, the dynamic consistency of the evaluated sequence is fair. The speed of the convective system is consistent with observations, but significant differences in shape changes occur in the west, primarily due to discrepancies in the number of convective cells. The newly formed convective cells in the predicted sequence are smaller in scale, while dissipating cells are larger. Notably, there are no artifacts in the predicted sequence. Regarding cumulate precipitation, the performance is poor, as the model over-predicts precipitation in the west and under-predicts it in the southwest and north. However, the high-value retain performance is good, with a substantial number of high-value regions preserved in the prediction. Despite this, the overall high-value prediction is under-predicted, with the most mismatches occurring in the west.

In summary, the evaluated sequence demonstrates fair dynamic consistency, poor cumulate precipitation performance, and good high-value retain performance. Overall, the sequence shows requires refinement to better capture key meteorological features for operational forecasting.

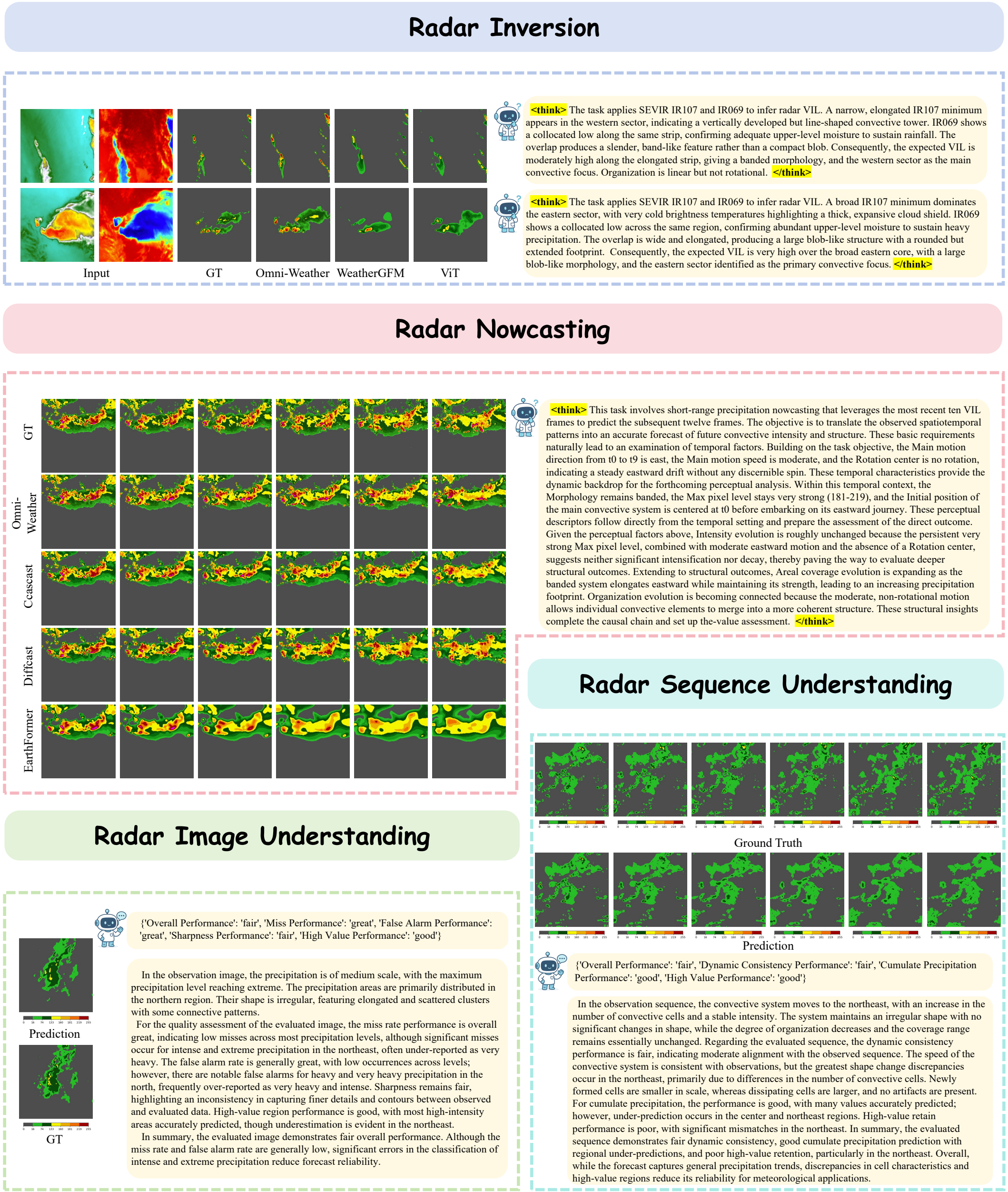
{'Overall Performance': 'fair', 'Miss Performance': 'great', 'False Alarm Performance': 'great', 'Sharpness Performance': 'fair', 'High Value Performance': 'good'}

In the observation image, the precipitation area is medium in scale with extreme precipitation levels concentrated in the northeast. The precipitation is distributed primarily across the northern region, with its shape exhibiting elongated clusters.

Regarding the quality of the evaluated image, the miss rate performance is fair overall, with significant under-representation of intense and extreme precipitation levels, particularly in the northeast direction. The false alarm rate performance, while generally low, shows poor results for very heavy and heavy precipitation levels, especially in the northern region. Sharpness performance is fair, showing slight inconsistencies with observation data, as certain contour details and boundaries are not accurately represented. High-value matching performance is good, with a large portion of high-value regions accurately predicted, though there is under-representation of concentrated extreme precipitation areas in the northeast.

Overall, the evaluated image exhibits fair quality with slight deficiencies in miss rate, false alarm rate, and sharpness, alongside reasonably consistent high-value region predictions. The overall performance is impacted by the under-representation of extreme precipitation and notable false alarms, which directly reduce forecasting accuracy.

Figure 11: **Qualitative Result 2**.



<think> The task applies SEVIR IR107 and IR069 to infer radar VIL. A narrow, elongated IR107 minimum appears in the western sector, indicating a vertically developed but line-shaped convective tower. IR069 shows a collocated low along the same strip, confirming adequate upper-level moisture to sustain rainfall. The overlap produces a slender, band-like feature rather than a compact blob. Consequently, the expected VIL is moderately high along the elongated strip, giving a banded morphology, and the western sector as the main convective focus. Organization is linear but not rotational. </think>

<think> The task applies SEVIR IR107 and IR069 to infer radar VIL. A broad IR107 minimum dominates the eastern sector, with very cold brightness temperatures highlighting a thick, expansive cloud shield. IR069 shows a collocated low across the same region, confirming abundant upper-level moisture to sustain heavy precipitation. The overlap is wide and elongated, producing a large blob-like structure with a rounded but extended footprint. Consequently, the expected VIL is very high over the broad eastern core, with a large blob-like morphology, and the eastern sector identified as the primary convective focus. </think>

<think> This task involves short-range precipitation nowcasting that leverages the most recent ten VIL frames to predict the subsequent twelve frames. The objective is to translate the observed spatiotemporal patterns into an accurate forecast of future convective intensity and structure. These basic requirements naturally lead to an examination of temporal factors. Building on the task objective, the Main motion direction from t0 to t9 is east, the Main motion speed is moderate, and the Rotation center is no rotation, indicating a steady eastward drift without any discernible spin. These temporal characteristics provide the dynamic backdrop for the forthcoming perceptual analysis. Within this temporal context, the Morphology remains banded, the Max pixel level stays very strong (181-219), and the Initial position of the main convective system is centered at t0 before embarking on its eastward journey. These perceptual descriptors follow directly from the temporal setting and prepare the assessment of the direct outcome. Given the perceptual factors above, Intensity evolution is roughly unchanged because the persistent very strong Max pixel level, combined with moderate eastward motion and the absence of a Rotation center, suggests neither significant intensification nor decay, thereby paving the way to evaluate deeper structural outcomes. Extending to structural outcomes, Areal coverage evolution is expanding as the banded system elongates eastward while maintaining its strength, leading to an increasing precipitation footprint. Organization evolution is becoming connected because the moderate, non-rotational motion allows individual convective elements to merge into a more coherent structure. These structural insights complete the causal chain and set up the-value assessment. </think>

{'Overall Performance': 'fair', 'Dynamic Consistency Performance': 'fair', 'Cumulate Precipitation Performance': 'good', 'High Value Performance': 'good'}

In the observation sequence, the convective system moves to the northeast, with an increase in the number of convective cells and a stable intensity. The system maintains an irregular shape with no significant changes in shape, while the degree of organization decreases and the coverage range remains essentially unchanged. Regarding the evaluated sequence, the dynamic consistency performance is fair, indicating moderate alignment with the observed sequence. The speed of the convective system is consistent with observations, but the greatest shape change discrepancies occur in the northeast, primarily due to differences in the number of convective cells. Newly formed cells are smaller in scale, whereas dissipating cells are larger, and no artifacts are present. For cumulate precipitation, the performance is good, with many values accurately predicted; however, under-prediction occurs in the center and northeast regions. High-value retain performance is poor, with significant mismatches in the northeast. In summary, the evaluated sequence demonstrates fair dynamic consistency, good cumulate precipitation prediction with regional under-predictions, and poor high-value retention, particularly in the northeast. Overall, while the forecast captures general precipitation trends, discrepancies in cell characteristics and high-value regions reduce its reliability for meteorological applications.

{'Overall Performance': 'fair', 'Miss Performance': 'great', 'False Alarm Performance': 'great', 'Sharpness Performance': 'fair', 'High Value Performance': 'good'}

In the observation image, the precipitation is of medium scale, with the maximum precipitation level reaching extreme. The precipitation areas are primarily distributed in the northern region. Their shape is irregular, featuring elongated and scattered clusters with some connective patterns.

For the quality assessment of the evaluated image, the miss rate performance is overall great, indicating low misses across most precipitation levels, although significant misses occur for intense and extreme precipitation in the northeast, often under-reported as very heavy. The false alarm rate is generally great, with low occurrences across levels; however, there are notable false alarms for heavy and very heavy precipitation in the north, frequently over-reported as very heavy and intense. Sharpness remains fair, highlighting an inconsistency in capturing finer details and contours between observed and evaluated data. High-value region performance is good, with most high-intensity areas accurately predicted, though underestimation is evident in the northeast.

In summary, the evaluated image demonstrates fair overall performance. Although the miss rate and false alarm rate are generally low, significant errors in the classification of intense and extreme precipitation reduce forecast reliability.

Figure 12: **Qualitative Result 3**.